Automatic Detection and Classification of Prosodic Events

Andrew Rosenberg

Submitted in partial fulfillment of the
Requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2009

Abstract

Automatic Detection and Classification of Prosodic Events

Andrew Rosenberg

Prosody, or intonation, is a critically important component of spoken communication. The automatic extraction of prosodic information is necessary for machines to process speech with human levels of proficiency. In this thesis we describe work on the automatic detection and classification of prosodic events – specifically, pitch accents and prosodic phrase boundaries. We present novel techniques, feature representations and state of the art performance in each of these tasks. We also present three proof-of-concept applications – speech summarization, story segmentation and non-native speech assessment – showing that access to hypothesized prosodic event information can be used to improve the performance of downstream spoken language processing tasks. We believe the contributions of this thesis advance the understanding of prosodic events and the use of prosody in spoken language processing towards the goal of human-like processing of speech by machines.

Contents

List of Figures

List of Tables

Acknowledgements

My advisor Julia Hirschberg has my endless gratitude. During my time at Columbia University, her support, compassion, friendship, guidance and patience never wavered. I am in her debt.

Equally as important has been the love and kindness of my best friend, Rachelle Bergstein. Without her, I would be worse at everything.

I would also like to thank the following for their roles in my academic and personal life: Kathy McKeown, Dan Ellis, Martin Jansche, Mari Ostendorf, Ed Binkowski, Michael Joyce, Mike Roam, Anna Verdi, Fadi Biadsy, Bob Coyne, Noemie Elhadad, David Elson, Frank Enos, Michel Galley, Agustín Gravano, Kevin Lerman, Jackson Liscombe, Sameer Maksey, Ani Nenkova, Kristen Parton, Sara Stolbach, Kapil Thadani, Stefan Benus, Raul Fernandez, Nizar Habash, Vida Ha, Terry Tai, Herb Rosenberg, Jean Rosenberg, Gimme! Coffee (Brooklyn), Mojo Coffee House (New Orleans), and everyone who reminds me that there is a world outside.

Chapter 1

Introduction

Prosody, or intonation, is a critically important component of spoken communication. The relationship between the words that are being spoken and their prosody is commonly understood to be the difference between “the what” and “the how”. Where the words that comprise an utterance describe the lexical content of speech, prosody describes many aspects of the manner in which these words are spoken.

This dichotomy is a useful simplification of the relationship between the lexical content of speech and prosody. It allows researchers to examine prosodic variation without reference to the words that are being spoken. However, this division carries an implication that the meaning or desired interpretation of an utterance can be obtained from the lexical content of an utterance, “the what”, while prosody, “the how”, modifies this meaning, as an ancillary process. This division belies the interaction between the information streams present in the speech signal. For example, the meaning of the declarative statement “John loves Mary.” substantially differs from the interrogative “John loves Mary?” The interrogative interpretation is typically signaled by a rising pitch at at the end of the utterance, not present in a typical production of the declarative utterance. Prosody is also used to disambiguate multiple syntactic interpretations of an utterance. For instance, “Bill doesn’t drink because he’s unhappy” may mean that Bill drinks, but for a reason other than unhappiness, or it may mean that Bill doesn't drink, and the reason for his abstinence is his unhappiness. The lexical content of this utterance does not provide enough information to disambiguate these possibilities. However, the second interpretation – that due to Bill's unhappiness, he doesn't drink – is typically produced with a phrase boundary after "drink", as in "Bill doesn't drink — because he's unhappy". Prosodic variation provides a mechanism for syntactic disambiguation that is not present in the lexical content of an utterance. These examples represent instances where prosody is used to modify or disambiguate the intended interpretation of an utterance. There are, of course, instances where prosody is used to transmit additional information to a listener. In Standard American English (SAE), indications of speaker state are frequently conveyed through prosodic variation [123]. Emotional states, such as anger, frustration and happiness are displayed in a speaker's speech through prosodic variation [102]. Prosody also carries indicators of speaker states of incredulity, uncertainty [224, 92] and sarcasm [209]. Prosody plays a critical role in the participation in and interpretation of dialog. In addition to directly impacting the meaning of some utterances, like 'okay' [66], prosody is used to structure spoken material [1, 71], affecting the intended context of interpretation. Smooth turn taking behavior is made possible in part due to prosodic variation [65]. These examples represent only a subset of the communicative uses of prosodic variation. Prosody is critically important to the transmission of information via speech and its interaction with the spoken lexical material continue to be discovered.

Due to its many roles in human communication, proper interpretation of prosody is important to many spoken language processing tasks. Prosody plays a significant role in syntactic, semantic and pragmatic interpretation, as well as discourse and topic structuring. The use of prosodic information is not optional for any spoken language processing task which seeks to perform these tasks. Without access to these prosodic information streams, human-like interpretation of or interaction with spoken material will remain out of reach.

Prosodic analysis has already been shown to be helpful to many spoken language processing tasks including, but not limited to, automatic speech recognition [74], speech synthesis [48], speaker identification [228], language [212] and dialect identification [174], story/topic segmentation [192, 170], sentence segmentation [181], discourse segmentation [86], extractive speech summarization [131], punctuation insertion [40], and speech act classification [180, 2]. Despite these demonstrations of its importance, access to prosodic information is rarely considered to be a necessary component of spoken language processing. One major exception to this is speech synthesis applications, where prosodic assignment modules are staples of all state-of-the-art systems.

We have used the term "prosody" to refer to non-lexical speech content. However, prosodic variation occurs across distinct dimensions. Speaking rate, voice quality, loudness, pitch, phone duration, phrasing and accenting are some of the most heavily investigated dimensions of prosodic variation. We restrict the focus of this thesis to accenting and phrasing behavior. These two phenomena, accenting and phrasing, are referred to collectively as "prosodic events" in this thesis. Accenting is the act of making a word acoustically prominent from its surroundings. Phrasing is defined as a perceived disjuncture between words. We focus on these two events for a number of reasons. First of all, there is consensus among researchers in intonation that these events are important components of prosody in English. Second, there is a substantial amount of research into the acoustic and syntactic correlates of these phenomena. Finally, there is available material that has been annotated by expert labelers for the presence of accent and phrase boundaries.

This thesis describes techniques for the detection and classification of prosodic events – accents and phrase boundaries. We pursue three main goals in this thesis.

- **Novel Techniques** Throughout this thesis, we use novel techniques in the detection and classification of prosodic events. These techniques include new representations of known phenomena, and new machine learning techniques to leverage lexico-syntactic and acoustic information to improved detection and classification performance.

- **Improved Understanding** Through error analysis, descriptive statistics and comparing classification performance of distinct feature sets, we provide data-driven insights that serve to improve the scientific understanding of prosodic events in SAE.

- **Application** In the motivation of this work and discussion of experimental results, we demonstrate the use of prosodic information in spoken language processing tasks. The techniques and evaluations described in this thesis show that the extraction of prosodic information with high performance is possible. The applications described within this thesis (cf. Chapter 7) and elsewhere prove that this information can be used to improve performance of downstream spoken language processing tasks.

We use the ToBI standard of intonation to define pitch accent and phrase-ending intonation types for classification. Chapter 2 presents the ToBI standard in detail, and describes the experimental material that is used throughout the thesis. The body of the thesis is partitioned into four chapters. Chapter 3 describes research into the automatic detection of pitch accents. The major contributions in this chapter are the comparison of syllable- and word-based detection of pitch accent, the use of spectral information to accent detection and the presentation of corrected classifiers, an ensemble technique where classifier performance is improved by the presence of a secondary classifier to predict if the first is trusted or not. On Boston University Radio News Corpus (BURNC) material, this ensemble technique achieves state-of-the-art speaker-independent acoustic pitch accent detection accuracy, 84.95%. The previous highest reported result on this task was 80.09% [187].

We address the detection of prosodic phrase boundaries in Chapter 4. In this chapter, we develop a number of novel representations of acoustic reset, compare a variety of part-of-speech modeling techniques, and investigate the use of a top-down phrase detection technique, where intonational phrases are detected first, and intermediate phrase boundaries follow. The combination of syntactic and acoustic cues are used to predict phrase boundaries with accuracy over 90% on all evaluated corpora. On BDC-read, using AdaBoost with acoustic and syntactic features, we predict intonational phrase boundaries with 95.75% accuracy and F_1 of 0.824 under speaker independent evaluation. On BDC-spontaneous, the performance of this approach yields 92.55% accuracy and F_1 of 0.786 under speaker independent evaluation. The previous best accuracy on the BDC material, combining read and spontaneous material, was 90.58%.

We present research on the classification of pitch accents in Chapter 5. In this chapter, we describe Quantized Contour Modeling, a Bayesian classification structure, to classify pitch accent types. This classification task poses some unique challenges due to a skewed class distribution. Evaluation of automatic pitch accent type classification is difficult – the majority class baseline is greater than the rate of human agreement on this task. We find ensemble sampling to be well suited to addressing this challenge and able to train high performing classifiers.

Chapter 6 addresses the task of classifying phrase-ending intonation. The major contributions included in this chapter are the use of syntactic parse-tree features and the assessment of a variety of regions of analysis preceding intonational phrase boundaries in the classification of phrase-ending intonation. On BURNC material, we are able to use Quantized Contour Model posteriors with other acoustic features to classify combinations of phrase accents and boundary tones with 77.64% accuracy under speaker independent evaluation. The best previously published accuracy on this task is 72.4% [173] evaluated using material from a single BURNC speaker.

To further demonstrate the use of prosodic event information in spoken language processing tasks, we present three proof-of-concept applications in Chapter 7. These include story/topic segmentation, extractive summarization of broadcast news, and non-native speech identification. In Chaper 8, we conclude and propose directions for future work.

Chapter 2

Materials

In this chapter, we present the material we use in the prosodic event detection and classification research contained in this thesis. The ToBI standard of describing Standard American English intonation is heavily used throughout this work. We in the pitch accent and phrase ending classification chapters, we automatically predict the ToBI tone or tones associated with these prosodic events. Also, the intonation of the experimental material as been annotated using the ToBI standard. In Section 2.1, we describe the ToBI standard in detail. We present the Boston Directions Corpus in Section 2.2, the Boston University Radio News Corpus in Section 2.3 and the TDT-4 corpus in Section 2.4.

2.1 ToBI Standard

ToBI, a shortening for Tones and Break Indices, is a standard of describing the intonation of Standard American English (SAE) [183]. The ToBI system of intonation description minimally consists of four parallel, time-aligned tiers. These four are the Tone, Orthographic, Breaks and Miscellaneous tiers.

The content of the Tone tier is a linear sequence of pitch events, aligned in time. These are closely based on the intonational phonology described by Pierrehumbert [151]. The tone sequence contained in the Tone tier describes both the prosodic tune as well as a two-tiered prosodic phrase structure. Five types of pitch accents – pitch movements that correspond to perceived prominence of an associated word – are defined in the standard: H*, L*, L+H*, L*+H, H+!H*. In addition to these five, high tones (H) can be produced in a compressed pitch range, a phenomenon called catathesis [159]. These "downstepped" tones can only occur following a previous high tone, the basis of the pitch range compression, and are annotated as !H*, L+!H* and L*+!H*. Moreover, the ToBI standard has a mechanism for indicating uncertainty in pitch accent type annotation. If a labeler is uncertain about the type of pitch accent, though confident that the word does bear an accent, he or she indicates this using a X*? annotation. Detailed discussion of the differences between pitch accent types can be found in Section 5.1. Also in included in this tier are phrase accents and boundary tones which describe the intonation preceding prosodic phrase boundaries.

Two levels of prosodic phrasing are defined: the intermediate phrase and the intonational phrase. The presence of a prosodic phrase boundary is indicated by perceived disjuncture between two words. Intonational phrases boundaries are defined by the highest degree of disjuncture, and are often associated with silence. Each intonational phrase is comprised of one or more intermediate phrases. Boundaries between intermediate phrases that are not also intonational phrase boundaries are marked by increased disjuncture, but less than intonational phrase boundaries. The level of disjuncture between words is indicated on the Breaks tier. Each word boundary has an associated "break index", which can take a value from 0 to 4, indicating increased disjuncture. Break indices of '4' indicate intonational phrase boundaries, while '3' indices indicate intermediate phrase boundaries. Typical word boundaries have a break index of '1'.

Each intermediate phrase has an associated phrase accent, describing the pitch movement between the ultimate pitch accent and the phrase boundary. Phrase accents can have High (H-), downstepped High (!H-) or low (L-) tones. Intonational phrase boundaries have an additional boundary tone, to describe a final pitch movement. These can be high (H%) or low (L%). As with pitch accent types, uncertainty in phrase accent type is indicated using the X-? annotation, while an annotator uses the X%? annotation to indicate a potentially ambiguous boundary tone. Since each intonational phrase boundary also terminates an intermediate phrase, intonational phrase boundaries have associated phrase accents *and* boundary tones. Each intermediate phrase must contain at least one pitch accent. Moreover, the pitch range of each intermediate phrase is marked with the annotation, HiF0, at the time point of the highest pitch (f0) value within the highest pitch accent in the phrase.

The Miscellaneous tier allows for annotation of other information not explicitly contained in the standard. Common elements of the Miscellaneous tier include indications of audible breaths, laughter, coughing, unconventional pronunciations, and annotations of disfluency.

The accents and phrase boundaries included in the ToBI standard represent the prosodic events that are detected and classified in this thesis. Figure 2.1 contains an example of a ToBI annotation as displayed by Praat [20]. In addition to the tiers, the waveform, and pitch (blue line) and intensity (green line) contours are displayed to facilitate annotation. Examples of each pitch accent type are discussed in Chapter 5. Each phrase final type is presented in detail in Chapter 6.

2.2 Boston Directions Corpus

The Boston Directions Corpus (BDC) [138] is made up of elicited monologues spoken by four non-professional speakers, three male and one female. The speakers were given written instructions to perform a set of increasingly difficult direction giving tasks around Boston. A silent confederate traced the route described on a map of which both the confederate and the speaker had a copy. The speech was subsequently orthographically transcribed with false starts and other speech errors removed. At least two weeks later the subjects returned and read this transcribed speech. The read and spontaneous material has been prosodically labeled using the ToBI standard (cf. Section 2.1 and[183]). The BDC corpus contains

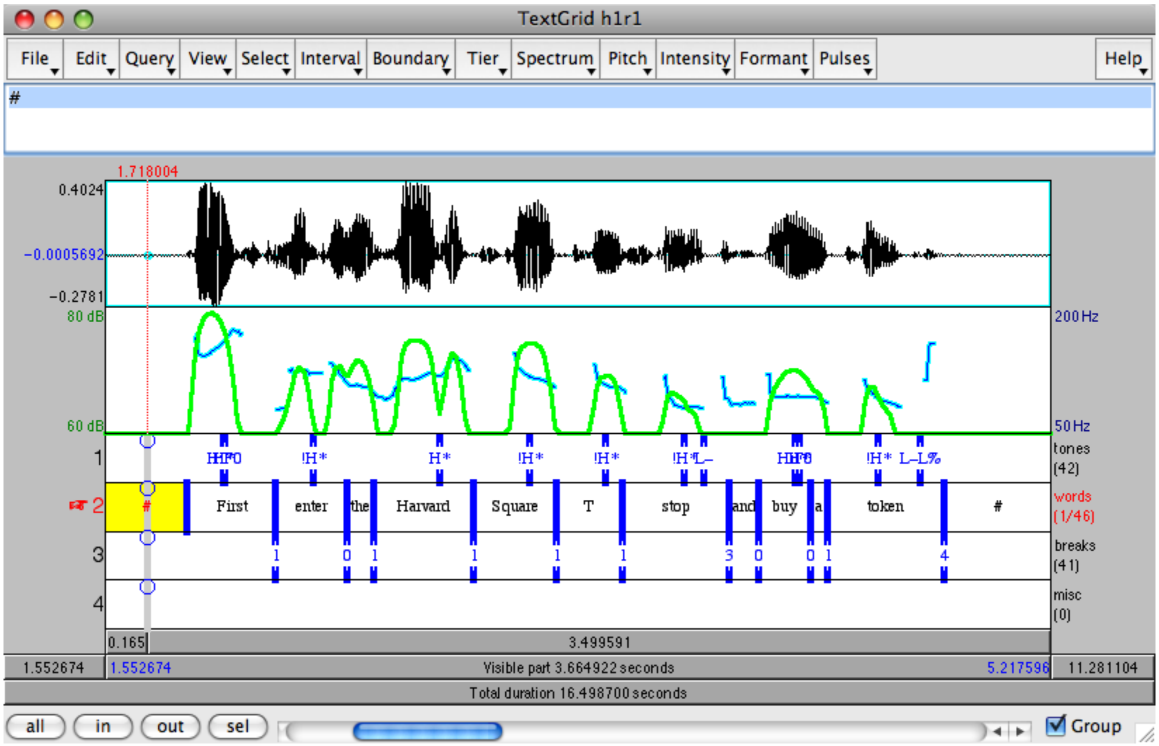


Figure 2.1: *Example of ToBI annotation of intonation. File h1r1 from the BDC-read corpus.*

approximately 60 minutes of annotated spontaneous speech and 50 minutes of read speech. Throughout this thesis, we treat the read and spontaneous material as separate corpora. This allows us to compare the performance of automatic detection and classification techniques on distinct genres of speech. The 50 minutes of read speech contain 10,831 words. There are 60 minutes of annotated spontaneous material containing 11,627 words. Distributions of ToBI tones present in these two corpora are presented in Tables 2.1, 2.2, 2.3 and 2.4.

H*	!H*	L+H*	L+!H*
2158 (0.47)	1344 (0.29)	555 (.12)	59 (0.01)

L*	L*+H	L*+!H	H+!H*	X*?
267 (0.06)	59 (0.01)	2 (0.00)	32 (0.01)	91 (0.02)

Table 2.1: *Distribution of pitch accent types in BDC-read. Accent rate: 42.2%*

L-L%	L-H%	H-L%	!H-L%	H-H%	X-?X%?
685 (0.49)	498 (0.36)	135 (0.10)	60 (0.04)	20 (0.01)	1 (0.00)

L-	H-	!H-	X-?
1661 (0.77)	409 (0.19)	84 (0.04)	1 (0.00)

L%	H%	X%?
880 (0.63)	518 (0.37)	1 (0.00)

Table 2.2: *Distributions of phrase ending tones (phrase accents and boundary tones) in BDC-read material. Mean intermediate phrase length: 5.03 words. Mean intonational phrase length: 7.74 words*

H*	!H*	L+H*	L+!H*
3322 (0.58)	1425 (0.25)	325 (0.06)	30 (0.01)

L*	L*+H	L*+!H	H+!H*	X*?
432 (0.08)	34 (0.01)	4 (0.00)	41 (0.01)	141 (0.02)

Table 2.3: *Distribution of pitch accent types in BDC-spontaneous. Accent rate: 49.5%*

2.3 Boston Radio News Corpus

The Boston University Radio News Corpus (BURNC) [143] is a corpus of professionally read radio news data. This corpus was collected primarily to support research in text-to-speech synthesis, particularly the synthesis of prosodic patterns. The BURNC contains over seven hours of speech from seven speakers, three female and four male. Of the seven professional speakers, two normally read news live, while the remaining five normally pre-record and edit their stories. There are two portions of this corpus. The BURNC consists of broadcast news speech collected in two ways. The *Radio News* portion contains news

L-L%	L-H%	H-L%	!H-L%	H-H%	H-X%?	X-?X%?
642 (0.29)	710 (0.32)	675 (0.31)	96 (0.04)	57 (0.03)	1 (0.00)	13 (0.01)

L-	H-	!H-	X-?
1836 (0.58)	1200 (0.38)	133 (0.04)	19 (0.01)

L%	H%	X%?
1412 (0.64)	767 (0.35)	13 (0.01)

Table 2.4: *Distributions of phrase ending tones (phrase accents and boundary tones) in BDC-spontaneous material. Mean intermediate phrase length: 3.73 words. Mean intonational phrase length: 5.32 words*

stories recorded in the WBUR radio station during broadcast. The *Lab News* material contains six speakers reading four stories and was recorded in a laboratory. We do not differentiate between these two sets of material in this thesis. One advantage of analyzing speech produced by professional radio announcers is that the speech tends to be clear and free of disfluencies. However, Bolinger [24] highlights some oddities of broadcast news speech, including the frequent accenting of discourse given words, and deaccenting of discourse new words. These observations suggest that broadcast news speech may be substantially different from more naturally occurring speech. A 2.35 hour, 29,578 word, subset from six speakers (three female and three male) has been prosodically labeled using the ToBI standard. Distributions of ToBI tones present in the BURNC material are described in Tables 2.5 and 2.6

H*	!H*	L+H*	L+!H*
7713 (0.47)	2284 (0.14)	2433 (0.15)	658 (0.04)

L*	L*+H	L*+!H	H+!H*	X*?
524 (0.3)	44 (0.00)	4 (0.00)	624 (0.04)	1957 (0.12)

Table 2.5: *Distribution of pitch accent types in BURNC. Accent rate: 54.7%*

L-L%	L-H%	H-L%	!H-L%	H-H%
3130 (0.55)	2139 (0.38)	199 (0.04)	38 (0.01)	67 (0.01)

X-?X%?	H-X%?	!H-X%?	L-X%?	X-?H%	X-?L%
52 (0.01)	9 (0.00)	7 (0.00)	14 (0.00)	3 (0.00)	8 (0.00)

L-	H-	!H-	X-?
6239 (0.75)	1169 (0.14)	857 (0.10)	99 (0.01)

L%	H%	X%?
3375 (0.60)	2209 (0.39)	82 (0.01)

Table 2.6: *Distributions of phrase ending tones (phrase accents and boundary tones) in BURNC material. Mean intermediate phrase length: 3.53 words. Mean intonational phrase length: 5.22 words*

Only a small subset of the prosodically annotated material in BURNC includes manually defined orthographic boundaries. However, all of the material contains the output of a forced-aligner. Forced alignment is a process in which a transcription of speech is time-aligned to a corresponding speech signal. The alignment of manual ToBI annotations with forced-alignment word boundaries occasionally introduces some problems. Where the location of break indices and word boundaries do not always align perfectly, we linearly assign break indices to word boundaries. That is, the *n*th break index is assigned to the *n*th word boundary, regardless of the time of either. In a few instances, the number of break indices do not equal the number of word boundaries. This is frequently due to the forced alignment routine treating compound words, such as “school-based” as a single word, while the ToBI labeler treated it as two, or vice versa. In these instances, a break index of ‘1’ was inserted, or the additional break was omitted, by hand. When the source of the error could not be determined, the file was omitted from the material.

In addition to the forced-alignment word and phone boundaries, and ToBI annotation, the BURNC also includes a lexicon which contains syllabification information for each lexical item spoken in the corpus. Unfortunately, the phonetic inventory used in the lexicon does not correspond to the phonetic inventory in the forced-alignment output. Therefore, to use this syllabification information, we align the syllable boundaries from the lexicon phone sequence to the phone sequence contained in the forced-alignment material. We use a minimal edit distance, dynamic programming, routine to perform the alignment between these two phone sequences. To ensure that the syllable nuclei (vowels) of the two sequences are aligned we set the substitution cost for any two vowels to zero. This technique generated 48,253 syllables, 16,781, 34.8%, of which bear accent. This approach is largely successful, though not without errors. 144 of the syllables created this way contain two forced-alignment vowels and 8 contain none. Moreover, 25 syllables contain two accented vowels.

2.4 TDT-4 Corpus

The TDT4 corpus [193] includes newswire text and broadcast news audio in English, Mandarin and Arabic. The TDT4 audio corpus includes 312.5 hours of English BN from 450 shows, 88.5 hours of Arabic BN from 109 shows and 134 hours of Mandarin BN from 205 programs. This material was drawn from six English news shows – ABC “World News Tonight”. CNN “Headline News”. NBC “Nightly News”, Public Radio International “The World”. MS-NBC “News with Brian Williams”, and Voice of America, English – three Mandarin newscasts – China National Radio, China Television Systems and Voice of America, Mandarin Chinese – and two Arabic newscasts – Nile TV and Voice of America, Modern Standard Arabic. All shows were aired between October 1, 2000 and January 31, 2001. In addition to the raw audio signal for each BN document, as part of the NIGHTINGALE team within the DARPA GALE program, Columbia University has access to a number of automatically produced annotations of the audio material, including automatic speech recognition transcripts with word boundaries [190] and inter-word durations, hypothesized sentence boundaries with confidence scores [124], and speaker segmentation (DIARIZATION) hypotheses [233].

For the purposes of prosodic event detection, one expert ToBI labeler annotated one 30 minute news broadcast, 20010131_1830_1900_ABC_WNT, for pitch accent presence and intonational phrase boundary locations. No classification of pitch accent type or boundary tones was annotated. The automatic annotations, including high-quality automatic transcripts, provided by our NIGHTINGALE collaborators of the TDT-4 material allow experiments on pitch accent detection in a completely automatic setting, with no human annotations. This show is 30 minutes long, and contains 3,326 words. 1658, 49.85%, of these words are accented, and the mean intonational phrase length is 5.98 words. Note, this material is only used in the pitch accent detection task (cf. Chapter 3).

Chapter 3

Pitch Accent Detection

3.1 Introduction

Accenting, or intonational prominence, has the effect of drawing a listener's attention to a particular section of a spoken utterance. Consider the following pair of utterances, where the words in small caps are accented.

My CLASS had a TEST yesterday. The students performed HORRIBLY.

The accenting of "class", "test" and "horribly" have the effect of highlighting the most salient information in these two utterances.

This accenting behavior is typically associated with an acoustic highlighting of a word, but the effect can be broader, drawing attention to the importance of a larger phrase. Take the following example, assumed to come after the two preceding utterances.

A student in the BACK fell ASLEEP in the MIDDLE of the exam.

The accenting of “back” may draw a listener’s attention to the location of the student or indicate prominence of the whole noun phrase (NP), “A student in the back”. The communicative effect of the accentuation of “back” is context dependent. If the previous utterance were discussing another student, the effect would likely be to draw contrast by highlighting the location of the subject of this utterance. On the other hand, without this context, accenting “back” serves to draw focus more broadly to the whole NP [83].

Fundamentally, accenting is an acoustic highlighting of a word through some modification of its associated speech signal. The effect of accenting, however, is not so simply understood. One theory posits that content words (e. g. nouns, verbs, adjectives, etc.) introducing “new” information are accented, while terms conveying “given” or already established information are not accented, or deaccented [70, 47]. This is something of an oversimplification, with empirical studies finding the existence of acoustically prominent “given” terms and non-prominent “new” terms [210, 68]. Though not complete in its description of accenting behavior, this theory does have *some* explanatory power; “new” content words (nouns in particular) are *often*, if not always, accented, and “given” nouns are *more* likely not to bear accent.

One limitation of any theory of how accenting impacts the interpretation of an utterance, comes from the fact that accenting is used for multiple purposes. In addition to correlations with information status [70], contrast is used to indicate the focus [72] and topic [76] of an utterance. These three may suggest different accenting behavior for the same utterance. Consider the following utterance in our series of examples.

This CLASS is the WORST in the SCHOOL.

If information status were the only force impacting accenting behavior, “class” would not be accented – it was introduced to the discourse in the first utterance. However, the class is the topic and focus of the utterance. To make this clear to the listener “class” is accented.

Accenting is also commonly used to highlight contrast, as in the following example.

YOUR class is a DREAM!

"Your" is accented to draw a contrast between the speaker's and the listener's class. While the entity referred to by "your class" is new to the discourse, and the topic of the utterance, "Your" is accented, rather than "class" to highlight the contrast between the two classes.

There are also instances in which accenting more clearly affects meaning. Consider the following set of utterances.

1. The math department offers classes in geometry, algebra and trigonometry.
2A. Most sophomores have only HEARD of algebra. (They haven't studied it.)
2B. Most sophomores have ONLY heard of ALGEBRA. (They've never heard of geometry or trigonometry.)

The interpretation of utterances 2A and 2B significantly differ. Utterance 2A implies nothing about the students' relationship to geometry or trigonometry. On the other hand, 2B contrasts their familiarity with algebra with a lack of awareness of the other material. While the complete breadth of communicative uses of accenting is not fully understood, there is consensus that accenting is an integral component of human speech.

Acoustically, an accented word stands out from its surroundings. Words bear accent if they contain an acoustic excursion away from a neutral voice [109, 25]. In Standard American English (SAE), accents are commonly associated with a pitch excursion, which has led them to be called "pitch accents". Until recently, these pitch excursions were largely taken to be the most perceptually important cue to pitch accent. Clark and Yallup [42] described pitch as "the most salient determinant of prominence." While Wightman and Ostendorf [230] found "[l]ittle discussion of energy cues...in the linguistics literature, however, probably because energy is less important than F0 and duration in human perception of prominence." However, recent studies have given more attention to the correlation between speech energy and pitch accent. Silipo [182] found that in spontaneous speech, duration and energy to be the most important acoustic parameters underlying accent with pitch playing only a minor role. Kochanski [103] expanded on this result with a lengthy analysis-by-classification of British and Irish English, finding f0 to be a weak predictor of prominence, with loudness and duration being much more discriminative of prominent and non-prominent syllables. Spectral tilt is a less studied acoustic quality that will be explored heavily in this chapter. In a pair of perception [186] and production [185] experiments, Sluijter and van Heuven showed that accent in Dutch strongly correlates with the energy within a particular frequency subband, that greater than 500Hz. This observation concerning "spectral emphasis" led to a number of other studies examining relationship between "spectral balance" or "spectral tilt" and pitch accent [80, 78, 53, 32], finding similar strong correlations. The terms, "spectral emphasis", "spectral tilt", "spectral balance" and "high-frequency emphasis", though they may be calculated in different ways, all extract energy information from a particular frequency subband and compare this energy to the total signal energy.

In this chapter we will explore a number of techniques to automatically detect pitch accents in Standard American English speech. The experiments will focus on approaches to acoustic based detection. Section 3.3 will establish a baseline by examining the automatic detection performance of pitch, energy and duration features and address the use of context in accent detection. Previous efforts to detect pitch accents are divided by their decision to detect accents on words, or syllables. In Section 3.4, we attempt to evaluate the impact of this decision, examining the performance of syllable- and word-based pitch accent detection. A close examination of energy features that may be useful for pitch accent detection will be described in Section 3.5. We present techniques using pitch and duration features in combination with the findings of these spectral experiments to detect pitch accents in Section 3.6. In Section 3.7, we assess the use of part-of-speech features in pitch accent detection.

We conclude in Section 3.8.

3.2 Related Work

Words are made prominent in their surrounding content by containing acoustic excursions away from an otherwise neutral voice [109, 25]. Pitch excursions, in particular, have beenlargely taken to be the most salient cue to pitch accent. Clark and Yallup [42] described pitch as "the most salient determinant of prominence." While Wightman and Ostendorf [230] found "[l]ittle discussion of energy cues...in the linguistics literature, however, probably because energy is less important than F0 and duration in human perception of prominence." However, recent studies has given more attention to the correlation between speech energy and pitch accent [32, 185, 186, 80, 182, 53, 78, 103]. We should note here that human agreement on the pitch accent detection task falls somewhere between 81% and 91% word-accuracy [156, 196], depending on the experience of the labelers and the genre of the material. Also, recall that a majority class baseline on the BURNC corpus yields 54.7% pitch accent detection accuracy, by hypothesizing that all words are accented. This section describes a substantial amount of previous work on the task of pitch accent detection, results of this related work is summarized at the end of this section in Table 3.1.

The task of automatically detecting pitch accent has received a considerable amount of research attention, supported by these and other studies of acoustic correlates to accenting. A wide range of supervised machine learning techniques have been applied to this problem. Chen and Withgott [37] experimented with using a supervised Hidden Markov Model (HMM) trained with smoothed pitch and intensity features to detect emphasis. This work used these emphasis hypotheses to improve summarization of a spontaneous two-party discourse. The results of the emphasis detection are presented as an ROC curve, making comparisons to other studies very difficult. That said, this approach is the first of many techniques to apply HMMs and other short-frame based models to this task.

For example, Conkie, et al. [44] used a HMM to detect pitch accent using speaker normalized pitch and energy values extracted at 10ms frames, with delta and delta deltas of these values. This approach was evaluated on the speech of a single professional speaker, reading three types of text: prompts, newspaper text, and phonetically balanced utterances. This acoustic HMM achieved 82.8% accuracy. A syntactic-prosodic model, associating part-of-speech sequences with prosodic labels, correctly predicted 84.0% of the accents, and in combination, the two performed at 88.3% accuracy. This represents one of the better results in detecting pitch accents using speaker dependent acoustic modeling.

Sequential modeling of acoustic modeling extracted at 10ms frames has been incredibly successful in automatic speech recognition (ASR) systems. It is, therefore, unsurprising that this approach would be used in the analysis of intonation. Sequential models take a number of forms. A distributed Time Delay Recursive Neural Network (TDRNN) was described by Ren [167]. The TDRNN is a different style of sequential modeling from the HMM, where the output of a Neural Network (NN) from one time step is used to provide features in the input layer of itself at a later time step. This distributed TDRNN incorporated four distinct TDRNN models, one for each feature type: pitch, intensity, duration, and filtered energy coefficients. The filtered energy coefficients were based on wavelet transformation of 14 frequency-band filters of the energy in the speech signal. A discrete cosine transform then converted these to a discrete set of coefficients, which the author's dubbed "spectral balance based cepstral coefficients" or SBCC. A fifth TDRNN model was used to combine the output of the independent TDRNN models. This distributed approach was able to detect pitch accents on syllables using only acoustic information with 83.64% accuracy, using three female speakers from the Boston University Radio News Corpus (BURNC) as data. When this detection was performed using a single TDRNN instead of four distributed models with a combining model, the performance was 2.43% worse.

Taking a similar approach, Ananthakrishnan, et al. [3] applied a coupled multistream HMM (CHMM) to the task of automatic pitch accent detection. Using this modeling technique, the HMM used inputs from the pitch, energy and duration domains independently, training a coupled model to combine the information. This work presented its results in terms of both syllable- and word-based accuracy. Evaluated on a single BURNC speaker, and using only acoustic information, the CHMM model performed with 72.03% word-based accuracy (73.97% syllable-based). In combination with a syntactic language model its word-based accuracy is improved to 79.5%.

Chaolei, et al. [35] developed a technique for pitch accent detection concurrent with phone recognition for language learning. While others have applied the same techniques for ASR and pitch accent detection, this work couples the two tasks, performing simultaneous accent detection and phone recognition. This technique used normalized pitch values along with Mel-Frequency Cepstral Coefficients (MFCC), typically used in automatic speech recognition (ASR), to recognize phones in accent-bearing and non-accent bearing forms. Using a standard HMM and evaluated on 7 BURNC speakers, this model was able to predict accents on syllables with 80.6% accuracy. For the accent detection evaluation, phone errors were ignored from the analysis. If an accent-bearing form of a phone was correctly detected, the detection was considered correct, whether or not the phone identity was accurate.

Evaluating the differences in short-frame HMM modeling, and maximum entropy modeling, Rangarajan et al. [188, 189, 187] applied both supervised detection approaches using acoustic and syntactic features. The acoustic features were f0 and RMS energy with delta and delta deltas. An HMM was evaluated on 4 BURNC speakers and correctly identified 70.58% of accents. The corresponding maximum entropy model achieved 80.09% word level accuracy. This paper explored the use of supertags [9], modeling uni-, bi- and trigrams as well as punctuation based features. The inclusion of these syntactic features, dramatically improves the performance of both the HMM and maximum entropy acoustic models to 85.13% and 84.53% accuracy respectively. While these results are impressive, there is one major caveat to the use of lexico-syntactic features for prosodic detection and classification on the BURNC material. This corpus is comprised of a professional news speakers each reading the same news stories. When performing a speaker-independent evaluation, the training and testing sets necessarily contain overlapping, if not identical, lexical information. While accent and phrasing decisions are not completely determined by lexical information [23], neither are they are independent from it. Therefore, when training lexico-syntactic models on BURNC data, it is critical to guarantee that any story appearing in the training data, does not also appear in the testing data, regardless of the speaker. Without this guarantee, it is quite likely that the syntactic results will be artificially inflated due to the consistency between train and test material. The approaches described in [189, 187, 4, 81] may have suffered from this issue. These papers do not directly address this potential problem, making it unclear to assess its impact.

While a significant number of approaches have used short frame (10ms) acoustic features for accent detection, others have extracted aggregated features of acoustic information over syllables or words. These features are then used to train supervised machine learning classifiers to detect which syllables or words bear accent. The work presented in this chapter falls into this class of automatic pitch accent detection approaches.

One of the earliest efforts at automatic pitch accent detection was performed by Wightman and Ostendorf [230]. They used decision trees to simultaneously predict accented syllables and phrase boundaries. The decision tree outputs were then used as input to an HMM to model the likelihood of observing a given sequence of prosodic labels. This was evaluated on a single speaker from the BURNC, and demonstrated 81.51% pitch accent detection accuracy. The decision tree model used pitch features – aggregations within the syllable, and differences from surrounding syllables – as well as syllable structure – position within a word, lexical stress – and the presence of a following pause or breath. Ostendorf and Ross [144] later proposed a stochastic modeling framework to simultaneously predict accent and boundary locations. The proposed input to this structure was pitch, duration, and energy features and segmental characteristics of the syllable sequence. Optimizing this joint modeling technique on material from a single BURNC speaker, Ostendorf and

Ross were able to detect pitch accents with 89% syllable-based accuracy. In this modeling, syllables were classified as unaccented, high, low or downstepped, while the 89% detection accuracy was derived by a *post hoc* collapsing of high, low and downstepped classes. This is an optimistic assessment of the automatic performance of the approach as it assumes knowledge of the presence of intermediate phrase boundaries. As noted in Chapter 4, the automatic detection of intermediate phrase boundaries is a very difficult task.

Ensemble learning techniques were explored by Sun [194]. Using boosting and bagging with CART (Classification And Regression Tree) models, high accuracy pitch accent detection was achieved using acoustic and syntactic features. This approach, like [144] classifies pitch accents as unaccented, high, low or downstepped. However, the publication includes confusion matrices from the experiments, enabling the construction of binary detection performance. Using acoustic features a syllable level accuracy of 89.90 was obtained on a single BURNC speaker (f2b). To the best of our knowledge this represents the highest published accuracy on pitch accent detection of a single speaker using only acoustic information. Using lexico-syntactic features, such as vowel identity, syllable stress of current and surrounding syllables, and part of speech, the bagged CART models performed with 86.99% accuracy – worse than the acoustic modeling. However, when combined they achieved a syllable level accuracy of 92.78%.

In general, words are considered to bear accent or not, however, some researchers have made finer distinctions. Emphatic pitch accents were defined as pitch accents that were perceived as more prominent than normal pitch accents, adding a third level to the level of prominence in [29]. It is unclear how consistently this distinction is labeled, but approximately 40% of pitch accents were considered *emphatic* in their experimental material – "child directed" stories read by a single female speaker. Brenier et al., however, used pitch, energy and duration features to detect emphatic pitch accents, noting that pitch *range* – defined by the height of the maximum pitch value – is particularly useful in distinguishing regular pitch accents from emphatic ones. Aggregated features were extracted from the current, previous and following words, and normalized by the surrounding intonational phrase. Lexical features – TF*IDF, whether the word was in an exclamation or negation, and word class information – were also explored. While this work was intended to detect emphatic accents, performance in detecting all accents was also reported. The acoustic and lexical feature sets performed comparably, detecting pitch accent with 78.2% and 78.4% word level accuracy, respectively. Combining them led to a significant improvement, up to 84.4% word level accuracy. The classifier was able to detect emphatic pitch accents with 87.8% accuracy using both acoustic and lexico-syntactic features, over a 79.8% baseline.

Manual annotation of prosody is very resource intensive. It can take an expert labeler up to 100-200 times real time to perform a full ToBI annotation [195]. To avoid the resource requirements of manual annotation of training data for supervised learning, a number of researchers have looked into unsupervised and semi-supervised learning techniques for pitch accent detection. Ananthakrishnan, et al. [4] examined the use of K-means, Fuzzy K-means and GMM clustering for pitch accent detection. Using intensity-, pitch-, duration-, pause-based acoustic features, and lexico-syntactic features capturing syllable identity, part of speech and lexical stress, the authors were able to detect accents with accuracy up to 77.8%. This technique was evaluated using 7 BURNC speakers, and potentially suffers from the pitfall of using BURNC syntactic features in speaker-independent evaluations. Tamburini [197, 199] defined a function, called the *Prom* function, which is a linear combination of energy, duration, pitch and spectral tilt features. An unsupervised thresholding of the value of this function was able to detect pitch accents with 80.2% accuracy at the syllable level. Extending this to a supervised approach, manually annotated data was used to learn weights for each of the terms in the function; this improved accuracy to 82.5%. Levow [116] evaluated the use of a semi-supervised technique, Manifold Regularization [17] in Laplacean Support Vector Machines, and a spectral clustering technique, asymmetric clustering [54] for the task of automatic pitch accent classification. Both of these techniques were applied using feature vectors containing only acoustic information. The maximum and mean of pitch and intensity was extracted from the current and surrounding syllable nuclei, as was the slope of the pitch. Differences between these features and those extracted from surrounding syllables was also included in the feature representation. Laplacean SVM, the semi-supervised technique, classified pitch accents with 81.5% syllable-based accuracy, while the unsupervised technique, asymmetric clustering, achieved 78.4% accuracy. Similar to [194] and [144], this technique classified pitch accents into four classes – unaccented, high, low and downstep. The binary, accented versus unaccented, accuracy, however, cannot be determined from the error reporting described in the paper. Braunschweiler [27, 28] developed a module dubbed the "Prosodizer" that uses manually written scoring rules over pitch energy duration and part of speech features to detect pitch accent. These rules, despite being designed by hand, performed remarkably well. They were able to recognize pitch accents at the syllable level with 81% accuracy on a single BURNC speaker.

The above approaches all used acoustic information to detect pitch accents, though some also employed lexico-syntactic features. The task of identifying pitch accent locations using only text based features is clearly related to acoustic accent detection, though it typically has a somewhat distinct use. In general, when approaches use only text features, it is for prosodic assignment, as opposed to prosodic analysis. Prosodic assignment is the task of describing a plausible intonation for a piece of text. This is frequently used in text-to-speech (TTS) systems, to assign the intended intonation for the synthesized speech.

From a more theoretical point of view, Chomsky posited that accent is entirely predictable from the surface of an utterance [39]. Bolinger took an opposing point of view, famously retorting "Accent is predictable (if you're a mind reader)" [23]. There do seem to be multiple valid accent placements for a given utterance. The repetition of lexical content in the BURNC even within a standard speaking style does not reveal entirely predictable accent and phrase boundary placement. However, lexico-syntactic analysis is able to predict prosodic labels with accuracy well above chance. This suggests that, while not completely predicable, there is significant information in the lexical content of an utterance that informs which tokens are *likely* to be accented. The following automatic approaches take advantage of features extracted only from text to predict pitch accent locations.

Hirschberg [81] described a technique for pitch accent assignment using part-of-speech information, complex nominal analysis and surface position information. Also, a technique for modeling the information status, Given vs. New [162] was proposed. Following the discourse structure theories posited in [71], a FIFO queue is used to store word roots to represent the concepts in the Local Focus of the current discourse. While the attentional and intentional structures are not explicitly modeled in this approach, word roots serve to represent concepts that may not be strictly present in the linguistic structure of the discourse. Orthographic cues, punctuation and paragraph boundaries are used to determine how the queue is manipulated throughout the course of the analysis – with different signals indicating whether tokens should be added or removed from the queue representing local focus. Using these features to train a CART classifier, a cross validation accuracy on 3 speakers of BURNC material of 76.5% was achieved. Evaluations on Audix – read news – and ATIS – travel domain data – showed improved performance, 80% and 85%, respectively.

Ross and Ostendorf [173] applied an HMM over decision tree posteriors to detect pitch accents from text. Similar to [230], the decision tree models the likelihood of a single syllable or word being accented, while the HMM models any sequential phenomena concerning accent and phrase final tone sequences. This system used a multi-stage approach, first detecting pitch accent, then assigning pitch accent type, and finally assigning phrase boundary intonation. In their experiments intonational phrase boundary location is assumed to be given. Based on the text from a single BURNC speaker (f2b), part-of-speech, prosodic phrase structure, Given/New status (simplified, though akin to the representation in [81]), lexical stress information, paragraph structure were all extracted for training the decision tree model. While this approach was trained using syllables, the paper reports accuracy at both the syllable and word level, facilitating comparison to other approaches. This technique was able to predict pitch accent placement from text with 87.7% syllable-based accuracy, corresponding to 82.5% word accuracy.

Many approaches to prosodic assignment operate similarly to these two early papers. The common form is to apply some supervised learning algorithm to associate features derived from part-of-speech tags, syntactic chunks, or syntactic parse trees to pitch accent locations. The differences in these approaches come down to the machine learning algorithm used and the features derived. Hirschberg and Rambow [91] evaluated the application of more sophisticated syntactic features to the task of prosodic assignment. They applied the RIPPER rule induction machine learning algorithm using part-of-speech tags, supertags and parse tree derived features, as well as the length and relative position of the current sentence. This technique was evaluated on a single professional speaker – material collected for use in an AT&T speech synthesis application – and 88.2% word-based accuracy was reported. They found word and part-of-speech features to be sufficient for predicting pitch accents, while the parse tree features were more useful for phrase boundary prediction.

Marsi et al. [128] compared the performance of CART modeling in pitch accent assignment to the performance to Memory Based Learning (MBL). The lexico-syntactic features that were examined included the word identity, punctuation, part of speech, NP and VP chunk location, Information Content (IC), TG*IDF, and distance to the previous occurrence of word and the sentence boundary. This was evaluated on 201 articles from the ILK corpus, a collection of Dutch newspaper text. While the performance of both of the modeling techniques are high, f-measure of 82.0 for CART and 83.6 for MBL, without corresponding accuracy information, it is difficult to assess the relative strength of these features and modeling techniques to others.

Gregory and Altun [69] applied Conditional Random Field (CRF) modeling to the task of accent assignment. Part-of-speech tags were collapsed to broad classes of Function, Verb, Noun and Other. Based on these broad features, probabilistic variables were extracted, representing the log probability of a word being accented given part-of-speech information. Unigram, bigram, reverse bigram, joint and reverse joint log probabilities were all used in the model, as were the number of phones and syllables in a word and the position in the utterance. Using a CRF with a window size of 5, 76.36% word-based accuracy was reported. This evaluation was performed on manually annotated Switchboard corpus data (telephone conversation) with male and female speakers with a variety of American English dialects. It is unclear from this paper how many speakers are represented in their prosodically annotated data but the whole Switchboard corpus contains over 500 speakers [64]. We will examine prosodic assignment techniques again in Chapter 4. Similar approaches have been applied with considerable success to the task of phrase boundary assignment in addition to the accent assignment approaches discussed here.

Recently, Nenkova, et al. [139] identified a simple lexical attribute which is remarkably successful in detecting pitch accent type given its simplicity. They define a term called *accent ratio*, capturing the accent rate of a given word. $AccentRatio(w) = \frac{k}{n}$ if $B(k, n, 0.5) \leq 0.005$ and 0.5s otherwise, where k is the number of times word w appears accented in the training data, n is the total number of times word w appeared. This feature yields 75.59% accent detection accuracy on Switchboard [64] data; inclusion of other lexical features increased accuracy to 76.65%.

In a slightly different application, Hasegawa-Johnson et al. [74] found that prosodic dependent acoustic and language models can be applied to improve ASR performance. They report an improved word error rate on the BURNC corpus from 24.8% using prosody independent models to 21.7% using prosody-dependent models. This approach used a prosodic assignment technique in the construction of its prosody-dependent language model. Syntactic parse tree and part-of-speech features from a range of windows were used to train an artificial neural network (ANN) model to assign pitch accents and phrase boundaries [43]. Evaluated on six speakers from the BURNC, this approach was able to predict pitch accents with 83.1% accuracy.

This section contains descriptions of many approaches to automatic pitch accent detection. To facilitate comparison to previous approaches and results, Table 3.1 contains a summary of the performance of the approaches described.

Paper	Features	# Speakers / Corpus	Domain	Model	Accuracy
Hirschberg 1993 [81]	**L**	3 / BURNC	Word	CART	76.5%
Hirschberg 1993 [81]	**L**	1 / Audix	Word	CART	80%
Hirschberg 1993 [81]	**L**	26 / ATIS	Word	CART	85%
Wightman 1994 [230]	**A+L**	1/ BURNC	Syl.	Decision Tree +HMM	81.51%
Ross 1996 [173]	**L**	1 / BURNC	Syl.	CART+HMM	87.7%
Ross 1996 [173]	**L**	1 / BURNC	Word	CART+HMM	82.5%
Ostendorf 1997[144]	**A+L+P**	1 / BURNC	Syl.	Stochastic Modeling	89%
Conkie 1999 [44]	**A**	1 / TTS & BN	Word	HMM	82.8%
Conkie 1999 [44]	**L**	1 / TTS & BN	Word	HMM	84.0%
Conkie 1999 [44]	**A+L**	1 / TTS & BN	Word	HMM	88.3%
Hirschberg 2001 [91]	**L**	1 / WSJ	Word	Ripper	88.2%
Sun 2002 [194]	**A**	1 / BURNC	Syl.	AdaBoost	89.90%
Sun 2002 [194]	**L**	1 / BURNC	Syl.	Bagging	86.99%
Sun 2002 [194]	**A+L**	1 / BURNC	Syl.	AdaBoost	92.78%
Tamburini 2003 [197]	**A**	1 / TIMIT	Syl.	*Prom* function	80.1%
Ren 2004 [167]	**A**	3 / BURNC	Syl.	TDRNN	83.64%
Cohen 2004 [43]	**L**	6 / BURNC	Word	ANN	83.1%
Gregory 2004 [69]	**L**	? / Switchboard	Word	CRF	76.36%
Brenier 2005 [29]	**A+P**	1 / NA	Word	Maxent	78.2%
Brenier 2005 [29]	**L**	1 / NA	Word	Maxent	78.4%
Brenier 2005 [29]	**A+L+P**	1 / NA	Word	Maxent	84.4%
Ananthakrishnan 2005 [3]	**A**	1 / BURNC	Syl.	CHMM	73.97%
Ananthakrishnan 2005 [3]	**A+L**	1 / BURNC	Syl.	CHMM	74.84%
Ananthakrishnan 2005 [3]	**A**	1 / BURNC	Word	CHMM	72.03%
Ananthakrishnan 2005 [3]	**L**	1 / BURNC	Word	CHMM	79.70%
Ananthakrishnan 2005 [3]	**A+L**	1 / BURNC	Word	CHMM	79.50%
Tamburini 2005 [199]	**A**	1 / TIMIT	Syl.	Weighted *Prom* function	82.5%
Braunschweiler 2006 [28]	**A+L**	1 / BURNC	Syl.	Hand crafted scoring rules	81%
Levow 2006 [116]	**A**	1 / BURNC	Syl	Asymmetrical clustering	78.4%*
Levow 2006 [116]	**A**	1 / BURNC	Syl	Laplacian SVM	81.5%*
Ananthakrishnan 2006 [4]	**A+L**	7 / BURNC	Syl.	K-means	77.8%
Ananthakrishnan 2006 [4]	**A+L**	7 / BURNC	Syl.	Fuzzy K-means	77.5%
Ananthakrishnan 2006 [4]	**A+L**	7 / BURNC	Syl.	GMM	77.8%
Chaolei 2007 [35]	**A**	7 / BURNC	Syl.	HMM	80.6%
Nenkova 2007 [139]	**L**	? / Switchboard	Word	J48	76.65%
Sridhar 2008 [187]	**A**	4 / BDC	Word	Maxent	74.51%
Sridhar 2008 [187]	**A+L**	4 / BDC	Word	Maxent	78.64%
Sridhar 2008 [187]	**A**	4 / BURNC	Word	Maxent	80.09%
Sridhar 2008 [187]	**A+L**	4 / BURNC	Word	Maxent	84.53%
Sridhar 2008 [187]	**A**	4 / BURNC	Word	HMM	70.58%
Sridhar 2008 [187]	**A+L**	4 / BURNC	Word	HMM	85.13%

Table 3.1: *Summary of Pitch Accent Detection Performance.* ***A*** *: Acoustic,* ***L*** *: Lexical,* ***P*** *: Phrasing*

3.3 Acoustic Detection

There is consensus and empirical confirmation that pitch, energy, and duration are major acoustic correlates of pitch accent. In this section, we evaluate the use of these acoustic features for the automatic detection of pitch accent. The experiments presented in this section are supervised machine learning approaches, trained and evaluated on BDC-read, BDC-spontaneous and BURNC material. These corpora are described thoroughly in Chapter 2. We treat pitch accent detection as a binary classification problem, classifying words as accent bearing or non-accent bearing. In this section, we limit our classification to the word level, while in Section 3.4, we examine the decision to detect pitch accents on words as opposed to syllables or syllable nuclei.

To establish baseline acoustic accent detection performance, we construct distinct feature sets using pitch, energy, duration and voice quality features. For the pitch feature set, we extract the minimum, maximum, mean, standard deviation, and z-score of the maximum pitch within the word from the raw and z-score speaker normalized pitch contours. The energy feature set extracts the same features from the raw and speaker normalized energy contours, as opposed to the pitch contour. The duration feature set consists of a single feature: the duration, in seconds, of the word. We evaluate these feature sets on each of the three corpora, BDC-read, BDC-spon and BURNC, using ten-fold cross-validation and J48 decision trees (a java implementation of Quinlan's C4.5 algorithm [164] distributed with the weka machine learning toolkit [232]), Logistic Regression [113] and sequential minimal optimization (SMO) trained Support Vector Machines with linear kernels [157]. Results from these experiments can be found in Table 3.2.

Across all corpora and classification techniques, we find energy features to be the most predictive of pitch accent. Duration features, in general perform with 2% less absolute accuracy. The relative performance of pitch features show a genre bias. On BURNC material, professional broadcast news speech, pitch features perform with accuracy approximately 10% lower than energy features. On BDC-read material, this relationship is less pronounced,

	J48	SVM	Lostistic
BDC-read			
Pitch	77.26 ± 0.754	77.13 ± 0.754	77.46 ± 0.558
Energy	79.13 ± 0.640	80.06 ± 0.459	80.08 ± 0.574
Duration	77.45 ± 0.426	76.24 ± 0.623	75.72 ± 0.738
All	78.73 ± 0.738	81.07 ± 0.394	81.16 ± 0.640
BDC-spon			
Pitch	75.48 ± 0.525	75.77 ± 0.705	75.78 ± 0.590
Energy	79.33 ± 0.410	80.59 ± 0.590	80.50 ± 0.508
Duration	78.30 ± 0.754	77.43 ± 0.492	77.23 ± 0.640
All	78.68 ± 0.558	80.60 ± 0.476	80.41 ± 0.689
BURNC			
Pitch	73.50 ± 0.394	72.43 ± 0.328	73.50 ± 0.262
Energy	83.00 ± 0.361	82.91 ± 0.262	82.50 ± 0.410
Duration	81.94 ± 0.262	80.89 ± 0.394	79.73 ± 0.310
All	83.00 ± 0.410	83.37 ± 0.279	83.35 ± 0.492

Table 3.2: *Classification Accuracy using acoustic feature sets with J48, Logistic Regression and SVM classification.*

with pitch features only 5% worse than energy features. On the BDC-spontaneous corpus, the pitch features perform as well or better than the duration features, roughly 2% worse than the energy feature set. This suggests that pitch is a stronger predictor of pitch accent in spontaneous speech than read speech. These results support those of Kochanski et al. [103] and Silipo and Greenberg [182] which also find that pitch is a weak predictor of pitch accent.

While the energy feature set is able to generate the best performance in isolation, the inclusion of pitch and duration features improves the overall detection accuracy on BDC-read and BURNC corpora. This indicates that these feature sets do not capture identical or redundant information.

The SVM and Logistic Regression classifiers perform as well or better than the J48 decision tree classification on this task. Moreover, they never differ significantly in their performance. As Logistic Regression models train notably faster than support vector machines, we favor this classification technique. Decision trees train still faster than Logistic Regression models, and have the advantage that their learned decision process is easily interpretable; the tree structure can be trivially read and interpreted as a series of if-then statements. We continue to explore J48 decision trees in this chapter, particularly in Section 3.6, but note that in general, Logistic Regression classification generates superior performance.

Accents are regions of speech that are perceived to stand out from their surroundings. This prominence relative to the surrounding speech material suggest that acoustic features which capture contextual acoustic information should be able to improve pitch accent detection. To test this hypothesis, we extend each of the pitch, energy, and duration feature sets with context based features. For the pitch and energy features, the mean and maximum value within the word is z-score normalized using the mean and standard deviation from a number of context regions surrounding the current word. In the duration feature set, the value of the duration is z-score normalized[1] relative to the duration of the words comprising the context region. For these normalizations, eight context-regions are defined: 1) one previous word, 2) one following word, 3) two previous words, 4) two following words, 5) one previous and one following word, 6) two previous and one following word, 7) one previous and two following words and 8) two previous and two following words. We repeat the evaluations in Table 3.2 with the context extended feature sets, result from these experiments are reported in Table 3.3.

We find the inclusion of context normalized acoustic features to significantly improve automatic pitch accent detection performance of SVM and Logistic Regression classifiers using a t-test. Using SVM and Logisitic Regression classifiers for detection, we see a similar relative strength of the feature sets as observed in the context-free experiments reported in Table 3.2. Namely, energy features yield the best performance in isolation, but combine with pitch and duration features to improved overall accuracy. All three feature sets show improved performance with the addition of context features. The use of context in pitch accent detection is well motivated by intonational theories – accented words are

[1] Z-score normalization is computed as $\frac{x-\mu}{\sigma}$, where x is a value to normalize, μ and σ are a mean and standard deviation of values to normalize by – e.g. all pitch values for a given speaker, or all pitch values within a particular region in time.

	J48	SVM	Logistic
BDC-read			
Pitch	77.01 ± 0.918	80.04 ± 0.590	79.96 ± 0.344
Energy	78.23 ± 0.640	81.76 ± 0.394	81.49 ± 0.230
Duration	78.07 ± 0.558	76.83 ± 0.443	77.21 ± 0.689
All	78.82 ± 0.607	83.44 ± 0.541	83.45 ± 0.410
BDC-spon			
Pitch	74.60 ± 0.640	77.85 ± 0.672	77.80 ± 0.804
Energy	78.75 ± 0.640	81.89 ± 0.574	81.58 ± 0.312
Duration	79.28 ± 0.558	78.71 ± 0.705	78.81 ± 0.836
All	78.36 ± 0.558	82.90 ± 0.509	82.88 ± 0.394
BURNC			
Pitch	73.26 ± 0.443	74.48 ± 0.361	75.19 ± 0.328
Energy	82.12 ± 0.295	83.51 ± 0.426	83.21 ± 0.476
Duration	82.13 ± 0.394	80.09 ± 0.541	80.41 ± 0.394
All	82.32 ± 0.344	85.08 ± 0.295	85.01 ± 0.443

Table 3.3: *Classification Accuracy using acoustic feature sets with context features with J48, Logistic Regression and SVM classification.*

acoustically prominent. Prominence implies a differentiation from the norm; representing the surrounding acoustic material in the feature representations serves to capture this difference from the norm. The value of context in pitch accent detection was also examined by Levow [115]. In a four way classification of pitch accent (unaccented, high, low, downstepped) from a single BURNC speaker (f2b), she found a syllable level improvement from 75.9% to 81.3% accuracy by incorporating acoustic features calculated relative to the left and right adjacent syllables. We do not observe so dramatic an improvement, but the overall result is confirmed: context is important to pitch accent detection.

Logistic Regression is able to detect pitch accent with the highest accuracy across corpora and feature sets. To evaluate the robustness to speaker differences, we evaluate the three feature sets with Logistic Regression detection of pitch accent using leave-one-speaker-out cross validation. This evaluates the speaker independence of these features. Results from this evaluation can be found in Table 3.4. In the ten-fold cross-validation evaluation, material from the same speaker appears in both training and testing folds. In the leave-one-speaker-out cross-validation evaluation, reported in Table 3.4, each test set

	BDC-read	BDC-spon	BURNC
Pitch	79.12 ± 2.034	76.77 ± 2.066	81.78 ± 1.082
Energy	75.97 ± 3.231	79.94 ± 3.756	82.62 ± 1.132
Duration	76.80 ± 2.673	78.44 ± 3.575	80.50 ± 0.804
All	78.96 ± 1.312	81.34 ± 2.903	84.26 ± 0.722

Table 3.4: *Classification Accuracy using acoustic feature sets with context features with Logistic Regression and leave-one-speaker-out cross-validation.*

contains the speech of a single speaker, while the training sets contain speech from all other speakers. This speaker independent evaluation should result in reduced performance for two reasons. First, if there are any speaker dependent effects in the production of accented words, this evaluation will not be able to learn them. Second, the amount of available training data is reduced. For example, in each of the BDC corpora, the ten-fold cross-validation sets each contain 90% of the full corpus, while in the speaker-independent evaluation the training sets comprise 75% of the material at each fold. Due to these two effects, we find reduced pitch accent detection performance on all corpora. However, the reduction in performance is not statistically significant on the BDC-spon (t-test $p = 0.411$)or BURNC material ($p = 0.180$). On the other hand, on BDC-read the reduction of 4.49% accuracy is quite significant ($p = 0.000458$). On this material, the performance of the pitch and duration feature sets do not significantly differ when evaluated under speaker dependent or independent evaluations. This indicates that the z-score speaker normalization of pitch is successful in normalizing out speaker differences in the pitch domain. However, it is the energy feature set which shows the most dramatic reduction in performance; the difference of 5.52% accuracy is approaches significance with p= 0.0209. In the speaker dependent evaluation, energy is the most predictive feature set for pitch accent detection, while on BDC-read these energy features show a strong speaker dependence. This effect is not consistent across genre; the same speakers are represented in the BDC-spontaneous corpus, and this speaker dependency is not observable in the evaluation of this material. This suggests a significant effect of genre on the use of energy to indicate pitch accent. Closer investigation of intonational differences across speaking styles – read or spontaneous – is required to more completely hypothesize about the source of this effect.

These acoustic experiments use a set of acoustic aggregations extracted over words to detect pitch accents. The importance of context in the detection of pitch accent has been observed on all corpora and feature sets, providing additional support for the findings of Levow [115]. These experiments serve as an acoustic baseline for pitch accent detection. The remainder of our investigation into pitch accent detection moves in three directions. We evaluate the decision to detect pitch accents at the word level, instead of at the syllable or syllable nucleus (vowel) level in Section 3.4. In the current section, we observed energy to be the most predictive feature for pitch accent detection. In Section 3.5, we examine the use of energy as a predictor of pitch accent — specifically the energy contained in different spectral regions — leading to the construction of a classifier combination approach to pitch accent detection in Section 3.6. Finally, we investigate the use of part-of-speech information for pitch accent detection in Section 3.7.

3.4 Acoustic Pitch Accent Detection at the Word, Syllable and Vowel Domains

As discussed previously, prosody in language like Standard American English can be used by speakers to convey semantic, pragmatic and paralinguistic information. Words are accented to convey information such as contrast [240], focus [72], topic [76], and information status [70, 47]. The communicative implications of accenting, in most cases, impact the interpretation of a word or phrase. However, the acoustic excursions that are associated with accenting behavior are typically aligned with the lexically stressed syllable of an accented word. This apparent disparity between the domains of acoustic properties and communicative impact has led to different approaches to automatic pitch accent detection.

In this section, we compare automatic pitch accent detection at the vowel, syllable, and word level to determine which approach is most accurate. Some of the research presented in this section was first described in [172]. For this comparison, we employ only acoustic features. Advocates of vowel-based detection propose that, if acoustic excursions associated with pitch accents are localized to lexically stressed syllables, then analyzing only the syllable nucleus – the vowel – will eliminate noise introduced by the surrounding syllable or word context. Examining the full syllable, in contrast, will include a relatively small amount of surrounding acoustic information, and is motivated by the assumption that all of the necessary information for pitch accent detection is available in the lexically stressed syllable of an accented word. Pitch accent prediction at the word level, on the other hand, includes additional context, on the assumption that there are broader effects of accenting within the word. Many accents reach their full acoustic excursion beyond the borders of the word's lexically stressed syllable. While the inclusion of other syllables in the word may introduce some noise to the analysis, capturing these additional effects may counter-balance this noise.

The domain of automatic pitch accent prediction also has an impact on how that prediction is to be used and how it is evaluated. While some downstream spoken language processing tasks benefit by knowing *which* syllable in a word is accented, such as distinguishing **des***ert* from *de***sert** or clarification of communication misunderstandings, such as "I said **un**lock the door – not lock it!", most applications care only about which *word* is intonationally prominent. For the identification of contrast, given/new status, or focus, only word level information is required. While nucleus-based or syllable-based predictions can be translated to word predictions, such a translation is not always performed, making it difficult to compare performance of different approaches.

In this section, we describe experiments in pitch accent detection on the Boston University Radio News Corpus (BURNC), comparing the use of vowel nuclei, syllables and words as units of analysis. Here, we briefly summarize the corpus stats from Chapter 2

The BURNC material used in these experiments comprises 157.9 minutes (29,578 words) from six professional speakers reading radio news [143]. 54.7% (16,178) of words are accented – a high percentage, but in line with what we know of accenting in broadcast news in general [24]. Time-aligned phone boundaries generated by forced alignment are also marked, and and are used to identify the vowel regions for analysis. There are 48,359 vowels in the corpus, 16,806 of these are accented, yielding an accent rate of 34.8%. To generate time-aligned syllable boundaries, we align the force-aligned phones with a syllabified lexicon included with the corpus. This lexicon also includes information on which syllable in the word receives primary lexical stress. Table 3.5 shows the distribution of pitch accents in this corpus. We have collapsed the *downstepped* versions of pitch accent types[2] with their non-downstepped counterparts, e.g. !H* with H*.

H*	L+H*	X*?	H+!H*	L*	L*+H
61.15%	18.67%	12.40%	3.87%	3.24%	0.31%
10038	3137	2084	650	545	52

Table 3.5: *BURNC pitch accent distribution with collapsed downstepped versions.*

The use of this corpus for accent prediction in our three domains is not straightforward, due to some anomalies in the corpus. First, the lexicon and forced-alignment output in BURNC use distinct phonetic inventories; to align these, we have employed a *minimum edit distance* procedure where aligning any two vowels incurs zero cost. This guarantees that, at a minimum, the vowels, and particularly, the lexical stressed vowels, will be aligned correctly. Also, the number of syllables per word in the lexicon does not always match the number of vowels in the forced-alignment. This leads to 114 syllables that contained two forced-aligned vowels, and 8 which contained none. We do not remove these from the data set. While these are obviously syllabification errors, including them in our analysis allows us to report the performance of the acoustic-based detection as well as the syllabification technique. This syllabification approach generated 48,253 syllables, 16,781 (34.8%) bearing accent.

BURNC labelers were allowed to annotate multiple accents within a word, resulting in

[2] Accents uttered in a compressed pitch range, cf. [183].

629 words that contain multiple accents (615 contain two, and 14 contain three). These cases include compounds where both elements are accented (e.g. “school-based”) and abbreviations, which are treated as single words, in which multiple components are accented (e.g. “S.J.C.”). They also include multiply-accented words apparently arising erroneously from discrepancies between the ToBI annotations and the boundaries of the forced alignment, which result in accents being aligned to the wrong word or 15 instances in which a single phone defined by the forced alignment boundaries which contains two accent annotations along with another 25 syllables that include two accented vowels. Without human corrected word boundaries it is not possible to recover the correct association between word or syllable and accent annotation. Finally, a word may, in fact, have multiple syllables which bear accent. When a region – word, syllable, or vowel – contains multiple accents, we consider it as “accented”. When describing accent type distributions, the type of the last accent is used.

To evaluate the discriminative power of acoustic information drawn from vowel, syllable and word regions, we train Logistic Regression – maximum entropy – models to detect pitch accent using acoustic features drawn from these different regions, using the Weka Machine Learning toolkit [232]. The features we use include pitch (f0), energy and duration information, which have been shown to correlate with pitch accent in English. To model these, we calculate pitch and energy contours for each token using Praat [20]. Duration information is derived using the vowel, syllable or word segmentation described earlier in this section. The feature vectors we construct include features derived from both raw and speaker z-score normalized pitch and energy contours. For vowel-nucleus-based analysis, material from surrounding consonant regions is excised from the contours. The feature vector used in all three analysis scenarios is comprised of basic aggregations of these raw and normalized acoustic contours, specifically minimum, maximum, mean, standard deviation and the z-score of the maximum given the region. The duration of the region in seconds is also included.

The results of ten-fold cross validation classification experiments are shown in Tables

3.6 and 3.7. When running ten-fold cross validation on syllables and vowels, we divide the folds by words. That is, each syllable within a word is a member of the same fold. To allow for direct comparison of these three approaches, we generate word-based results from vowel- and syllable-based experiments. If any syllable or vowel in a word is hypothesized as accented, the containing word is predicted to be accented. Conversely, we generate syllable- and vowel-based results from word-based experiments: Word-based predictions are transferred to the syllable containing primary lexical stress in the lexicon – or to its vowel nucleus. (Note that this approach makes it impossible to predict more than one accent per word.) All syllables that are *not* assigned primary lexical stress in the lexicon are predicted to be not-accented.

Region	Accuracy (%)	F-Measure
Vowel	73.3 ± 0.845	0.540 ± 0.0163
Syllable	80.0 ± 0.774	0.692 ± 0.0139
Word	85.5 ± 0.800	0.791 ± 0.0129

Table 3.6: *Syllable-level accuracy and F-Measure*

Region	Accuracy (%)	F-Measure
Vowel	68.5 ± 1.66	0.651 ± 0.0171
Syllable	75.6 ± 0.648	0.756 ± 0.00976
Word	82.9 ± 0.872	0.845 ± 0.00838

Table 3.7: *Word-level accuracy and F-Measure*

Vowel/syllable-evaluated accuracies should be higher than word-based accuracies since the baseline is significantly higher. However, we find that the F-measure for detecting accented units is consistently higher for word-based results. A prediction of **accented** on any of the $N \geq 1$ component syllables is sufficient to generate a correct word prediction. Moreover, syllable-based recall is a soft lower bound to word-based recall: if a pitch accent is identified at the syllable/vowel level it is necessarily identified at the word level.

Our results suggest, first of all, that there is discriminative information beyond the syllable nucleus of a syllable. Syllable-based classification is significantly better than vowel-nucleus-based classification, whether we consider accuracy or f-measure, and whether we calculate these measures over words or syllables. Moreover, we find that there is discriminative information outside the lexically stressed syllable of a given word. In both word- and syllable/vowel-based, word-based classification yields significantly better pitch accent detection than either vowel- or syllable-based strategies. While this may be due to errors in the forced-alignment phone boundaries, until automatic phone alignment improves, word-based prediction appears to be more reliable. It is also possible that an acoustic syllable nucleus detection approach could generate more discriminative regions of analysis for pitch accent detection than the forced-alignment and lexicon alignment technique used here. However, in Chapters 5 and 6, we find these approaches to perform poorly for pitch accent type and phrase ending intonation classification.

One explanation for the superiority of word-based prediction over syllable- or vowel-based strategiesis is that the acoustic excursions correlated with accent occur outside a word's lexically stressed syllable. In particular, complex pitch accents are generally realized on multiple syllables. To examine this possibility, we looked at the error distribution of the three classification scenarios. The distribution of pitch accent types of missed detections using syllable-based evaluation of the three scenarios is shown in Table 3.8.[3] In the ToBI framework, the complex pitch accents include L+H*, L*+H, H+!H* and their downstepped variants. As we suspected, larger units of analysis lead to improved performance on complex tones. For vowel-based error distributions, χ^2 tests support the hypothesis that the syllable and word error distributions are drawn from a distinct population from the vowel error distribution with $p \leq 0.0001$; $\chi^2 = 56.641$ and $\chi^2 = 99.563$, respectively. Moreover, χ^2 analysis of the difference between the syllable and word error distributions yields a χ^2 of 42.108, corresponding to a p-value less than 0.0001.

The results presented above are from classification experiments using acoustic information from the target domains only. Accenting, however, is the perception of a word as more

[3] Again, downstepped and non-downstepped versions of each accent type, e.g. H* and !H*, are collapsed. Tokens annotated as X*? are considered accented for pitch accent detection evaluation, but are omitted from this analysis.

Region	H*	L*	Complex
Vowel	68.25% (3732)	6.86% (375)	24.89% (1361)
Syllable	70.33% (2422)	8.51% (293)	21.17% (729)
Word	74.22% (2002)	6.10% (165)	19.86% (537)

Table 3.8: *Error Distribution of missed H*, L* and complex pitch accents under syllable/vowel based evaluation.*

prominent than surrounding words. Extracting features that incorporate local contextual acoustic information thus might improve detection accuracy at all levels. To represent surrounding acoustic context in feature vectors, we calculate the z-score of the maximum and mean pitch and energy over six regions. Three of these are "short range" regions: one previous region, one following region, and both the previous and following region. The other three are "long range" regions. For words, these regions are defined as two previous words, two following words, and both two previous and two following words. To give syllable- and vowel-based classification scenarios access to a comparable amount of acoustic context, the "long range" regions covered ranges of three syllables or vowels. There are approximately 1.63 syllables/vowels per word in the BURNC corpus; thus, on balance, a window of two words is equivalent to one of three syllables. Duration is also normalized relative to the duration of regions within the contextual regions. Accuracy and f-measure results from ten-fold cross validation experiments are shown in Tables 3.9 and 3.10.

Analysis Region	Accuracy (%)	F-Measure
Vowel	77.5 ± 1.138	0.651 ± 0.0215
Syllable	83.9 ± 0.462	0.764 ± 0.00891
Word	86.4 ± 0.854	0.804 ± 0.0154

Table 3.9: *Syllable-level Accuracy and F-Measure with Contextual Features*

Analysis Region	Accuracy (%)	F-Measure
Vowel	77.4 ± 1.371	0.774 ± 0.0192
Syllable	81.9 ± 1.0201	0.829 ± 0.0101
Word	84.2 ± 1.279	0.858 ± 0.0143

Table 3.10: *Word-level accuracy and F-Measure with Contextual Features*

These tables show dramatic increases in the performance of vowel- and syllable-based pitch accent detection when we include contextual features. Vowel-based classification shows nearly 10% absolute increase accuracy when translated to the word level. The improvements to word-based classification, however, are not statistically significant. It may be that word-based analysis already incorporates the contextual information that is helpful for detecting pitch accents.

Issues in Translating Accent from Words to Syllables

When transferring accents from words to syllables, we assign the word-based prediction to the lexically stressed syllable, as determined by the BURNC lexicon. In cases of double-accenting, or accents that do not fall on lexically stressed syllables, this transfer introduces some error, as noted Section 3.4. In fact, 7.6% (1,274 of 16,781) of accents in the corpus do *not* fall on lexically stressed syllables. So, when we transfer accent prediction from word to lexically stressed syllable, we will not predict the correct syllable to be accented. Non-lexically stressed syllables may bear accent for a number of reasons: 1) Contrast, as in, “I said *un*lock the window, not lock it”. 2) Stress shift. In complex nominals, lexical stress can be shifted to canonically unstressed syllables to avoid *stress clash*. For example, in “eighteen ninety” the canonical lexical stress would dictate the a stress pattern as in “eight**teen** **nine**ty”, but the likely production would be “**eight**een **nine**ty”. 3) Double-accenting. A word may, in fact, have multiple accent bearing syllables, such as “**Mass**a**chu**setts”. 4) Forced-alignment errors. The location of the accent may, in fact, correspond to the lexically stressed syllable, but incorrect forced-alignment may result in an assignment to a syllable that is unstressed. 5) Incorrect lexical stress assignment in compound words and abbreviations. Since compound words and abbreviations are treated as single words in the corpus, a single prediction for that item can be transferred to only one syllable, when multiple syllables may be accented. In these situations, we place lexical stress on the final stressed syllable as indicated by the lexicon.

Addressing these sources of error is a challenging task, especially those due to forced-alignment noise. However, we *can* measure the error due to compound words and abbreviations. If we omit such items from our analysis entirely, our syllable-based performance does not significantly improve – with contextual features the accuracy is 86.5% ± 1.12, f-measure 0.805 ± 0.0161. However, when we examine the syllable-based error rates on compound words and abbreviations, we find error rates of 38.9% and 50.0% respectively, in comparison to a 13.9% error rate on syllables in other words. While there are not enough syllables in abbreviations (266 syllables in 92 tokens) and compound words (570 syllables in 182 tokens) to significantly depress overall performance, the prediction transfer clearly demonstrates difficulty in identifying the accent bearing syllable on these tokens via lexicon-based stress assignment

If we assume the existence of an oracle to determine any and all syllables that should be accented when a word is predicted as accented, we can measure the error introduced by the prediction transfer across all tokens. We find that, using no contextual features, an oracle-based prediction transfer from word-based hypotheses to syllables we can detect pitch accented syllables with 89.4% ± 0.67 accuracy and an f-measure of 0.850 ± 0.0105. Including contextual features improves performance, without statistical significance, to 90.1% ± 1.012 accuracy and an f-measure of 0.860 ± 0.0149. Thus the assumption that all accented words bear accent on their lexically stressed syllable (using forced-alignment segmentation and lexicon-based syllabification and stress assignment) introduces approximately 4% absolute error to accuracy and 0.06 to f-measure. Recall that human agreement on the pitch accent detection task is between 81% and 91% word-accuracy [156, 196], depending on the experience of the labelers and the genre of the material.

Acoustic-based word-to-syllable prediction transfer

Many of the errors arising from lexicon-based transfer of word-based predictions to lexically stressed syllables come from situations where the syllable indicated as lexically stressed is not the syllable bearing the accent. Four of the five explanations for this phenomenon arise from situations where the lexicon does not accurately describe the correct location of the stressed syllable. If the lexicon driven lexical stress assignment correctly identified the accented syllable, only double accents would introduce errors in the transfer of correctly detected accents. To more accurately identify the accent bearing syllable within a word, we experiment with acoustic based prediction transfer techniques. In this section, we present the impact of transferring word based predictions to syllables based on acoustic properties of the component syllables, specifically, their pitch, energy and duration – the same features that indicate the presence or absence of a pitch accent.

We explore transferring word-based predictions of the presence of a pitch accent to the syllable with the maximum duration, the maximum energy, the greatest mean energy, the maximum pitch and the greatest mean pitch. These experiments are based on the results of the word-based analysis ten-fold cross validation results reported in Table 3.9 and 3.10.

In addition to these single-rule based techniques we explored one data driven approach. For each fold, we used the training data points to train a Logistic Regression classifier to predict which syllable within a word would bear an accent. To train these models, we use only syllables of those words which are accented. The task is thus, differentiating accent-bearing syllables from non-accent-bearing syllables within accented words. The trained model is used to assign a transfer target to all testing set words. However, this target is only used to transfer a word-based prediction when the pitch accent detection model predicts the word to be accented. The feature vector for this model contains the syllable duration, minimum, mean, maximum, and standard deviation of pitch and energy, as well as the zscore of the mean and maximum relative to the syllable acoustic material, and the content of the surrounding word.

The model achieves 88.77% accuracy on those syllables within accented words. While the model is evaluated on all syllables, those within accented and unaccented words alike, the model predictions are only used if a word is predicted to be accent bearing. Moreover, there is only evaluation data for this accent location modeling for accented words; there is no way to evaluate the model performance on unaccented words. Recall, however, that the lexicon-based technique, identifies the correct location of 92.4% of accents – only 7.6% of accents fall on syllables that do not bear lexical stress according to the lexicon. In Table 3.11, we present the syllable level accuracy generated by using these acoustic-based prediction transfer techniques.

Transfer Strategy	Accuracy (%)	F-Measure
Oracle	90.1 ± 1.012	0.860 ± 0.0149
Lexicon	86.4 ± 0.854	0.804 ± 0.0154
Max Dur	77.97 ± 0.938	0.681 ± 0.0150
Max Pitch	74.47 ± 0.686	0.588 ± 0.0190
Mean Pitch	73.46 ± 0.895	0.572 ± 0.0171
Max Energy	80.96 ± 0.861	0.724 ± 0.0131
Mean Energy	70.48 ± 1.292	0.572 ± 0.0171
Classifier	82.22 ± 1.023	0.742 ± 0.0164

Table 3.11: *Syllable-level accuracy and F-Measure using a transfer strategy from Word-based predictions.*

None of the acoustic approaches are able to provide better accent location information that the lexicon-based transfer approach. However, we know that there are two classes of words that cause particular problems for the lexicon based transfer approach. Compound words and abbreviations have ambiguous lexical stress annotations in the lexicon. Moreover, recall that we observe error rates of 50% of syllables in abbreviated words and 38.7% in compound words. Since we can identify these classes of words unambiguously, we are able to selectively apply these acoustic models to only these classes of words. While these represent only 1.73% of tokens in the corpus, this is a section of the corpus where these approaches may reduce the error introduced by the prediction transfer technique. Note, oracular transfer errors come from instances where the word-based prediction produced an miss.

In Table 3.12, we report the overall detection error on syllables in abbreviations and compound words under each transfer technique. We find on these tokens the accent placement classifier produces the best transfer results. While significantly worse on "normal"

Transfer Strategy	Abbrev.	Compound
Oracle	95.9%	98.8%
Lexicon	50.0%	61.1%
Max Dur	71.4%	57.5%
Max Pitch	65.4%	68.3%
Mean Pitch	70.0%	70.3%
Max Energy	68.0%	71.8%
Mean Energy	45.8%	57.5%
Classifier	81.6%	78.1%

Table 3.12: *Syllable-level detection accuracy on abbreviations and compound words.*

words, when the lexicon is unable to identify the lexical stress within the word, this classification technique is able to improve the accent location identification on difficult tokens. By applying the lexicon-based transfer technique on standard words and this classifier based transfer routine on abbreviations and compound words, we are able to improve the syllable level pitch accent prediction accuracy to 86.6% with an f-measure 0.807. This is a modest, though not statistically significant, improvement.

Conclusion

In this section, we describe experiments detecting the presence of pitch accent on words, syllables and vowels. To allow for direct comparison between these three scenarios, we generate word level and syllable level results for each scenario. To construct syllable level results, word-based predictions are transfered to the lexically stressed component syllable. Comparable word level results are constructed by considering a word as accented if any component syllable or vowel is predicted as accented.

These experiments show that, without context, word-based detection significantly outperforms syllable- or vowel-based approaches whether evaluated either at the word or syllable level. Extracting features that incorporate acoustic information from surrounding context, improves performance in all three scenarios and on both types of evaluation. However, the improvement when including these contextual features in word-based detection is not significant.

We find that we can generate syllable level pitch accent hypotheses using word-based detection and transferring the word-based predictions to the syllable level. Given an oracle that could determine which component syllable bears the accent(s) – within a word hypothesized to be accented – this transfer approach would be able to generate a syllable level accuracy of 90.1%. Without this oracle, however, we transfer word-based predictions to the lexically stressed syllable, as determined by a lexicon. This technique performs significantly better than detecting pitch accents at the syllable level directly: 86.4% over 83.9%. Moreover, if we augment this approach with an acoustic classifier to place accents on difficult tokens – abbreviations and compound words – we can improve this accuracy to 86.6%.

3.5 Using Filtered Energy Features to Detect Pitch Accents

In this section, we examine the relationship between the energy of a speech signal and pitch accents realized in its component utterances. We examine this relationship using an analysis-by-classification approach, by constructing simple decision-tree pitch accent classifiers using only energy features. Those energy features that correlate strongly with pitch accent, will yield greater classification accuracy than those that vary freely. It has been known since the 1950s [21] that speech signal amplitude is a significant aspect of prosody in general and accent in particular [14]. Additionally, Sluijter and van Heuven [185, 186] showed that the energy component of a high frequency subband – greater than 500Hz – highly correlates with stress in Swedish speech. These papers examined the energy components of four subbands, those greater than 500Hz, 1kHz, and 2kHz. The experiments presented in this section elaborate on this result by closely examining the correlation between pitch accent and the energy components of a larger range of frequency subbands.

As discussed in Section 3.2, a great deal of research attention has been given to the task of identifying stressed, accented and prominent words or syllables within a given utterance. While there is concensus among work on the detection of accent that the energy of a word or syllable correlates with accent, some the best way to leverage energy information of a speech signal into a reliable predictor of pitch accent has not yet been determined. Sluijter and van Heuven [186] showed that accent strongly correlates with the energy within a particular frequency subband namely that greater than 500Hz in Swedish. Heldner [78, 80] and Fant [53] continued to examine the effect of this “spectral emphasis” on pitch accent in read Swedish speech. This work shows thoroughly that “spectral emphasis” is an excellent predictor of picth accent, where spectral emphasis is relationship between the energy in a particular spectral region and the overall energy of the signal. On English speech, Tamburini [198, 199], based on the findings of Sluijter and van Heuven [186], reported that the energy components of the 500Hz to 2kHz frequency band were more predictive of prominence than the energy components from either 0 to 500Hz or above 2kHz. Also, Tepperman [207] used the RMS energy extracted from between 60 and 400Hz as a feature in his syllablic stress detection system on non-native British English speech. This research indicates that energy extracted from specific frequency regions as opposed to the entire spectrum is helpful in the automatic prediction of pitch accent in English. The work presented in this section examines the energy component of a large set of frequency bands, and determines which of these are the most predictive of pitch accent. Through this we intend to offer a more thorough assessment of the correlation between energy and pitch accent in read English speech.

The experimental method is presented in Section 3.5.1. In Section 3.5.2, our results are reported and discussed.

3.5.1 Method

In this investigation into the correlation between energy and pitch accent detection, we use data from the read portion of the Boston Directions Corpus (BDC), collected by Nakatani, Grosz, and Hirschberg for a study of the relationship between intonation and discourse structure [86]. A more thorough description of this corpus can be found in Section 2. The BDC material has been manually ToBI [183] labeled and also labeled for discourse structure.

The BDC-read subcorpus contains 50 minutes of speech and 10831 words. We employ the hand-segmented word boundaries from the ToBI orthographic tier during the extraction of energy features, and we assume that word boundaries are available in both training and test sets. Moreover, we use the ToBI accent tier as the ground truth pitch accent labels for the training and testing of our classifiers. 42.2% of words in this material bear accent.

In order to examine the correlation between energy and pitch accent, we use an analysis-by-classification approach. We construct a feature vector for each manually-segmented word whose elements contain only features derived from the energy of the speech signal. Using the pitch accent annotation from the manual ToBI labeling, we assign a binary class to each feature vector indicating whether the word is uttered with a pitch accent or not. Using this labeled data and ten-fold cross validation, we run classification experiments to determine how predictive of pitch accent the energy components of various frequency subbands are. We use the *weka* machine learning environment's [232] C4.5 implementation, J48, a decision-tree algorithm, for this classification.

The features we examine are computed from the energy component of a variety of frequency subbands. The frequency subbands are derived from the Bark scale [4] [51]. We vary the lowest frequency of the subbands from the bark edges 0 to 19 and vary the bandwidth from 1 to 20 bark. The maximum frequency of any subband is 20 bark due to the 8kHz Nyquist rate of the BDC speech material. These combinations yield 210 frequency subbands from which we extract energy features for analysis by classification. We perform the filtering and energy extraction using the Praat speech analysis tools [20].

The examined energy features include the minimum, maximum, mean, standard deviation, and root mean squared (RMS) of the energy, as well as features designed to capture the dynamics of the energy within the word. These features included the z-score of the maximum energy in the context of the current word, and the mean slope.

Whether a word is perceived as accented or not is determined by its acoustic properties

[4] We used a Bark-to-Hertz transformation function of $hertz = 600 * \sinh(bark/6)$

relative to its surrounding intonational context [109]. Therefore we include in the feature vector eight normalized energy features based on the surrounding region. These are the same context regions examined in Section 3.3. We vary the size of this contextual window in eight different ways: 1) one previous word, 2) one following word, 3) two previous words, 4) two following words, 5) one previous and one following word, 6) two previous and one following word, 7) one previous and two following words and 8) two previous and two following words. The energy features calculated over these regions included:

- The difference between maximum energy in the current word and the mean energy in the region, normalized by the standard deviation of the energy in the contextual window.
- The difference between mean energy in the current word and the mean energy in the region, normalized by the standard deviation of the energy in the contextual window.
- The maximum energy in the current word normalized by the energy range realized in the contextual window.
- The mean energy in the current word normalized by the energy range realized in the contextual window.

We follow the American School of intonational description for SAE (e.g. [152]) in assuming that pitch accents, while interpreted as a property of the word, are realized on a particular syllable of that word. Therefore, the detection of pitch accents in SAE acoustic analysis of the speech signal for detecting prosodic events might profit from information found at the syllable or syllable nucleus level. To that end we automatically determine syllable boundaries, as well as start and end times of syllable nuclei using algorithms based on [133] and [150], respectively. In order to identify the most predictive region of analysis within a word we run the classification experiments under three different configurations: using energy information extracted from 1) the entire word, 2) the longest syllable in the word – the candidate for accenting and 3) the longest syllable nucleus in the word. We chose to include the longest syllable and syllable nucleus with the two larger regions due to previous work that indicate that pitch accent correlates with a lengthening of the accented vowel (e. g. [230]) and that the canonically stressed syllable of a multi-syllabic word tends to be the longest syllable [221].

The BDC has not been annotated for syllable or phone identity and boundaries. We therefore cannot provide precise error rates for these automatic segmentation approaches. However, we are able to compare the automatically derived syllable counts to the canonical pronunciation forms of the ToBI orthographic tier. The syllable boundary detector [219] has an insertion rate of 20% and a deletion rate of 36%. The nuclei detector has an insertion rate of 14% and a deletion rate of 39%. As the actual pronunciations may differ significantly from the cannonical forms, and pairs of deletion/insertion errors may be attributable to alignment problems, we make no claims as to the veracity of these error rates; they should be taken merely as estimates of the true accuracy of the automatic syllable segmentation systems.

In total, we explore six experimental configurations. We constructed 210 classifiers, one for each frequency subband, extracting energy features from either 1) the whole word, 2) only the longest syllable, and 3) only the longest syllable nucleus. Additionally, for each of these 3 configurations, we constructed either speaker-dependent classifiers, one for each speaker, or a single classifier using tokens from all 4 speakers.

3.5.2 Results and Discussion

The classification accuracy produced by our machine learning experiments indicates significant[5] differences in the descriminative power of energy information extracted from distinct frequency subbands. Across all experimental scenarios – extracting energy from four different regions within a word, and looking at speakers individually or all together –

[5] Statistical significance was determined by χ^2 with $p \leq 0.001$.

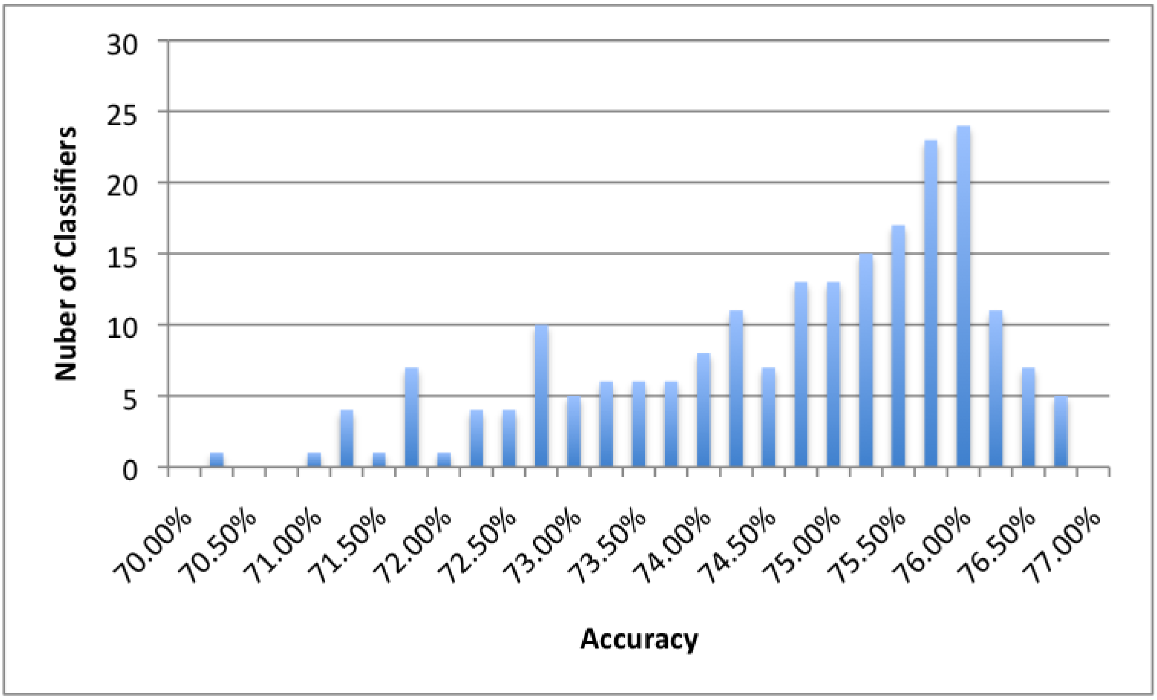


Figure 3.1: *Histogram of filtered energy classifier performance accuracy with word level features*

the mean relative improvement of the most predictive subband over the least predictive is 14.8%. A histogram of classifier accuracy using ten-fold cross-validation over material from all speakers can be found in Figure 3.1.

While our experiments confirm the claim that accent is realized through increased energy in a particular frequency subband(s), the classification results do not support the regions employed in previous research on automatic prodody prediction.

Tamburini [198, 199] used the frequency region between 500 and 2000Hz. Sluijter and ven Heuven [186] found the region above 500Hz to correlate with pitch accenting. Tepperman et al. [207] extracted energy from 60 to 400Hz. Energy features extracted from these regions do not yield the best performing pitch accent detection accuracy for any of the four speakers represented in BDC-read. That said, there are experimental configurations in which the classifications based on energy contributions from the above frequency regions are not significantly worse than the most predictive band. We find that the frequency range that yields the most predictive features is between 3 and 18 bark (312Hz to 6000Hz) with energy information drawn from the entire word. This band correctly classifies 76% of all words on average over 10-fold cross-validation, above a baseline of 57.6%. The precision and recall for detecting accented words is 71.6% and 73.4%, respectively. The most predictive features used in this classifcation are the z-score of the maximum energy and difference of maxima normalized by the standard deviation based on regions including 1) 1 previous word, and 1 following word, 2) 2 previous words and 1 following word, and 3) 2 previous and 2 following words. However, this subband generates results significantly worse than the best performing subband for one of the four speakers – regardless of the region of analysis within each word. Interestingly enough this speaker is one of the three male speakers, not the female speaker. The band from 2 to 20 bark (203Hz to 8kHz)[6], while not being the most predictive region in any experimental configuration, is only significantly[7] worse than the best performing band in one configuration[8]. With energy drawn from the entire word, the subband between 2 and 20 bark correctly classifies 75.5% of all words. The precision and recall for detecting accented words is 70.5% and 72.5%. The most predictive features used in this classification are identical to those used for the 3 to 18 bark band. Due to both the predicive power and robustness of this energy component within this frequency subband to a variety of speakers and types of analysis, we believe this to be the best region from which to extract energy information for the prediction of pitch accent in read speech. The Nyquist rate of our corpus is between 19 and 20 bark, therefore it is impossible to tell whether the band is more accurately described as "greater than 2 bark" or strictly between 2 and 20 bark.

Observing that distinct subbands were predicted pitch accent with varying accuracies, we analyze the classification results to determine what degree the correct classifications overlapped. We expect a high degree of overlap if there existed a set of words that are easy to classify and one of difficult words. We find that examining the hypotheses produced for all data points by every analysed subband with energy taken from the full word, there is a

[6] NB:8kHz was the Nyquist rate of the corpus.

[7] Statistical significance determined by χ^2 indicating $p <= 0.05$

[8] Classification of speaker h1, a male, with features extracted from the longest nuclei of the word.

relatively small intersection of predictions, even between overlapping or adjacent subbands. Overall, the pairwise mean overlap between two classifiers is 74.71%. Two arbitrarily chosen classifiers will agree on the assessment of approximately three fourths of the data points. If we break down this analysis by predictions generated by overlapping or adjacent regions, we find that classifiers trained on overlapping energy data share 74.87% of their predictions – not significantly different from the mean. Moreover, classifiers trained on adjacent regions share on average 72.37% of their classifications – *below* the mean. We observe a similar effect when we compare pairwise differences in overall accuracy. The overall mean pairwise accuracy difference is 1.781%. Classifiers trained on overlapping regions differ by more than the mean, 1.961%, while those trained on adjacent regions differ by 2.066%. Both of these measures show a significant effect of the region type – overlapping, adjacent, or neither – when evaluated using ANOVA, both with $p < 2.2 * 10{-}16$.

Moreover, 10806 out of 10831 data points are correctly classified by at least one of the 210 classifiers. A plot of the coverage of the energy-based predictors can be found in Figure 3.2. To exploit the predictive power of the individual classifiers, we employ a voting scenario, where each classifier classifies a given data point, and the majority classification is used as the final hypothesis. Using this voting classifier, the accuracy improves to 79.9%. This is a quite high accuracy, considering that the classification ignores $f0$ and duration information – known to be strong correlates to pitch accent.

The preceeding discussion of results describes only those energy features extracted from the entire word. In addition to this, we run classification experiments extracting the energy from 1) only the longest syllable, 2) only the longest syllable nucleus. However, each of these configurations yield classification performances significantly worse than those obtained by generating features based on the energy in the entire word. The best performing single classifier using word features achieves 76.6% accuracy, while the best syllable-based classifier performs at 72.8% accuracy, with nucleus-based at 72.6%. The mean accuracies of the word-based, syllable-based and nucleus-based classifiers are 74.5%, 68.6% and 68.0%,

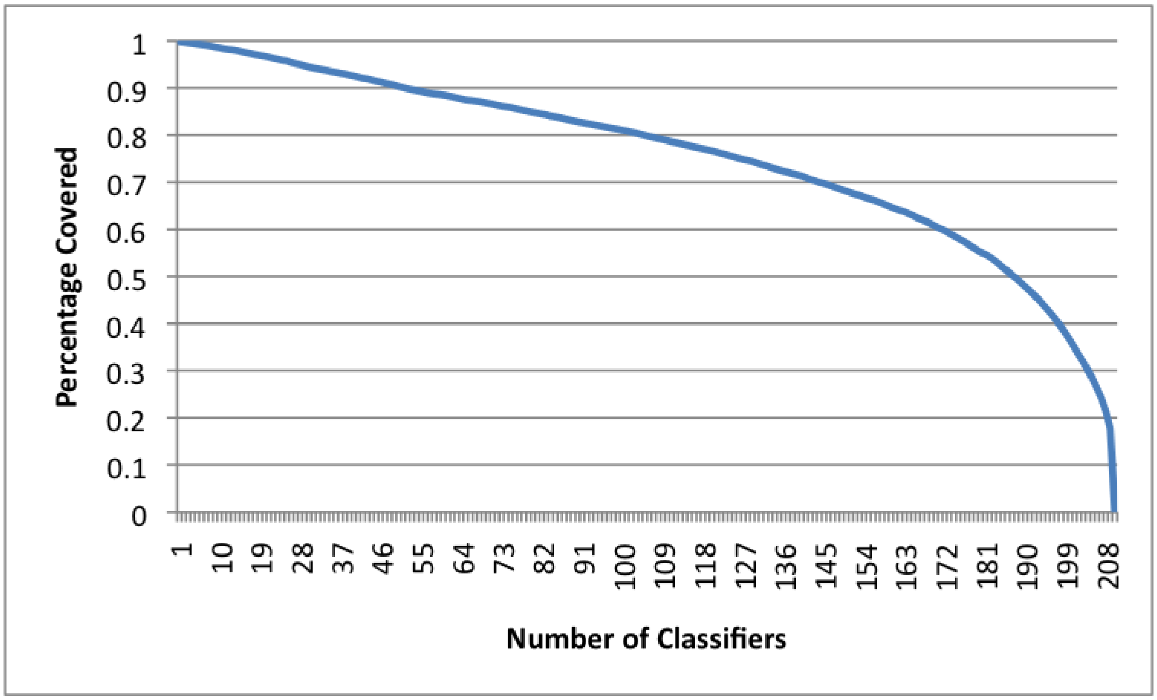


Figure 3.2: *Plot of the portion of data points correctly classified by at least N energy classifiers.*

respectively. Moreover, we find that the classification accuracies obtained by using the longest syllable and the longest syllable nucleus as the region of analysis did not significantly differ. This is an interesting result, especially given that the duration of the syllable nucleus and the syllable obtained "almost the same ratio of correct classifications" of syllablic prominence reported by [198]. These two findings would suggest that the syllable nucleus is a sufficient region of analysis to extract both duration and energy information for the purposes of detecting pitch accent at the syllable level.

3.5.3 Conclusion

In this section, we describe an analysis-by-classification approach to determining if the energy contributions from different frequency bands correlate to pitch accent to different degrees in read English speech, and if so, which bands are the most predictive. These experiments confirm that the energy component from differing frequency subbands predict pitch accent with differing degrees of success. Specifically, we find that the band between 3 and 18 bark to be the most predictive on BDC-read material, containing tokens from all four speakers. However, the band between 2 and 20 bark predictes pitch accent significantly better than the band from 3 to 18 for one speaker, while not prediciting significantly worse for the other three. As the Nyquist rate of our corpus is between 19 and 20 bark, it is moot whether this band is more accurately reported as “greater than 2 bark” or strictly between 2 and 20 bark. This result may be due to the removal of noise that may be present in low-frequency bands. Specifically the low-frequency energy subband between 0 and 2 bark may show significant covariation with f0 while the higher energy regions may be more sensitive to harmonic structure while being relatively invariant to changes in f0.

We observe that the differences in predictive power between frequency subbands is not merely one of varied accuracy, but rather that different subbands can accurately detect pitch accents on different sets of words. Using a voting scheme, we are able to construct a classifier based on every examined subband – base frequency from 0 to 19 bark, bandwidth from 1 to 20 bark. This voting classifier predicts pitch accent with 79.9% accuracy. Extending this approach we investigate an automatic method to determine which frequency region contains the most predictive energy information for the prediction of pitch accent in Section 3.6.

3.6 Corrected Energy Based Classifier

As discussed in Sections 3.2 and 3.3, three major acoustic correlates to pitch accent are pitch excursions, increased intensity and prolonged vowel duration [21, 14]. In Section 3.5, we explored the discriminative properties of energy features extracted from a range of frequency subbands. We found that energy features extracted from different frequency subbands, even adjacent and overlapping ones, predict pitch accent with varying degrees of accuracy, and moreover produce correct predictions on different subsets of data points. It was determined that the frequency region between 2 and 20 bark was the most accurate, and robust predictor to pitch accent. Additionally, we observed that at least one of the energy-based predictions was correct for upwards of 99% of all words. In this section, we build upon these results, investigating techniques to leverage these predictions along with pitch and duration information to the ends of constructing a robust, high-accuracy pitch accent detector.

3.6.1 Methods

We explore a number of techniques of combining results from the filtered energy experiments with pitch and duration features in order to create a robust pitch accent detection module. To separate the learning architecture from the features used, we extract the same acoustic features for each classification experiment.

Pitch and Duration Features

We compute, for each word, the minimum, maximum, mean, root mean squared and standard deviation of pitch (f0) values extracted using Praat's [20] Get Pitch (ac)... function. We also compute each of these features based on speaker normalized pitch values. This normalization is performed using z-score normalization. We include in the feature set, the above features calculated over the slope of both the raw and speaker normalized pitch tracks.

Additionally, we use eight contextual windows to account for local context. These are the same contextual windows explored in Sections 3.3 and 3.5, namely, 1) one previous word, 2) one following word, 3) two previous words, 4) two following words, 5) one previous and one following word, 6) two previous and one following word, 7) one previous and two following words and 8) two previous and two following words. Based on the pitch content of these regions we perform z-score and range normalization on the maximum and mean raw and speaker normalized f0 and slope of f0 of the current word. We extract three timing features: the duration of the current word in seconds, the duration of the pause between the current and following word, and the duration of the pause between the current and previous word.

Energy Features

We extract energy information from 210 distinct frequency bands. These frequency bands are constructed by varying the minimum frequency from 0 bark to 19 bark, and the maximum frequency from 1 bark to 20 bark. 20 bark is the maximum frequency in all of our corpora (cf. Section 2) due to Nyquist rates of 8kHz.

For each word, we extract the maximum, minimum, mean, standard deviation of energy, and the z score of the maximum relative to the word. Additionally, we use the same eight contextual windows to account for local pitch content to normalize out local context from the energy information. Based on the content of these regions we perform z-score and range normalization on the maximum and mean energy of the current word.

Simple decision trees

In order, to have a point of comparison for our experiments with filtered energy features, we first perform pitch accent detection using feature vectors containing the pitch, duration and unfiltered energy features.

In [169] and Section 3.5, based on experiments with the BDC-read corpus, we hypothesized that the frequency region between 2 and 20 bark contains energy information that is the most robustly discriminative of pitch accent. To evaluate this claim, we run classification experiments on all three corpora with feature vectors containing energy features drawn from the 2-20 bark frequency subband along with pitch and duration features.

Voting classifiers

Using an ensemble of classifiers, each trained using only energy features extracted from a single frequency subband, we construct a simple majority voting classifier. For each data point, 210 predictions are obtained – one from each filtered energy-based classifier. The ultimate prediction for each data point was the class (‘accented’ or ‘non-accented’) predicted by at least 106 energy-based classifiers. In the case of a tie, the data point is assigned to the

the majority class in the corpora. We also evaluate the performance of a weighted majority voting classifier, where we weight the predictions by the J48 confidence scores.

We observe that on both BDC-read and BDC-spontaneous, the oracular coverage of the 210 predictors is over 99%. That is, at least one energy-based classifier produces a correct prediction for nearly every word in both corpora. We perform two experiments examining ways of using pitch and duration information to determine which predictors will be correct for a given word.

In the first experiment, we construct our feature vector using the pitch and duration features along with the 210 raw predictions from the filtered energy-based classifiers. When evaluating this type of classifier in a cross-validation setting, particular attention is paid to guarantee that none of the elements of the testing set were used in constructing the predictions included in the training set feature vector. To that end, for each training and testing set, an additional ten-fold cross validation scenario was run over the training set in order to produce predictions for use in the training feature vector. The testing set predictions were based on energy-based classifiers trained on the full training set.

The expectation in constructing this type of classifier is that rules would automatically be learned that could either associate predictions from frequency bands or associate pitch features that might distinguish when one frequency band might be more predictive than another. In Figure 3.3 we can observe an instance of the former relationship. The behavior represented by this clipping of the decision tree says that, for a given word, following some number of previous decisions, if the speaker normalized mean pitch is below 0.6, then predict non-accented. If this pitch value is greater than or equal to 0.6, then trust the prediction made by the energy classifier trained on energy information within the frequency band between 8 and 16 bark. One possible explanation behind this type of decision is that this particular energy-based classifier is fairly accurate in a specific pitch environment, but fairly inaccurate in others. This type of branch inspires the next type of classification scheme, in which we make explicit the use of pitch-based features to correct energy-based predictions.

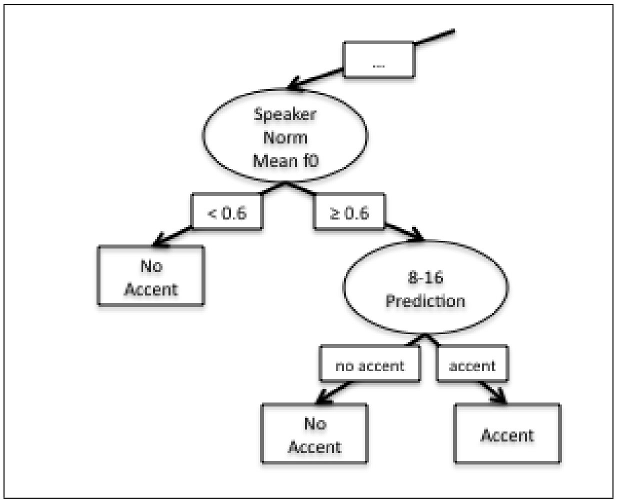


Figure 3.3: *Detail view of a portion of a single pitch-based classifier*

In our final classifier design, we make the relationship between pitch and duration information and filtered energy based predictions explicit. For each frequency band, we build a pitch and duration-based classifier that predicts when the energy-based prediction from the given frequency band will be correct, and when it will be incorrect. If a prediction is predicted to be incorrect, it will be inverted before contributing to the voting decision.

Again, when performing the ten-fold cross-validation on this two stage classifier, we pay particular attention to ensuring that no data point in the test set is ever used in producing a training set prediction. For each training set, we use ten-fold cross-validation to generate filtered energy-based pitch accent predictions for each frequency region. We then, for each energy-based classifier, train a second classifier using pitch and duration features that classifies each training-set energy prediction as either ‘correct’ or ‘incorrect’. Predictions that are classified as ‘incorrect’ are inverted. Thus, a ‘accent’ prediction classified as ‘incorrect’ becomes ‘non-accented’ and vice versa. Since, this correction is performed independently for each filtered energy-based classifier, we are left with 210 ‘corrected’ pitch accent predictions. We then combine these into a final prediction using a majority voting scheme. There are some examples of correction in previously published classifier combination and classifier fusion studies which served to inspire this work (cf. [108, 141, 135, 175]). However, the precise formulation of a correcting classifier described in this chapter does not appear in any of these previous works. Equation 3.1 contains a formula representing this decision process, where N is the number of corrected classifiers, $\phi(A|X_i)$ is the confidence of the energy-based classifier that the token is 'accented' and $\psi(C|Y_i, \phi(A|X_i))$ is the confidence of the correcting classifier that the energy prediction is correct. In the simple majority voting case, the confidence of each classifier is constrained to be either zero or one. In the confidence weighted voting, the confidence scores are determined by the decision tree algorithm. The confidence of a decision tree prediction is the accuracy of the decision tree leaf when evaluated on training data – while this is not strictly a posterior probability (i. e. $p(class|data)$), it serves to represent the expected performance of a given decision tree leaf.

$$c(A) = \sum_{i}^{N} \phi(A|X_i)\psi(C|Y_i, \phi(A|X_i)) * (1 - \phi(\neg A|X_i))(1 - \psi(\neg C|Y_i, \phi(A|X_i))) \qquad (3.1)$$

3.6.2 Results and Discussion

In this section, we report results of the ten-fold cross-validation evaluation of the techniques described in Section 3.6.1. In addition to evaluation on the two BDC subcorpora, read and spontaneous, we also evaluate the pitch accent detection performance of each approach on the BURNC material, and a subset of TDT4 [193] material.

The TDT-4 corpus [193] was constructed by the LDC for the Topic Detection and Tracking shared task, and was provided for use in the DARPA GALE project. As part of the SRI NIGHTENGALE team, Columbia University was provided with automatic speech recognition (ASR) transcriptions of the corpus by SRI [190] and hypothesized speaker diarization results by ICSI Berkeley [233]. The TDT-4 corpus as a whole comprises material from English, Mandarin and Arabic broadcast news (BN) sources aired between October

1, 2000 and January 2, 2001. However, for the experiments presented in this paper, we had one 30-minute broadcast, 20010131_1830_1900_ABC_WNT, annotated for pitch accent. The annotation was performed by a single experienced ToBI labeler. The annotator was asked to annotate the ASR transcript with pitch accent labels. Since ASR hypothesized word boundaries may not align with those perceived by a human listener, the annotator was asked to mark an ASR hypothesized word as containing a pitch accent if he believed any syllable within the ASR word to contain the realization of a pitch accent. After omitting regions of ASR error, silence and music, the TDT4 material for use contained approximately 20 minutes of annotated speech and 3326 hypothesized words. Note that we use the ASR hypotheses only for word boundaries – not for lexical content. The output of the automatic speaker diarization system identified 25 speakers within this show. These hypothesized identities are used to normalize acoustic information to account for speaker differences.

These four corpora include different amounts of human annotation. The BDC corpora use manual word boundaries, while the BURNC material word boundaries are generated by forced alignment recognition of manual transcripts. For these three corpora, BURNC, BDC-read and BDC-spontaneous, speaker identities are also known. On the TDT material, however, the word boundaries are generated from unconstrained automatic speech recognition (ASR), and speaker identities are obtained from an automatic speaker diarization module. Ideally, the value of human annotation or word boundaries and speaker identities would be performed on a single corpus with multiple word boundary and speaker identity decisions obtained from different sources. However, without access to high quality ASR and diarization systems, this is impossible. We were fortunate to have collaborators to recognize and diarize the TDT data as part of the DARPA GALE program. Diarization is a process which detects speaker turns, and groups those produced by the same speaker, in multiparty speech such as broadcast news, or meetings. The collaboration, however, did not extend to obtaining ASR and diarization hypotheses of the BDC and BURNC material. Absent this annotation, evaluation of these pitch accent detection techniques on corpora with reduced amounts of human annotation allows us to make some observations about how sensitive the evaluated techniques are to errors in word boundary placement and speaker identities. Table 3.13 contains the detection accuracy of each experiment configuration.

	BDC-read	BDC-spon	BURNC	TDT
Weighted Corrected Voting	84.38 ± 0.541	83.20 ± 0.394	85.51 ± 0.213	83.73 ± 1.164
Corrected Voting	84.13 ± 0.541	83.29 ± 0.377	85.46 ± 0.180	83.52 ± 1.230
Pitch/Dur + Predictions	78.47 ± 0.771	77.34 ± 0.574	80.55 ± 1.722	78.89 ± 1.132
Weighted Majority Voting	79.96 ± 0.492	80.67 ± 0.508	83.18 ± 0.377	82.74 ± 1.214
Majority Voting	79.87 ± 0.525	80.48 ± 0.574	82.93 ± 0.328	82.50 ± 1.033
'Best' Band Energy	79.12 ± 0.656	78.19 ± 0.689	82.12 ± 0.443	80.37 ± 1.607
No Filtering	78.82 ± 0.607	78.36 ± 0.558	82.32 ± 0.344	81.39 ± 1.296

Table 3.13: Pitch Accent Detection Accuracy using J48 Prediction

The baseline experiment, 'No Filtering', which uses pitch, duration and unfiltered energy features to train a standard decision tree, yields the lowest accuracy on all corpora. Replacing the unfiltered energy features with corresponding energy features extracted from the frequency band between 2 and 20 bark ("Best' Band Energy') does not yield significantly different results on any corpus. The hypothesis that the band between 2 and 20 bark would yield the most robust and discriminative energy features was based on experiments on the BDC-read corpus. On this corpus, we observe a statistically insignificant gain in accuracy of 0.30%. This band does not improve the accuracy on either other corpora – even insignificantly reducing it on BDC-spon. While the energy features extracted from the frequency region between 2 and 20 bark are able to predict pitch accent significantly better than unfiltered energy features, when combined with pitch and duration information, the impact of this improvement is insignificant under J48 classification.

Based on the 210 predictions per data point using exclusively those energy features extracted from each frequency subband ('Majority Voting'), a simple majority voting classifier achieves classification accuracy that is significantly better than the baseline experiment on the TDT and BDC-spon corpora. Weighted voting classifiers, where each prediction is weighted by either J48 confidence score, cross validation accuracy, or the product of the two, do not yield significantly different results from the majority voting classifier.

When we included the 210 energy-based predictions in a feature vector (‘Pitch/Dur + Predictions’) along with the pitch and duration features, the classification accuracy was reduced below that of the majority voting classifier. We expected the decision tree to learn associations between pitch features and energy predictions, or to identify mutually reinforcing sets of predictions. However, even the baseline classifier outperforms this approach.

The two-stage classification technique (‘Weighted Corrected Voting’), where pitch information is used to correct energy-based predictions before voting, demonstrates the best classification results on all corpora. On BDC-read, the accuracy of this technique is 84.38%. On the BDC-spontaneous, BURNC and TDT corpora the accuracy is 83.20%, 85.51% and 83.73% respectively. The human agreement on pitch accent identification is generally taken to be somewhere between 81% and 91%, depending on genre, recording conditions and particular labelers [230, 183]. These results thus represent a significant improvement over the baseline classifier, and approach human levels of competence. The fact that the accuracy on the TDT corpus is not significantly different from that obtained on the BDC material suggests that the technique is relatively indifferent to the fine grained accuracy of word boundary placement and speaker identity information. Recall that the BDC corpus word boundaries are manually defined, the TDT word boundaries are a result of ASR output. We also evaluate the correction technique using hard decisions rather than confidence scores (‘Corrected Voting’). This correction technique generates poorer results than the ‘Weighted Corrected Voting’ technique on all corpora, though t-tests do not show that this difference is statistically significant. These results raise the question of why decision tree classifiers (‘Pitch/Dur + Predictions’) are not able to learn the correction and voting structure that is manually enforced in the voting classifiers. While the decision tree classifiers have access to the same information, the local decision making at each decision tree node prevents a decision tree from modeling voting decisions. Voting decisions require information about the values of multiple features in a vector. Decision trees examine each feature in isolation

– when examining the prediction of a particular energy-based classifier, the decision tree does not ask “how many other features, other classifier predictions, have the same value as the current features”. This locality constraint prevents decision trees from modeling voting decisions. Discussion of why classifier combination can yield improved performance than extending a feature vector is beyond the scope of this thesis. Interested readers are directed to [58, 176] among others for further discussion of this issue.

This high accuracy performance on disparate corpora demonstrates that the technique is robust to genre, speaker and recording condition differences, as well as noise in word boundary locations. One drawback of this technique presented in this paper is that it is quite resource consuming to train and test. While there are many opportunities for parallelization, each data point requires 420 classifications in order for pitch accent to be detected.

The remainder of this section is devoted to other evaluations of this classification technique. We evaluate the robustness to speaker identity, genre as well as the application of different learning techniques to replace energy-based predictors, correcting classifiers and aggregation (voting) technique.

Classification Training Modifications

To train the correcting classifier, we require input predictions from an energy-based classifier in order to learn which predictions are correct and which are incorrect. These training predictions for the correcting classifier are trained using the elements of a single training fold. We explore two ways of generating these simulated training predictions. In the description of the technique described in Section 3.6, to approximate predictions based on unseen data, we generate training predictions using an approach similar to cross validation. We divide the *training* data, into another ten stratified partitions. For each partition, we train an energy-based model on the remaining nine partitions of the training data. This classifier is then used to generate predictions for the omitted partition of the training data. Thus, the predictions that are used as training data for the correcting classifier have not been used in the training of the classifiers which generated them. The goal of this technique is to approximate the behavior of the fully trained energy-based classifier on unseen data. One limitation of this approach is that it requires the training of an additional ten energy-based classifiers for each of the 210 frequency bands. The training of an additional 2100 decision trees per cross-validation fold leads to a significant increase in the time taken to run the full evaluation. If this care in approximating the energy-based classifier performance does not lead to improved classification performance, the simpler training method is obviously preferred.

To compare this training approach to the technique described in Section 3.6, we train each energy-based classifier on the full training set, and apply this classifier to each training point. The predictions obtained on these training points are then used in training the correcting classifier. However, these predictions over-estimate the performance of the energy-based classifier and may not accurately reflect the input that will be encountered by the correction classifier. We evaluate these two prediction generation approaches – the full training set and stratified partitioning – using ten-fold cross-validation with identical fold assignments. We find the accuracy of the system trained using predictions derived from the full training set to have an accuracy of 83.37 ± 0.49528 compared to an accuracy of 84.13 ± 0.51168 obtained using predictions generated by the stratified partitioning approach. Using a paired t-test we find these to differ significantly with p=0.0176. Thus there is a significant gain to training the correcting classifiers on cross-validated predictions, as opposed to predictions derived from the full training set. While not necessarily surprising, it is valuable to know that the extra effort in generating higher quality approximations of unseen data leads to improved overall performance of the system.

In previous experiments (cf. Section 3.3), we found Logistic Regression to be a better modeling technique than J48 for detecting pitch accent using acoustic features. To evaluate the use of Logistic Regression in this ensemble technique, we train the energy-based classifiers and correcting classifiers using Logistic Regression models instead of J48 decision trees. The performance of majority voting of energy-based Logistic Regression classifiers and corrected classifiers are reported in Table 3.14.

Modeling	Energy	Corrected	Change
J48	79.87 ± 0.492	84.38 ± 0.512	4.51
Logistic	79.22 ± 0.595	79.56 ± 0.458	0.34

Table 3.14: *Corrected Energy-Based Classifier accuracy (%) using J48 and Logistic Regression classifiers*

We observe that the voting classifier using energy-based decision tree components – without correcting classifiers – predicts pitch accent with slightly higher (0.76%) accuracy than an identical classifier using Logistic Regression components. While this difference is small, a paired t-test indicates that this difference is significant with p=0.00401. Based on the improved performance of Logistic Regression over J48 classification in a standard classifier, we expect to see this improvement reflected in a voting scenario. This raises the question of why are J48, decision tree, models more suited to learning correcting classifiers than Logistic Regression models. When we examine some of the correcting classifier trees, we often find branches that lead to ‘accented’ predictions being treated differently than ‘non-accented’ predictions. The decision tree learning approach – recursive partitioning – is particularly suited to model these non-linearities. Logistic Regression on the other hand, is a linear model, and is incapable of addressing these issues directly. One way of addressing this, would be, for each correcting classifier, to build a distinct Logistic Regression model for ‘accented’ predictions and one for ‘not-accented’ predictions. Manually encoding this non-linearity in the model may allow Logistic Regression models to be successfully applied to the correcting task. These claims about the non-linearity of the correcting task are, however, only evaluated using decision trees and Logistic Regression learners. Though it seems likely that similar results will be found under other linear and non-linear learning techniques, this has not been evaluated more broadly.

Prediction Combination Techniques

In the previous experiments, we combined the 210 corrected energy predictions into a final hypothesis using majority voting and weighted majority voting. This is, of course, not the only available technique for prediction combination. In this section, we present the result of experiments on BDC-read that explore the use of classification algorithms for prediction combination.

We run experiments using both standard voting as well as a voting approach where each vote is weighted by the confidence of the corrected prediction. In addition to these, we train combination classifiers to generate a final prediction based on a feature vector comprising the 210 predictions and confidence values from the corrected energy based classifiers, as well as the uncorrected predictions and confidences. For each of the ten evaluation folds, we train a supervised combining classifier based on the outputs from running training data through the classifier. For this combining classifier, we used SVM classification with a linear kernel, SVM classification with a radial basis function (rbf) kernel, Logistic Regression and J48 classification. These classification techniques cannot model voting decisions – they have no explicit mechanism to count the number of features with a particular value. Since we previously observed the success of voting decisions, we also evaluate these combination techniques with a feature vector that includes the previously mentioned information as well as the proportion of energy-based predictors that predicted ‘accented’ and the proportion of corrected energy-based predictors that predicted ‘accented’, as well as the weighted voting equivalents. The results of these experiments can be found in Table 3.15.

We find that weighted majority voting leads to the best results. The classification routines are unable to combine the predictions to improve pitch accent detection accuracy. When given access to the voting decisions, the J48 and SVM with linear kernel performance is improved, but the SVM with rbf kernels and Logistic Regression performance does not significantly differ. While the best performance is obtained using confidence weighted majority

Technique	Accuracy w/o Voting Feature	Accuracy w/ Voting Feature
J48	78.32 ± 0.492	80.56 ± 0.672
SVM linear	81.81 ± 0.377	82.12 ± 0.525
SVM rbf	84.28 ± 0.361	84.26 ± 0.312
Logistic	82.37 ± 0.394	82.17 ± 0.312
Voting	84.13 ± 0.508	
Weighted Voting	84.38 ± 0.508	

Table 3.15: *Ten-fold cross-validation accuracy using combination techniques.*

voting, the accuracy using a regular voting decision rule and the SVM with rbf kernels do not perform at rates that differ significantly. The SVM classifier is more computationally demanding in both training and evaluation requirements. Without significant performance improvements, we find this additional classification technique to be superfluous. We prefer the weighted majority voting over the standard majority voting, as the two have equivalent computational requirements and the weighted version performed better – if insignificantly – in these evaluations.

Evaluating Robustness to Genre

To evaluate the robustness to genre we evaluate the material from each corpus, BDC-read, BDC-spontaneous and BURNC on corrected classifiers trained on the other corpora. The BDC-read and BDC-spontaneous have the same speakers, and nearly identical lexical content. The main difference in these corpora is the speaking style. These corpora differ more dramatically from the BURNC corpus, which is professionally read news material. Due to these inherent qualities, we expect the performance on the two BDC corpora to be more similar to each other than either corpus with the BURNC material. Table 3.16 contains the results from this evaluation. Ten-fold cross-validation was used to evaluate a classifier on its training data. These evaluations appear along the diagonal of Table 3.16. The remaining cells report the accuracy of a classifier trained on the full training set of one corpus and evaluated on the full set of another. These evaluations contain no confidence interval information.

Test Corpus	BDC-read	BDC-spon	BURNC
BDC-read	84.38 ± 0.512	83.86	81.89
BDC-spon	83.33	83.20 ± 0.370	81.44
BURNC	81.67	82.35	85.51 ± 0.200

Table 3.16: *Corrected Energy-Based Classifier Detection Accuracy (%)*

We find that the performance on BURNC material is significantly lower when evaluated using either of the BDC trained classifiers, and, moreover, that the BDC accuracy is lower when evaluated on the BURNC classifier. We find that the detection accuracy of BDC-read material is not significantly different when we use a classifier trained on BDC-read material or BDC-spontaneous material. However, we find a significant degradation when a classifier trained on BURNC data is used. The same relationship is observed on the BDC-spontaneous material. That is, the accuracy is not significantly different when applying a model trained on BDC-spontaneous or BDC-read material, but the accuracy is significantly worse when applying the BURNC trained model. Recall, however, that the two BDC subcorpora contain material spoken by the same four speakers, producing nearly identical lexical material. This suggests that the difference in genre pale in comparison to speaker and domain differences. We find no genre difference given identical speakers and domain.

However, we can compare the non-professional read and spontaneous models to the professional read speech contained in the BURNC corpus. Here we find some minor evidence that non-professional spontaneous speech is more similar to professionally read speech than non-professional read speech is. First of all, we notice that the BURNC accuracy is 0.68% higher when detected using the classifier trained on BDC-spontaneous data compared to the accuracy using the classifier trained on BDC-read material. Using a difference of proportion test we find this difference to approach significance with p=0.0626. Second, we examine the performance of the classifier trained on BURNC material in detecting accents in BDC-read and BDC-spontaneous. Here we find no significant difference between the performance of the BURNC model on the BDC-read or -spontaneous subcorpora.

By calculating the weighted mean and confidence intervals of the performance of the BDC-read classifier on BURNC data, the BDC-spontaneous classifier on BURNC data, and the BURNC classifier on the two BDC subcorpora, we estimate the performance of this classification technique on unseen data, from dissimilar recording conditions, using unknown speakers as 81.97% ± 2.5584%. Due to the disparities in training and evaluation data, we consider this to be a lower bound on the expected performance of this pitch accent detection technique. We are not aware of similar evaluations of other pitch accent detection approaches.

Evaluating Robustness to Speaker

The previous evaluations used ten-fold cross-validation. To evaluate this modeling technique's robustness to speaker identity, we evaluate each model using leave-one-speaker-out validation. The BDC data contains speech produced by four speakers, thus, for each validation fold, we train on the material from three speakers, and test on the other. This produces four validation folds. The BURNC data contains six speakers, leading to a six-fold evaluation. The results of this evaluation appear below. Since the fold sizes are different, the reported mean and confidence intervals are calculated by weighting the contributions from each fold by the fold size.

Corpus	Energy	Corrected	Change
BDC-read	80.43 ± 2.837	82.61 ± 1.279	2.18
BDC-spon	80.99 ± 2.214	82.62 ± 2.066	1.63
BURNC	83.05 ± 0.918	84.95 ± 0.787	1.90

Table 3.17: *Leave-one-speaker-out evaluation*

Notice that the confidence intervals associated with this evaluation are much larger than those from ten-fold cross-validation. This is probably due to the reduced number of samples in the validation technique. While the confidence interval is constructed using the variance from ten folds under the cross validation technique, only four or six folds are available in this leave-one-speaker-out validation technique. This results in overlapping confidence intervals of the energy-based voting classifier and the corrected energy-based classifier. Therefore, we cannot conclude that that the correction significantly improves performance in a speaker-independent context from this evaluation. Using a proportion test, however, we find these differences to be quite significant: p=$2.237 * 10^{-6}$ on BDC-read, p=$6.78 * 10^{-4}$ on BDC-spontaneous and p=$1.57 * 10^{-10}$ on BURNC. Moreover, we are encouraged by consistent improvement on all speakers across all corpora when applying the correction classifiers.

In general, we find the pitch based correction to improve performance by approximately 1% under this isolated speaker configuration. The inclusion of correcting classifiers lead to a improvement of 1.90% on BURNC material, this difference approaches significance with p=0.0299. While this difference is not statistically significant on the BDC-spon and BDC-read corpora, the correction technique leads to consistently improved performance on both of these. That said, the wide confidence intervals on the spontaneous material, in particular, indicates that the accents of at least one speaker were quite accurately detected using a classifier trained on data from the other three. Interestingly enough, the energy based performance is slightly, though not statistically significantly, improved in the speaker independent evaluation. This suggest that, while the energy based features are largely speaker independent, it is the correction features – which are pitch and duration based – that show some speaker dependence.

We examine the pitch accent type distribution of hits – true positives – and misses – false positives – on the ten-fold cross-validation results and the leave-one-speaker-out results. Chi-squared tests comparing these distributions showed no significant differences between the two training scenarios. That is, while the performance is lower under the speaker independent modeling, it is consistently lower. The ability to detect particular accent types (tones) does not show more or less speaker dependence.

We compare this speaker independent – leave-one-speaker-out cross-validation – evaluation of the corrected energy-based classifier (CEBC) to two baseline speaker independent classifiers. The baseline classifiers are trained on a feature vector using the same pitch and duration features used in the training of the correction classifiers, and unfiltered energy features. For comparison, we include results from leave-one-speaker-out validation using both J48 decision trees – the classifier used in the combination model described here – as well as Logistic Regression classifiers. The results of this evaluation can be found in Table 3.18.

Corpus	J48	Logistic	CEBC
BDC-read	77.41 ± 0.794	76.62 ± 1.640	82.61 ± 1.279
BDC-spon	77.25 ± 0.829	80.78 ± 1.984	82.62 ± 2.066
BURNC	81.49 ± 0.471	84.43 ± 0.576	84.95 ± 0.789

Table 3.18: *Leave-one-speaker-out evaluation of baseline classifiers and the corrected energy-based combination technique*

Again, in interpreting these results we are faced with the problem of large confidence intervals due to the small number of samples used in the calculation of the mean. We observe overlapping confidence intervals on the CEBC and Logistic Regression classifiers on both BDC-spontaneous and BURNC corpora. Only on the BDC-read do we observe a clear, statistically significant improvement of the corrected energy-based classifier over the standard Logistic Regression when evaluated using leave-one-speaker-out cross-validation. On all corpora we find the highest accuracies to be obtained by the CEBC, however, these differences are not always statistically significant.

For comparison with syllable level approaches on BURNC, we convert these word-based predictions to the component syllables. We perform this conversion using the lexicon based transfer technique described in 3.4. The most common evaluation techniques used in previously published work are leave-one-speaker-out cross-validation and n-fold cross-validation using single-speaker training and testing using speaker f2b from the BURNC (cf. Section 3.2). For easy comparison with previous work, Table 3.19 contains evaluations of the corrected energy-based classifier at the word and syllable level evaluated using speaker-dependent ten-fold cross-validation, speaker-independent leave-one-speaker-out cross-validation and speaker-dependent single-speaker (f2b) evaluation.

Domain	Leave-one-speaker-out	Ten-fold	Single Speaker (f2b)
Word	84.95 ± 0.787	85.51 ± 0.197	85.00 ± 0.640
Syllable	86.57 ± 0.672	87.10 ± 0.279	87.94 ± 0.328

Table 3.19: *BURNC evaluation of Corrected Energy-Based Classifier performance on pitch accent detection at the word and syllable level using ten-fold cross-validation over all speakers, speaker independent and single speaker evaluation.*

As expected, the pitch accent detection accuracy is improved when evaluated at the syllable level. The corrected energy-based classification approach, with lexicon based transfer of predictions, yields pitch accent detection accuracy of 86.57% under speaker independent evaluation and 87.94% in the single speaker condition. The speaker independent accuracy is better than any previously published result on the BURNC data set whether evaluated on the word or syllable level. The single-speaker evaluation, however, performs worse than that published by Sun [194] who reported a pitch accent detection accuracy of 92.78%. This approach used AdaBoost with CART decision trees and acoustic and lexico-syntactic features. The approach using only acoustic information also performs better than the corrected energy-based classification approach, with accuracy of 89.90%. In addition to the classification approach, one major difference between the work presented here, and that in [194] is that Sun classified pitch accents into four classes – unaccented, high, low and downstepped. It is possible that this four way classification is able to produce better pitch accent results than the binary classification employed here. Also Sun's work defines a number of pitch features based on a derivation of 'underlying pitch targets' – features unexplored in this chapter. While we find pitch features to be weak indicators of pitch accent, this work suggests that more sophisticated feature engineering can be used to leverage discriminative information from the pitch contour.

In this section, we have presented a classifier combination technique building on the energy observations reported in Section 3.5. This technique uses filtered energy features to generate 210 pitch accent detection hypotheses for each word. Then, using pitch and duration features, each of these hypotheses are classified as ‘correct’ or ‘incorrect’. Predictions classified as ‘incorrect’ are inverted, and a weighted majority voting decision is used to combine these corrected predictions into a single pitch accent hypothesis. This acoustic pitch accent detection technique achieves state-of-the-art performance. On BURNC the speaker-independent accuracy is 84.95%, and the single speaker (f2b) accuracy is 85.00%. The best previously published results under these evaluations using only acoustic features are 80.75% [117] and 82.8% [44].

We experimented with using other classifiers within the classifier combination structure. We found decision trees to be particularly well suited to learning correction rules, and weighted majority voting to be the best prediction aggregation routine. We also performed evaluations to determine the robustness of these techniques to speaker and genre differences. The speaker independent evaluation did not reveal sufficient evidence of any speaker dependence in the performance of the energy-based classifiers. However, the correction classifiers, showed some speaker dependence, though this effect was not always statistically significant. We found some effect of genre or speaking style on the performance of the detector – with BDC classifiers performing worse on BURNC material and vice versa. However, when we evaluated the BDC-read material on BDC-spontaneous trained models, we found no significant effect of genre. These two corpora share similar recording conditions, speakers and were annotated by the same labelers. This suggests either that the professional broadcast news style of speaking is more different from non-professional read and spontaneous speech than they are from each other, or that the difference in performance is more due to differences in speakers, domain of the speeach, recording conditions or labeler idiosyncrasies across the BDC and BURNC corpora.

3.7 Using Part-of-speech tags in Pitch Accent Detection

Intonation, and accenting in particular, is an acoustic phenomenon. The lexical identity of a word does not determine whether or not it will bear an accent or not. However, certain classes of words are more likely to be accented than others. For example, nouns tend to be accented more frequently than prepositions, adjectives more than determiners. This follows from the pragmatic uses of accenting. Accenting is used to draw a listeners attention to a some aspect of the accented word whether new information or to indicate contrast or for some other effect. Nouns and adjectives may be a common source of contrast, and may be more likely to vary in terms of information status. Moreover, they are more likely to be the topic or focus of an utterance than function words such as “the”, “in” and “of”. Therefore, a system with syntactic information can hypothesize *a priori* if a noun is encountered that it is more likely to be accented than not, and vice versa for determiners. All techniques using syntactic information in prosodic analysis essentially encode this prior into the decision making process. When building a prosodic assignment module for a text-to-speech (TTS) system, for example, lexico-syntactic information and formatting, such as punctuation and capitalization are the only available input. Thus, the model is essentially only a lexico-syntactic prior. However, when analyzing intonation in speech, the prior determined by syntactic information must be integrated with acoustic information drawn from the speech signal.

In this section, we examine the use of part-of-speech (POS) information to improve the performance of acoustic pitch accent detection. We investigate three techniques for using POS tag information with acoustic features in training pitch accent detector. In the first, we construct a feature vector which contains both acoustic and POS-based features. In the second, we train two classifiers, one using POS-based features and another using acoustic features and merge their hypotheses. The final technique involves the training a distinct acoustic classifier for each POS tag. The automatic tagger, the Stanford Tagger [213], used in the experiments, annotates each word as one of 35 Penn Treebank [127, 19] POS tags. We find that using this full set of tags shows some evidence of data sparsity problems. Therefore, we also examine pitch accent detection performance using five techniques collapsing the full set of tags into broader part-of-speech based word classes.

In the context of a fully automated intonation analysis system, the presence of syntactic information makes two assumptions. First, that high quality lexical identity information is available, and second that this stream of lexical items can be tagged accurately. By using human transcriptions and word boundaries, we make the former assumption – that high quality lexical identity and word boundaries are available. The performance of automatic POS taggers suffers when processing spoken input, particularly when this input contains disfluencies as the BDC-spontaneous material does. By using an automatic tagger to generate POS tag annotations, these experiments avoid making overly optimistic assumptions about the quality and availability of high quality POS information.

3.7.1 Part-of-speech-based Word Classes

In order to avoid potential data sparsity problems that arise from using 35 POS tags, we examined a number of strategies of collapsing the tag set into fewer classes. We use two techniques that are based on the syntactic class of the POS tags. First, we collapse tags into six broad classes: *NOUN*, *VERB*, *ADJECTIVE*, *ADVERB*, *CARDINAL*, *FUNCTION*. A similar method of collapsing was used by Ross and Ostendorf [173]. The distribution of these tags, and associated accent rates can be found in Tables 3.20 and 3.21. Second, we collapse the tag set into two categories, function words and content words. Function words comprise a closed set in English that are used to indicate a grammatical relationship as opposed to having semantic content in isolation. These include, for example, determiners (“the”, ‘a’), prepositions (‘over’, ‘through’), modal verbs (“have” as in “I have seen X”). For these experiments we define function words as any word tagged with any tag other than, *NOUN*, *VERB* (not modal or auxiliary), *ADJECTIVE*, *ADVERB*, and *CARDINAL*. The distribution of these and associated accent rates can be found in Tables 3.22 and 3.23.

POS Class	BDC-read	BDC-spon	BURNC
NOUN	27.4 (2971)	27.5 (3202)	35.7 (10567)
VERB	15.2 (1642)	13.1 (1759)	16.1 (4751)
ADJECTIVE	4.3 (469)	7.5 (872)	7.4 (2186)
ADVERB	7.8 (847)	4.1 (473)	3.6 (1069)
CARDINAL	1.1 (123)	1.1 (126)	2.4 (706)
FUNCTION	44.1 (4779)	44.7 (5195)	34.8 (10299)

Table 3.20: *Distribution (%) of* broad class *POS aggregations.*

POS Class	BDC-read	BDC-spon	BURNC
NOUN	66.2	73.3	81.0
VERB	53.9	61.3	60.7
ADJECTIVE	79.5	84.4	84.9
ADVERB	65.9	69.2	79.9
CARDINAL	70.7	73.8	81.6
FUNCTION	14.9	23.8	14.0

Table 3.21: *Accent rate (%) of* broad class *POS aggregations.*

When examining the accent rate of the raw tags that are collapsed together under these two techniques, we found some outlying data points – tags with accent rates very different from the other tags included in the collapsed category. For example, the WRB tag – wh-adverbs, e.g. “**when** you get to Thayer Hall” – is accented 16% of the time, while other adverbs have an accent rate of 70%. Third person singular verbs, tagged as VBZ, have an accent rate of 29%, while other verbs are accented 57% of the time. This led us to reconsider collapsing POS tags based strictly on syntactic classes.

We explored three aggregations of POS tags based not on syntactic class, but on their accent rate in the training data. Each of these techniques identifies a threshold value and groups the POS tags into a set having an accent rate greater than or equal to the threshold value, and a set with an accent rate below the value. The first technique uses a static threshold of 50%. This technique is related to the *AccentRatio* feature defined in [139], but defined over part-of-speech tags as opposed to words. The second sorts the tags by their accent rate and then defines the threshold such that the two groups are as close to equally sized as possible. The third defines the threshold using the information gain criteria, as in the C4.5 algorithm [164]. The threshold is defined as the point which generates subsets of S, S_1 and

POS Class	BDC-read	BDC-spon	BURNC
CONTENT	55.9 (6052)	55.3 (6432)	65.2 (19279)
FUNCTION	44.1 (4779)	44.7 (5195)	34.8 (10299)

Table 3.22: *Distribution (%) of* function/content *POS aggregations.*

POS Class	BDC-read	BDC-spon	BURNC
CONTENT	63.7	70.2	76.4
FUNCTION	14.9	23.8	14.0

Table 3.23: *Accent rate (%) of* function/content *POS aggregations.*

S_2 such that $S_1 \cup S_2 = S$, $S_1 \cap S_2 = \emptyset$ and $H(S) - (H(S_1) + H(S_2))$ is maximized. This can be seen as a single node decision tree quantization of the pos tag set. For brevity, these three aggregations are referred to as *static*, *equal_size* and *info_gain*. The distribution and accent rates of these distributions based on the full training set are reported in Tables 3.24 and 3.25. However, when the pitch accent detector is evaluated using cross-validation, these aggregations are reconstructed for each training fold, and are therefore not strictly identical to what is reported in this table.

POS Class	BDC-read	BDC-spon	BURNC
static (+)	53.9 (5842)	55.3 (6429)	60.0(17758)
static (-)	46.1 (4989)	44.6 (5198)	40.0 (11820)
equal size (+)	47.1 (5100)	46.8 (5437)	46.1 (13626))
equal size (-)	52.8 (5731)	53.2 (6190)	53.9 (15952)
info gain (+)	43.5 (4707)	43.3 (5032)	42.6 (12605)
info gain (-)	56.5 (6124)	56.7 (6595)	57.4 (16973)

Table 3.24: *Distribution (%) of data-driven thresholded POS aggregations. (+) represent aggregations of tags with accent rates above the threshold, (-), those below.*

3.7.2 Acoustic Features

In the experiments presented in Section 3.7.3, we use a consistent set of acoustic features. These features are similar to the context-sensitive features explored in Section 3.3. As mentioned throughout this chapter, the major acoustic correlates of pitch accent are duration, pitch and energy. From the pitch contour, we extract the minimum, maximum, mean, and standard deviation within each word, as well as the z-score of the maximum. In addition to

POS Class	BDC-read	BDC-spon	BURNC
static (+)	66.0	71.8	79.3
static (-)	14.2	21.9	17.7
equal size (+)	68.0	73.9	83.1
equal size (-)	19.2	28.0	30.4
info gain (+)	68.8	74.7	83.4
info gain (-)	21.7	30.2	33.3

Table 3.25: *Accent rate (%) of data-driven thresholded POS aggregations. (+) represent aggregations of tags with accent rates above the threshold, (-), those below.*

these features extracted at the word level we calculate a set of features which incorporate acoustic information outside the word boundaries. For each word, we calculate the z-score of the mean and maximum pitch value relative to the pitch data extracted from six context regions. These context regions are defined as follows, 1) the previous word, 2) the following word, 3) the two previous words, 4) the two following words, 5) the two surrounding words, and 6) a window of four words surrounding the current word – two previous and two following. We extract these features from the raw pitch contour as well as its slope, and from a speaker normalized pitch contour, and its slope. Similar features are extracted from the energy contour. We include in the feature vector the duration of the word, as well as the word's duration z-score normalized by the context regions previously defined. (We do not speaker normalize word duration or extract features based on the delta of word durations.) We also extract the length, in seconds, of any preceding or following silent region.

3.7.3 Combining Acoustic and Word Class Information

In Section 3.7.1 we define six POS tag set aggregations: *raw* - no aggregation, *broad* - six classes based on syntactic class, *fc* - function v. content words, *static* - static thresholding, *equal_size* - equal sized class thresholding, and *info_gain* - information gain thresholding. In the experiments we describe here, we evaluate each of these tag sets in isolation. Moreover, we also examine unigram, bigram and trigram features of each of these. This leads to the evaluation of 18 POS tag configurations.

The most basic approach we take to incorporating POS tag and acoustic features is to simply extend the acoustic feature vector with POS tag based features. The first approach we take was to represent the POS tags as nominal features. In implementation these nominal features are expanded to N binary numeric features where N is the number of POS tags. When using bigram and trigram features especially, this leads to an explosion in the size of the feature vector resulting in both data sparsity and memory demand issues.

To make explicit the syntactic contribution, we train a set of binomial models using POS tags, $p(accent|pos, \theta)$. This tag model is applied in two ways. First, we include the posterior of the model in the feature vector along with the acoustic features. This extended feature vector is then used to train a Logistic Regression model. Second, we keep the models distinct and combine the likelihoods using Equation 3.2.

$$\arg\max_{accent} p(accent|pos) * p(accent|acoustic, \theta) \tag{3.2}$$

We train the acoustic model using Logistic Regression. Each of these experiments is evaluated using ten-fold cross validation, requiring us to retrain the tag model for each fold.

The final technique we explore to combine word class and acoustic information is via class-based modeling. For each word class, we train a distinct acoustic model. During evaluation, a test data point is first directed to an acoustic model based on its word class, the prediction of the system is the prediction returned by that word class-based acoustic model.

3.7.4 Results and Discussion

In this section we present the results of the experiments described in Section 3.7.3 using each of the aggregations defined in Section 3.7.1. All of the results reported in this section are based on ten-fold cross validation. The confidence intervals are based on the accuracies observed over the ten-folds. The BURNC material has a great deal of repeated lexical material. The same stories are read by multiple speakers, sometimes up to six times. To avoid any influence of this repetition on the evaluation, we guarantee that data from each story is represented in a single fold, regardless of its speaker. That is, no material from the same story ever appears in corresponding training and testing data.

Unigram POS tag experiments

In this section, we present the results of the experiments described in Section 3.7.3. Exhaustive results based on unigram POS tag sets are reported and discussed. These experiments were also run using bigram and trigrams of these same POS tags.

First, we evaluate the performance of the syntactic models in isolation, without integration with any acoustic information. We train a Logistic Regression model using nominal part-of-speech tag features. These nominal features are represented as a set of *N* binary numeric features, with each feature corresponding to a part of speech tag. A value of 1 indicates that the data point is annotated by the corresponding part-of-speech tag, 0 if it is not. The performance of this syntactic modeling approach is reported in Table 3.26. Across all corpora, the raw POS tag set shows some of the highest accuracy pitch accent detection. This indicates that the 35 classes represented here are not introducing data sparsity issues at the unigram level. Quite the contrary, the granularity of the classes is enabling this tag set to perform better than the broad and function/content aggregations. The data driven aggregations fail to perform significantly better than the raw POS tag set. However, they avoid the degradations observed in the broad and function/content aggregation performance. The static threshold aggregation, in particular, does not perform significantly differently from the raw POS tag set in any experimental configuration. This suggests that we have not lost

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	***75.0* ± *0.886***	***74.6* ± *0.476***	***80.2* ± *0.246***
Broad	73.1 ± 0.836	72.9 ± 0.853	79.7 ± 0.410
Fn/Content	73.1 ± 0.771	72.9 ± 0.820	79.7 ± 0.410
Static	***75.0* ± *0.607***	***74.6* ± *0.525***	***80.5* ± *0.426***
Equal Size	74.7 ± 0.410	72.8 ± 0.672	75.7 ± 0.295
Info Gain	74.0 ± 0.722	72.8 ± 0.574	74.6 ± 0.935

Table 3.26: *Accuracy (%) of Logistic Regression modeling using only nominal POS tags to detect pitch accent.*

any information with respect to pitch accent detection by collapsing POS tags with accent rates ≥ 50% and those with accent rates < 50%. The information gain derived thresholds lead to the poorest pitch accent detection rates of the three thresholding approaches.

Additionally, we evaluate the syntactic model implemented as a set of binomial models – $p(accent|POS)$. That is, for each POS tag, a binomial model representing the accent rate of training tokens annotated with this tag is constructed. At evaluation, the posterior of the model corresponding to a test point's POS tag is the likelihood that the point is accented. The performance of this technique is presented in Table 3.27.

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	74.7 ± 0.574	74.5 ± 0.541	80.5 ± 0.213
Broad	73.1 ± 0.426	72.9 ± 0.771	79.7 ± 0.197
Fn/Content	73.1 ± 0.623	72.9 ± 0.672	79.7 ± 0.361
Static	75.0 ± 0.476	74.6 ± 0.771	80.5 ± 0.426
Equal Size	74.6 ± 0.722	72.8 ± 0.525	75.6 ± 0.312
Info Gain	74.0 ± 0.672	71.2 ± 0.886	74.7 ± 0.836

Table 3.27: *Accuracy (%) of a set of binomial models.*

We observe that the two approaches do not significantly differ in their ability to model the relationship between POS tag and presence of pitch accent. All else being equal, this would lead to a preference of including the posterior of the selected POS model in the feature vector of combination with acoustic information. This would result in the inclusion of a single **numeric** feature, as opposed to a single **nominal** feature – which in the case of Logistic Regression (and many other classification techniques) will be expanded to $N > 1$ binary numeric features.

However, the nominal representation of POS tag information is clearly richer than the posterior of the mixture model. It is possible that a learning algorithm will have more success predicting pitch accents if it has access to this representation when acoustic information is included in the feature vector. The results of Logistic Regression classification experiments with a feature vector containing a nominal POS tag feature, and the acoustic features described in Section 3.7.2 can be found in Table 3.28. Results obtained by including the posterior of the mixture model instead of the nominal POS attribute are reported in Table 3.29. For comparison, the baseline Logistic Regression classification accuracy using only acoustic features has been included in these two tables.

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	84.0 ± 0.771	82.7 ± 2.13	84.2 ± 0.410
Broad	83.4 ± 0.640	82.8 ± 0.476	84.0 ± 0.476
Fn/Content	83.3 ± 0.476	82.6 ± 0.623	83.8 ± 0.410
Static	83.5 ± 0.476	***83.0 ± 0.525***	83.9 ± 0.220
Equal Size	***83.6 ± 0.410***	82.9 ± 0.262	***84.1 ± 0.0328***
Info Gain	83.3 ± 0.722	82.6 ± 0.623	***84.1 ± 0.476***
No POS	***83.6 ± 0.640***	82.4 ± 0.525	83.7 ± 2.13

Table 3.28: *Accuracy (%) of Logistic Regression with a feature vector containing acoustic features and a POS tag as a nominal value.*

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	83.5 ± 0.672	82.3 ± 0.476	84.8 ± 0.623
Broad	82.4 ± 0.623	***82.9 ± 0.623***	83.2 ± 0.312
Fn/Content	81.9 ± 0.410	82.7 ± 0.722	***85.3 ± 0.312***
Static	***83.6 ± 0.623***	82.8 ± 0.623	84.5 ± 0.6340
Equal Size	84.0 ± 0.623	82.8 ± 0.771	84.9 ± 0.410
Info Gain	83.3 ± 0.640	***82.9 ± 0.722***	85.0 ± 0.476
No POS	***83.6 ± 0.640***	82.4 ± 0.525	83.7 ± 0.672

Table 3.29: *Accuracy (%) of Logistic Regression with a feature vector containing acoustic features and the posterior of the tag mixture model.*

Though there is some fluctuation, the performances obtained by extending the acoustic feature vector with either mixture model posteriors and nominal POS tag features do not dramatically differ. This again supports the notion that the model posterior is capable of concisely representing the discriminative information contained in the POS tag data. On the BURNC data, inclusion of the model posterior yields superior results under all tag sets except *broad*. As these posteriors are necessarily simplifications of the nominal POS tag features, the most probably explanation for this is the ability of Logistic Regression to model this information. The increased number of features in the nominal value representation leads to a sparser feature space which can be harder to model. This phenomenon is not consistent, as we do not observe this effect in the results on the BDC subcorpora.

On BDC-read, the inclusion of POS tag information, either as nominal features or model posteriors, does not consistently improve pitch accent detection performance. The *static threshold*, *equal size* and *raw* tag sets show the best results, but these are at best 0.4% over baseline, while other tag sets fail to surpass the acoustic-only baseline. This is a surprising observation, as it would indicate the the discriminative information contained in the POS tags (as reported in Tables 3.26 and 3.27), is either captured by or is redundant with the acoustic information. The only acoustic qualities which might overlap with information represented by the POS tag are duration and pause information. Different parts-of-speech have different mean durations, and occur at differing rates before and after pauses. However, it is remarkable that these relatively basic acoustic qualities would subsume the informativeness of syntactic class labels. For the *function/content* distinction, this is consistent with findings of Batliner, et al. [12].

Experiments on the BURNC corpus show consistent improvement by the inclusion of POS tag information. In Tables 3.26 and 3.27, we see that the *static threshold*, *raw* and syntactic class-based tag sets can predict pitch accent at 80% accuracy, roughly 5% higher than is observed on the BDC subcorpora. When included in the acoustic feature vector as nominal features, these result in improvements from 0.1% up to 0.5%. As mentioned, this improvement is greater when mixture model posteriors are included in the feature vector, ranging from 0.8 to 1.6%. This improvement is not, however, observed when using the *broad class* tag set which fails to perform over baseline. This may be statistical noise; on all other tasks, performance tends to monotonically decrease from *raw* to *broad* to *function/content* tag sets. Moreover, only here does the *function/content* tag set outperform both the *raw* and *broad class* sets. These model posteriors interact with the acoustic features in a unique way on this material.

Disregarding for the moment the anomalous success of the *function/content* tag set on the BURNC data, we observe that the data-driven aggregations – *static threshold*, *equal size*, *info gain* – perform approximately as well the syntactic class based aggregations. These data-driven aggregations often perform as well or better than even the most informative, *raw* POS tag set. The most consistent of these is the *equal size* threshold. Recall that this method thresholds the accent rate of raw POS tags at a point such that there are approximately the same number of training points above and below the threshold point. By dividing the feature space into approximately equal sized regions, the *equal sized* threshold transmits POS tag information in a way the Logistic Regression algorithm can most easily take advantage of.

Next we evaluate an approach in which the syntactic models are combined with the acoustic model in a *post hoc* step. A Logistic Regression acoustic model is trained independently from the set of binomial models. At evaluation, the model likelihoods are combined using Equation 3.2. The results of these experiments, evaluated using ten-fold cross-validation, are reported in Table 3.30.

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	83.2 ± 0.623	82.9 ± 0.525	85.4 ± 0.361
Broad	83.1 ± 0.623	82.5 ± 0.525	85.1 ± 0.410
Fn/Content	83.0 ± 0.574	82.4 ± 1.476	84.9 ± 0.361
Static	83.1 ± 0.623	82.7 ± 0.525	84.8 ± 0.295
Equal Size	82.8 ± 0.623	82.7 ± 0.410	85.2 ± 0.262
Info Gain	83.0 ± 0.672	82.5 ± 0.410	85.1 ± 0.361
No POS	83.6 ± 0.574	82.4 ± 0.525	83.7 ± 0.672

Table 3.30: *Accuracy (%) of Syntactic and Acoustic model combination.*

In these results, we observe what is becoming a consistent trend. Prediction of pitch accent on the read subcorpus of the BDC does not benefit from the inclusion of POS tag information and often suffers due to it. The spontaneous portion shows modest improvement, while the BURNC shows definitive improvement. The difference in the BDC-read and BDC-spontaneous corpora is somewhat remarkable, as the two have nearly identical lexical content. Recall that the read material is transcription of the spontaneous material with disfluencies removed. The range of performance when combining syntactic and acoustic models is fairly narrow, between 0.4 and 0.6%. While the *raw* POS tag yields the best performance, this narrow range indicates that the contribution from the syntactic model is relatively consistent regardless of the specific tag set used.

Sridhar et al. [187] report a similar relationship in their pitch accent detection results. On the BDC corpus – they combine both the spontaneous and read data – they report that their acoustic model predicts pitch accent with 68.57% accuracy, and that their syntactic model predicts with 79.81% accuracy. Their combined performance however, is 80.01% an improvement of only 0.2% over the syntactic model. On the BURNC, however, their acoustic and syntactic models yield accuracies of 70.58% and 84.6% respectively and combine to 85.13%, a more substantial improvement of 0.53%. However, there is a methodological concern with the use of syntactic information in this paper. The evaluation is speaker-independent, using leave-one-speaker-out cross-validation. Using BURNC material with leave-one-speaker-out cross-validation leads to the presence of identical lexical material in the training and testing data. While spoken by different speakers, the material is completely identical. It is unclear if the authors have addressed this issue. If the test data contains lexical material found in the training data, the syntactic results are likely to be inflated.

The final technique we explore to combine POS tags and acoustic features is class-based modeling. Recall, in this technique, we train a distinct acoustic model for each POS tag. If different POS tags have different acoustic properties, this modeling approach should enable the training of better acoustic models. However, if they do not, the reduced amount training data will generate models which perform worse than a single acoustic model. The results of these experiments can be found in Table 3.31.

POS Tag Set	BDC-read	BDC-spon	BURNC
Raw	80.2 ± 0.722	78.8 ± 0.525	84.1 ± 0.312
Broad	83.5 ± 0.558	82.6 ± 1.968	85.8 ± 0.361
Fn/Content	84.1 ± 1.148	82.8 ± 0.213	85.2 ± 0.312
Static	84.1 ± 0.312	83.0 ± 0.558	85.6 ± 0.312
Equal Size	83.9 ± 0.525	83.0 ± 0.672	85.3 ± 0.213
Info Gain	84.1 ± 0.279	82.8 ± 0.623	85.1 ± 0.213
No POS	83.6 ± 0.558	82.4 ± 0.525	83.7 ± 0.689

Table 3.31: *Accuracy (%) of POS tag class-based modeling.*

This modeling technique generates the best performing pitch accent detecting classifiers. This is the only combination of syntactic and acoustic information to perform better than the acoustic-only baseline on the BDC-read subcorpus. Across all corpora, the *static* thresholding aggregation yields the best and most consistent performance. Using this aggregation, based on the training data accent rate, two models are trained, one with a prior indicating that a majority of the data is accented, and another with the the majority of the data being unaccented. This representation of POS tag information is sufficient to allow the technique to perform better than syntactic-class driven partitioning of the data. The only exception to this is *broad class* POS tags on the BURNC corpus. This partition yields the best performance on this data, at 85.8% accuracy. However, this tag set yields a more mediocre performance on the BDC data – even below baseline on BDC-read. This result, along with the fact that the *static threshold* tag set does not perform significantly worse, only 0.2% lower accuracy, on BURNC leads us to prefer the *static* tag set.

Across all corpora, the *raw* tag set class-based modeling performs the poorest. This suggests that there are acoustic similarities across raw POS categories, and the acoustic model's ability to predict pitch accents is diluted by the lack of available training data. This is not surprising. The raw POS tag set distinguishes, for instance, singular and plural nouns. There is no reason to expect these two syntactic classes to have different acoustic qualities with respect to accenting behavior. This is borne out by the fact that the *broad class* tag set performs better on all corpora. On the BDC corpora, the fact that the *function/content* tag set yields still better pitch accent detection indicates that there are acoustic similarities across words tagged as *NOUN*, *VERB*, *ADJECTIVE*, *ADVERB*, and *CARDINAL*. On the BURNC data, however, collapsing *broad class* tags to the *function/content* aggregation negatively impacts performance. This finding indicates that, on this data, there are significant acoustic differences associated with accenting between some of these syntactic word classes.

Bi- and Trigram Experiments

In Section 3.7.4, we presented the results of experiments exploring a number of techniques combining part of speech tag information and acoustic features using unigram POS tags. In this section, we present the results from these same experiments using bigrams and trigrams of the POS tag sets presented in Section 3.7.1.

The graphs in Figures 3.4, 3.5 and 3.6 contain the unigram, bigram and trigram results of Logistic Regression classification with only nominal POS tag features. On all three corpora, we observe a sharp decline in the raw POS tag set performance using trigrams. This is a clear indicator of a data sparsity problem. In general, the inclusion of bi- and trigram context does not drastically affect performance. On the BDC-read data, the *broad* aggregation is a notable exception to this observation. This tag set shows monotonic improvement with the inclusion of more context.

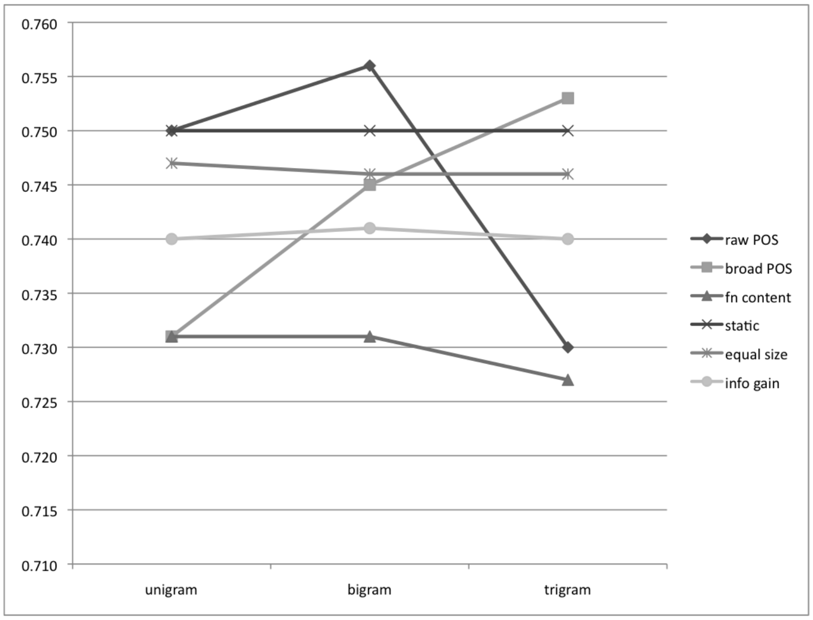


Figure 3.4: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag features on BDC-read.*

We again compare the performance of the Logistic Regression with nominal attributes to the likelihood of a mixture of binomial POS tag models. The results of the mixture model experiments with bi- and trigram features are presented in Figures 3.7, 3.8 and 3.9.

We see much the same trends in the model results as we do in the nominal attribute modeling. The raw POS tag class shows some evidence of data sparsity at the trigram level. However, overall, the impact of this is less drastic in the modeling scenario. Across all of

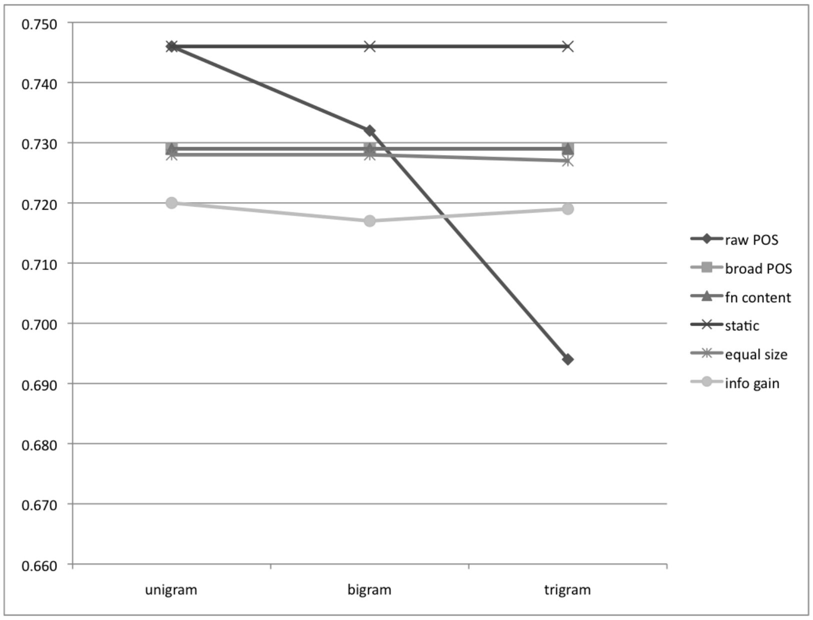


Figure 3.5: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag features on BDC-spontaneous.*

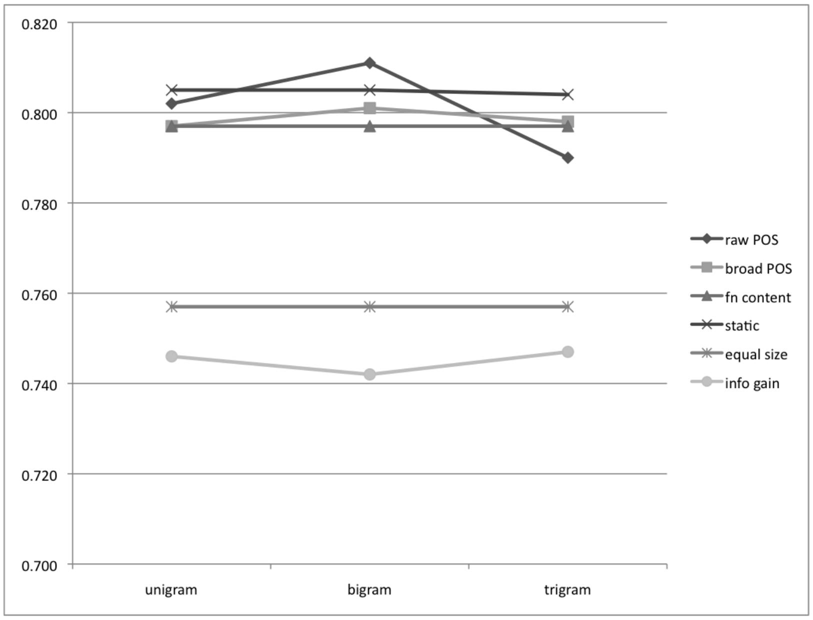


Figure 3.6: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag features on BURNC.*

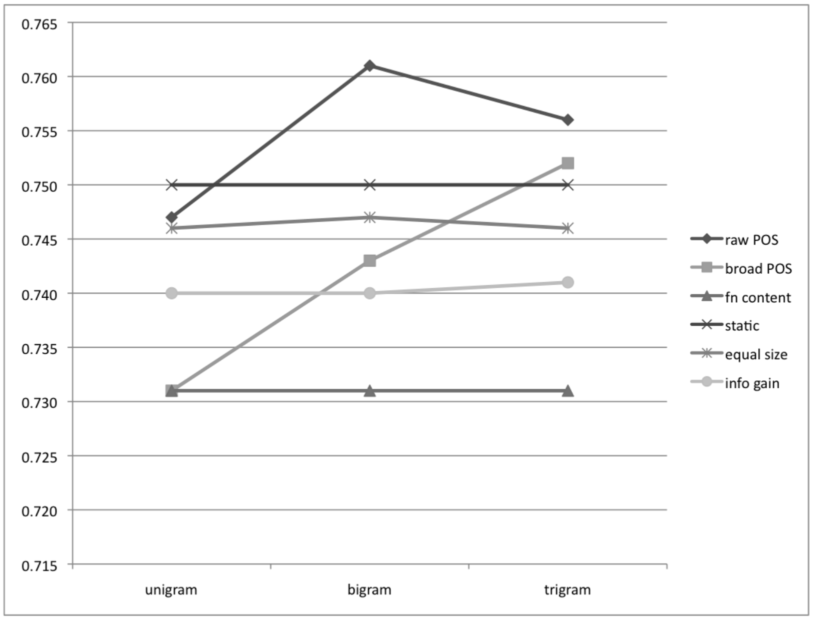


Figure 3.7: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag model posterior features on BDC-read.*

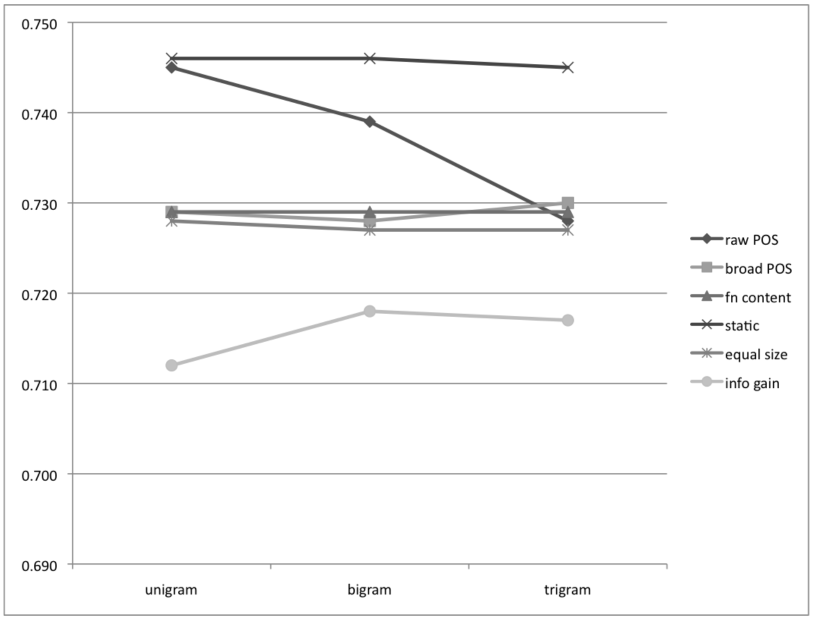


Figure 3.8: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag model posterior features on BDC-spontaneous.*

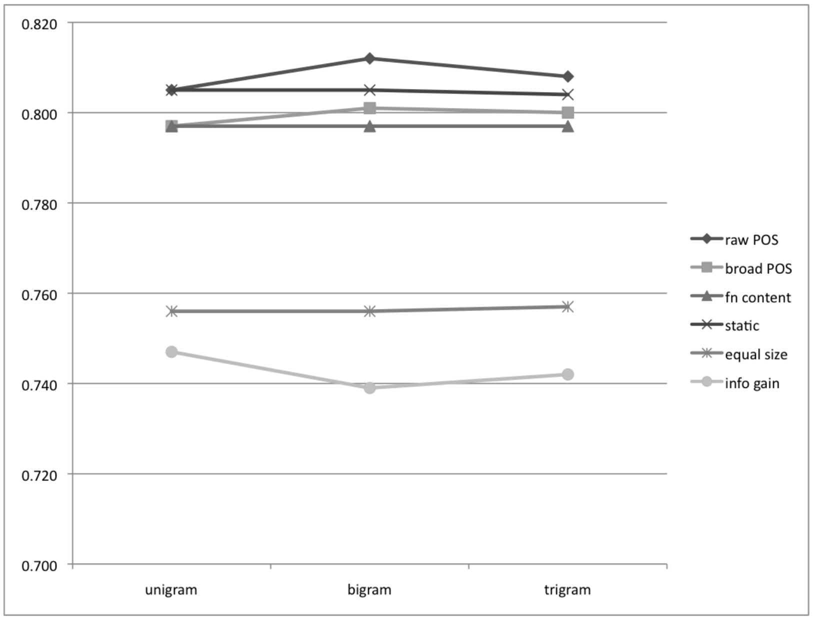


Figure 3.9: *Logistic Regression accuracy using nominal Uni-, Bi- and Trigram POS tag model posterior features on BURNC.*

these charts, the best performing features are bigram raw POS tags and *static threshold*. The raw tag bigrams have higher performance on BDC-read and BURNC, but have a data sparsity problem on BDC-spon. On the other hand, the static threshold tags are robust to corpus, performing successfully on all three.

Results from including bigram and trigram features and model likelihoods in the acoustic feature vector are presented in Figures 3.10, 3.11, 3.12, 3.13, 3.14, and 3.15.

When including nominal representations of bi- and trigram POS tag features in the acoustic feature vector, we observe a slight upward trend with increased context on the BDC sub corpora. The exception to this is again in raw POS trigrams. The data sparsity issue we observed before integrating POS information with acoustic features is manifested here as well. In the BURNC data, there is a trend for performance to increase with bigrams but *de*crease with trigram features. This would suggest that the data sparsity issues that may occur in the BURNC data are not so severe as to occur when examined in isolation, but only when the feature space is expanded, for example, with acoustic features. The *static threshold* aggregation is resistant to this sparsity issue, however, showing increased performance with

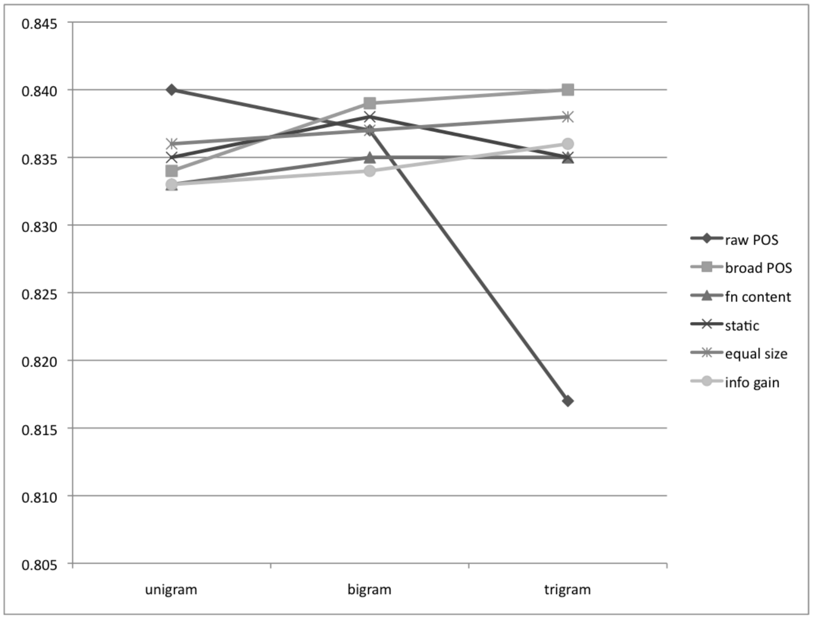


Figure 3.10: *Logistic Regression accuracy using acoustic features and nominal Uni-, Bi- and Trigram POS tags on BDC-read.*

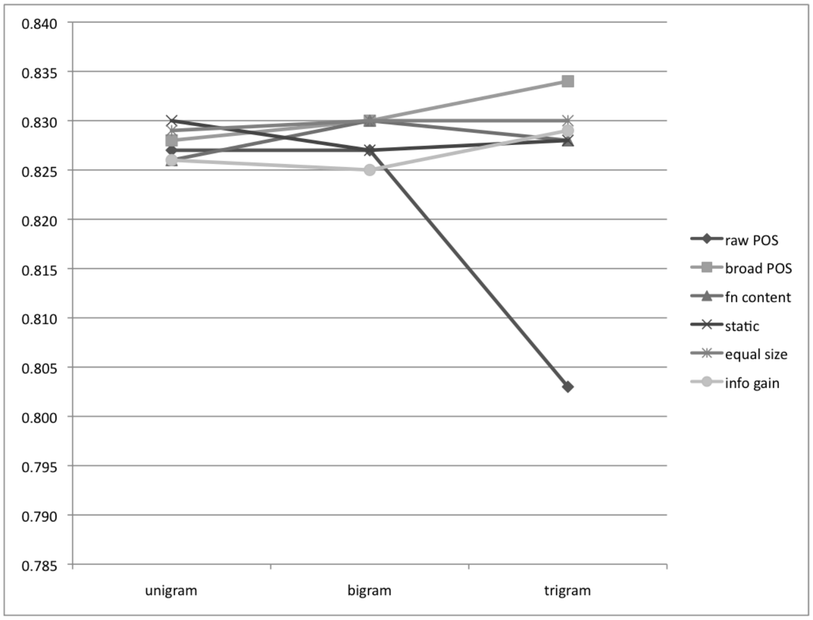


Figure 3.11: *Logistic Regression accuracy using acoustic features and nominal Uni-, Bi- and Trigram POS tags on BDC-spontaneous.*

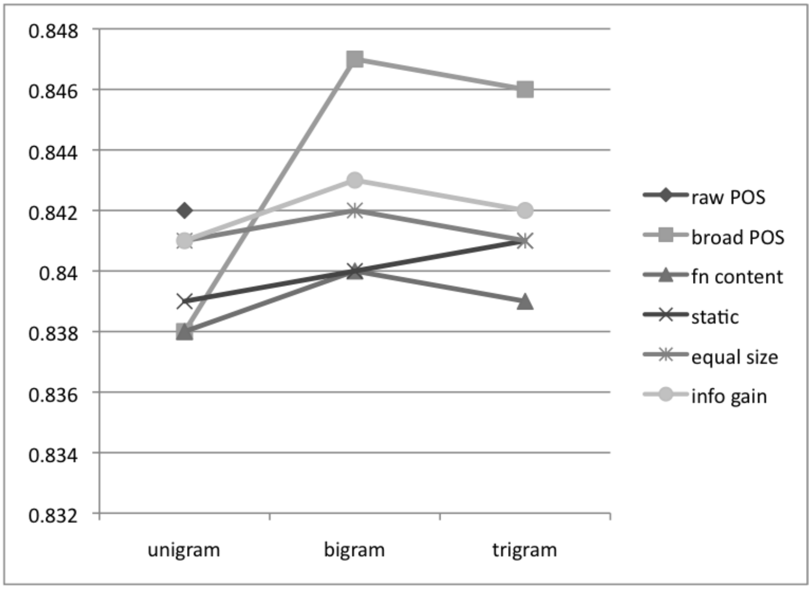


Figure 3.12: *Logistic Regression accuracy using acoustic features and nominal Uni-, Bi- and Trigram POS tags on BURNC.*

trigram features. The distributions of trigrams on this tag set are not notably more uniform than others; in fact, the entropy of the *function/content* tag set trigrams (2.94) is lower than that observed in the *static threshold* tag set (3.04). Note that the maximum entropy of these trigram distributions is greater than 3 despite being triples of binary features – the distribution includes trigrams that begin with a **null** tag for those data points which do not have previous or previous-previous tokens. This creates a trigram inventory of $12 = 2^3$ (standard trigrams) $+2 * 2$ (null tags) with a maximum entropy of 3.58. The performance of the *static threshold* tag set leads us to conclude that there is a unique relationship between this tag set and the acoustic features that allow this classifier to benefit from trigram features over bigrams.

One surprise in these experiments is the performance of the *broad class* tag set on BURNC. On the BDC data the performance of this tag set is relatively stable. However, on BURNC, we observe a remarkable increase in accuracy when including bigrams as opposed to trigrams. In isolation, the unigram and bigram performances do not significantly impact pitch accent detection accuracy (cf. Figure 3.6). In analyzing the results of the combination (cf. Figure 3.18) and class-based modeling experiments (cf. Figure 3.21), we do not see the same impact of *broad class* bigrams in other experimental scenarios.

We examine the tokens that were classified differently by the classifiers using unigram and bigram POS features. There are 1445 classifications in which the two classifiers differ, with 59% of these being correctly produced by the bigram classifier. Within these tokens that are classified differently, a chi-square test reveals an interaction between the accent type of the token, and the correctness of the bigram classification that approaches significance (p=0.07151). More of the NON-ACCENTED tokens that are classified differently by the unigram and bigram classifier are correctly classified by the bigram classifier (69.7%) than expected. In these cases, the unigram classifier is unable to correctly detect that a word is deaccented, but the bigram classifier can. A remarkable portion of these tokens, 40.7%, are the second token of NOUN/NOUN bigrams. In SAE it is common for the first noun in a compound noun to be accented. This result indicates that BROAD CLASS bigram information is able to identify these compound nouns, and significantly improve the detection of pitch accents on these tokens. The classification of L* tokens is also improved using the bigram classifier; 70.2% of L* tokens which yielded different classifications were judged correctly by the bigram classifier. This result indicates that while L* accents may be difficult to detect acoustically, access to more syntactic information can facilitate their detection. Also we observe fewer than expected tokens with an ambiguous tone, X*? (39.7%) were judged correctly by the bigram classifier. The use of X*? is quite labeler dependent. Without greater understanding of the use of this annotation, it is difficult to describe how the presence of increased POS context impacts the classification of these tokens.

Despite being more stable in isolation than the nominal features, when combined with acoustic information, the POS model likelihoods show more erratic pitch accent detection behavior. Based on the relationships observed in Figures 3.7, 3.8 and 3.9, we expect the raw tag set to exhibit performance degradation at the trigram level, but most other POS tag sets showed relatively consistent performance across inclusion of context. The exception

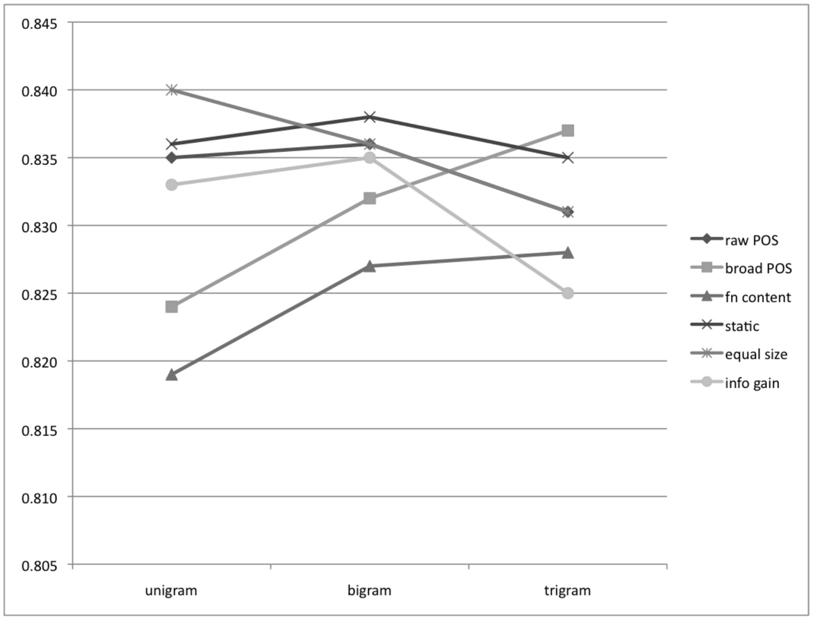


Figure 3.13: *Logistic Regression accuracy using acoustic features and mixture model likelihoods from Uni-, Bi- and Trigram POS tags on BDC-read.*

to this was the performance of the *broad class* set on the BDC-read subcorpus, which demonstrated monotonic increase with the inclusion of context. This is consistent with the inclusion of acoustic information. The BURNC, once again, demonstrates some surprising results, with the *raw* POS tag set improving with the inclusion of trigram likelihoods, and the *function/content* tag set degrading.

As observed in the unigram results, the extension of the acoustic feature vector with POS tag information only modestly (if at all) improves performance over acoustic only modeling on the BDC data. This effect is observed whether representing POS tag information as nominal values or syntactic model likelihoods. The inclusion of POS tag information in a feature vector along with acoustic features is somewhat more successful in detecting pitch accents on the BURNC material.

On the BDC corpus, using unigrams, *post hoc* combination of acoustic and syntactic models performed approximately equivalently to extending the acoustic feature vector with either nominal features or model likelihoods. On the BURNC, there was a slight improvement by model combination. When repeating the model combination experiments with bi-

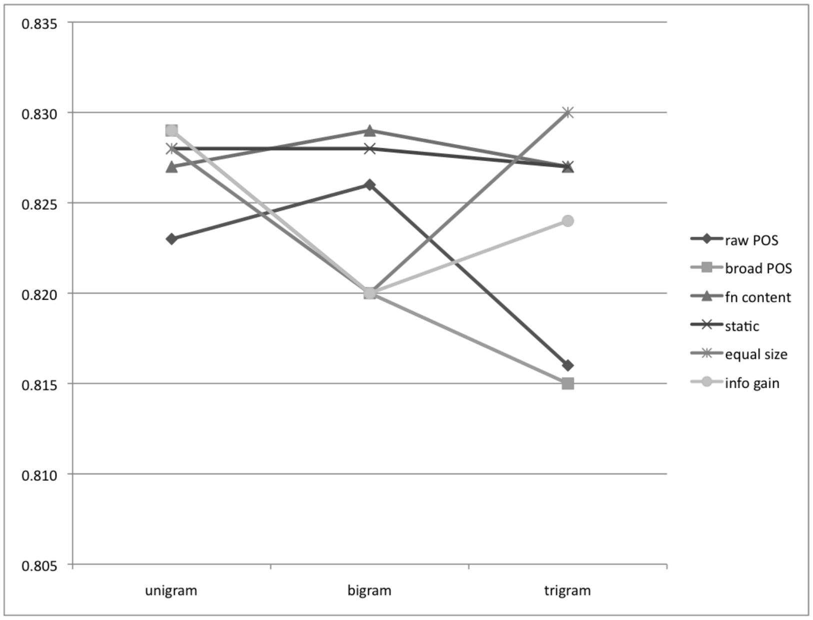


Figure 3.14: *Logistic Regression accuracy using acoustic features and mixture model likelihoods from Uni-, Bi- and Trigram POS tags on BDC-spontaneous.*

and trigram POS tags, across corpora and tag sets, we tend to see modest improvement in the bigram case, and either no improvement or degradation using trigrams. The one exception to this is that, in the BDC data, bigrams of raw POS tags show evidence of data sparsity problems. There is no evidence of this in the BURNC data. This is likely due to the fact that the BURNC is approximately 2.7 times larger than the BDC subcorpora, making it somewhat more robust to this issue. Results of model combination experiments using syntactic mixture models trained on bi- and trigram POS tags can be see in Figure 3.16, 3.17, and 3.18. On BDC-read the best performing model combination technique combines the acoustic model with a trigram *static threshold* statistical model, achieving an accuracy of 83.5%, still below the acoustic only performance of 83.6%. Regarding the BDC-spontaneous subcorpus, the best performing syntactic model, with accuracy of 83.0% over an acoustic-only baseline of 82.4% is trained on trigrams of *equal size* POS tags. As further evidence of the lack of data sparsity problems in the BURNC data, the best performing model combination uses a syntactic model trained on *raw*, unaggregated POS tag bigrams, detecting pitch accent correctly 85.8% over an acoustic-only baseline of 83.7%. This is 0.4% higher than the

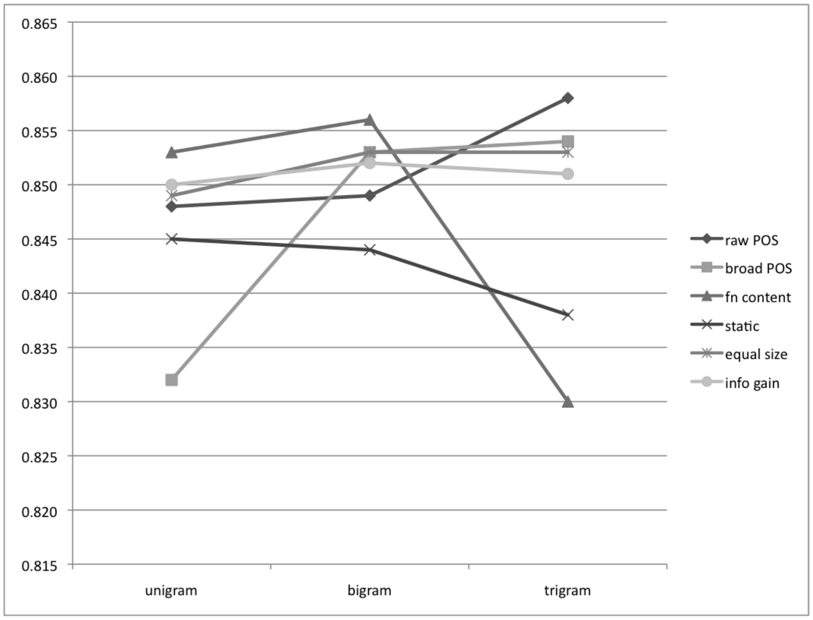


Figure 3.15: *Logistic Regression accuracy using acoustic features and mixture model likelihoods from Uni-, Bi- and Trigram POS tags on BURNC.*

next best model combination scenario, involving a syntactic model trained on *broad class* trigrams. This suggests that, given sufficient data, the information available in raw POS tags is more valuable than any aggregation. That said, with a tag set with cardinality of 35, the data requirements to avoid data sparsity problems in trigram modeling are significant.

Class-based modeling is the most successful strategy for incorporating syntactic and acoustic information at the unigram level (cf. Section 3.7.4). The results of extending these class based experiments to bi- and trigrams can be found in Figures 3.19, 3.20, and 3.21.

Here as in the unigram case, we find that the raw POS tags lead to poorer acoustic modeling, due to the reduced training data for the component acoustic models. This reduction in performance becomes still more pronounced in the bigram- and trigram-based modeling. Moreover, we see the same data sparsity problems appear in the *broad class* bigram- and trigram-based acoustic modeling. The remaining aggregations are all binary. This allows for some modest improvements under bigram-based acoustic modeling. In the trigram-based experiments, we tend to observe a reduction in pitch accent detection accuracy, but this artifact is less pronounced than we see in the larger *raw* and *broad class*

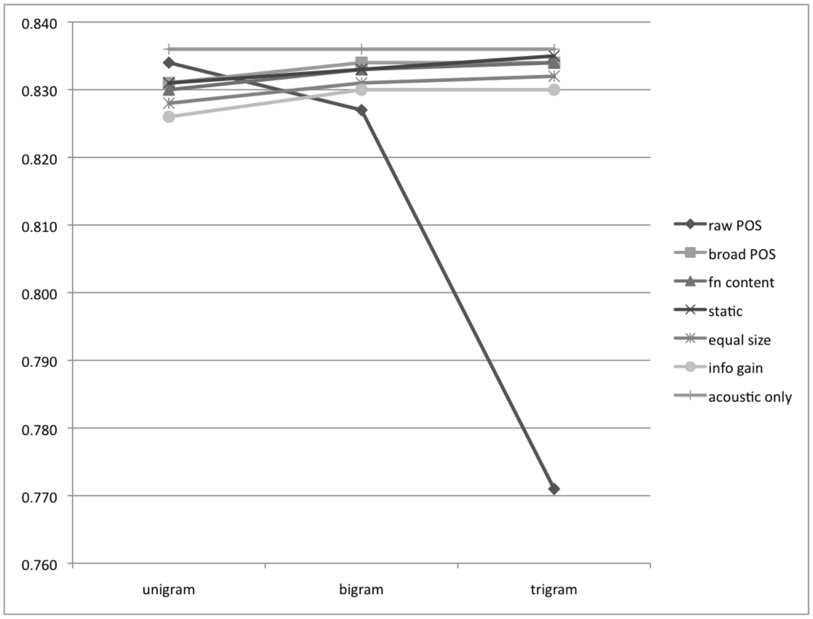


Figure 3.16: Post hoc *combination of acoustic and syntactic models using Uni-, Bi- and Trigram POS tags on BDC-read.*

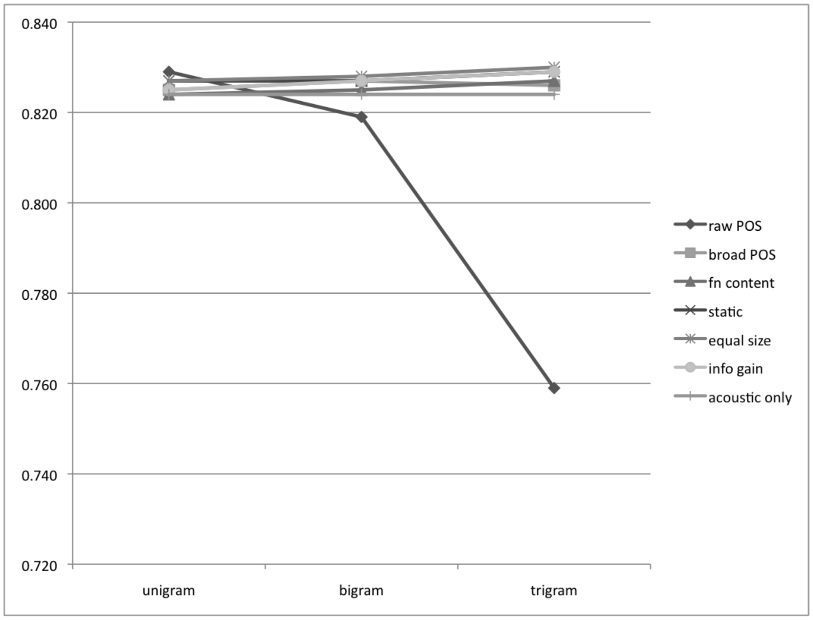


Figure 3.17: Post hoc *combination of acoustic and syntactic models using Uni-, Bi- and Trigram POS tags on BDC-spontaneous.*

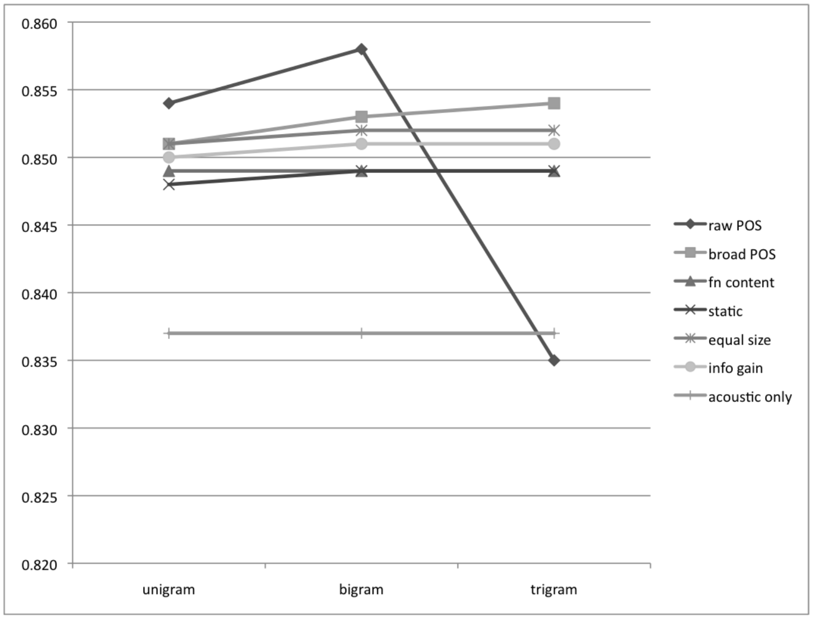


Figure 3.18: Post hoc *combination of acoustic and syntactic models using Uni-, Bi- and Trigram POS tags on BURNC.*

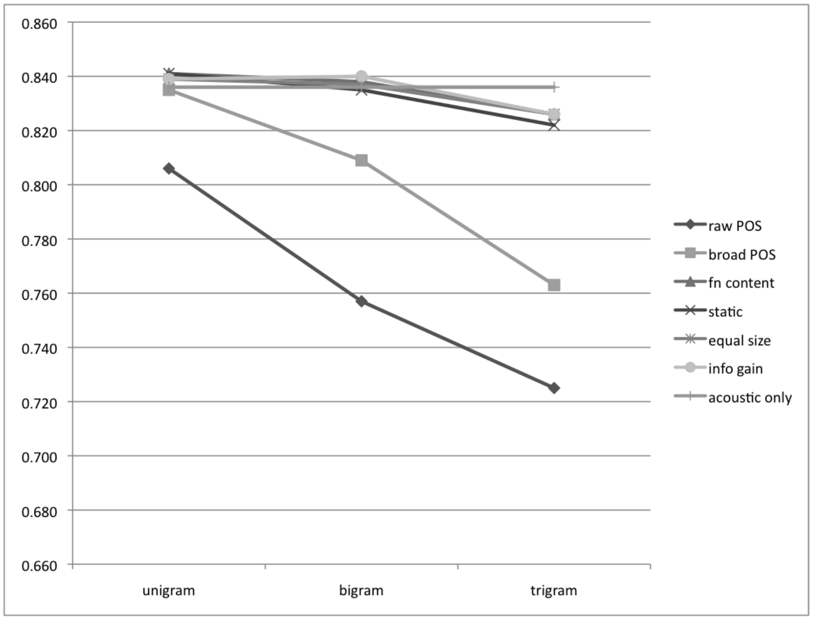


Figure 3.19: *POS tag class-based modeling using Uni-, Bi- and Trigram POS tags on BDC-read.*

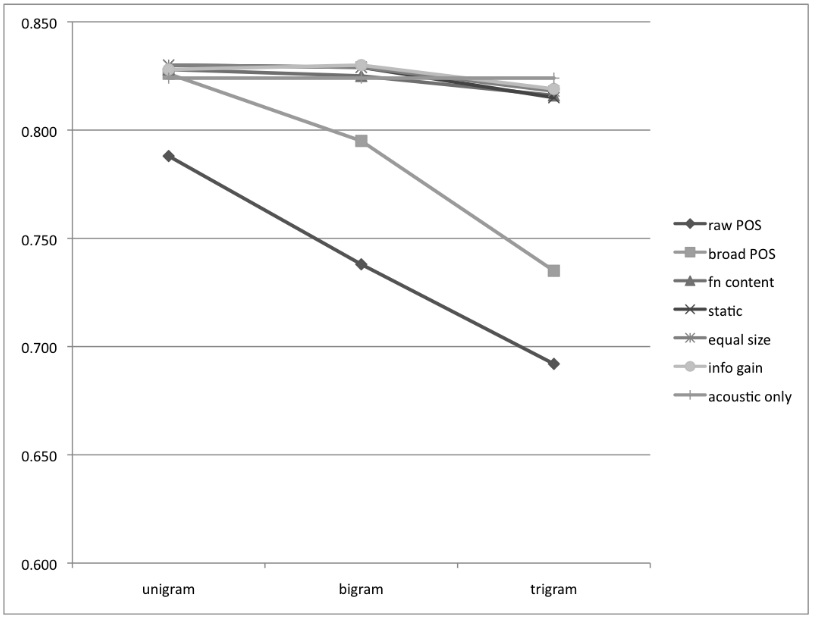


Figure 3.20: *POS tag class-based modeling using Uni-, Bi- and Trigram POS tags on BDC-spontaneous.*

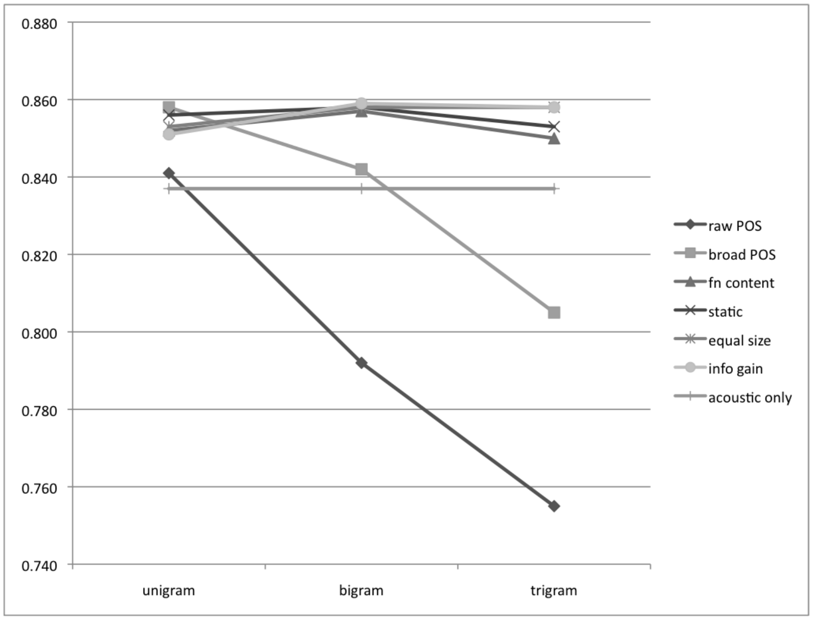


Figure 3.21: *POS tag class-based modeling using Uni-, Bi- and Trigram POS tags on BURNC.*

tag sets.

Overall, however, we do not see significant improvements from bigram-based modeling on the BDC subcorpora. The acoustic qualities that are discriminative to pitch accent are not distinct enough across bigrams of POS tags to significantly benefit from being modeled separately. On the BDC-read subcorpus, the best performing pitch accent detection scenarios use class-based modeling under either *function/content* or *static threshold* unigram tag sets. Both of these achieve accuracies of 84.1%. On the BDC-spontaneous subcorpus, the best preformance is achieved by class based modeling using *static threshold* unigrams or *information gain* bigrams, with accuracy of 83.0%. Only on the BURNC do we find bigram-based modeling to yield the best results, with an accuracy of 85.8% achieved by *information gain* bigrams. This provides some weak evidence that, given sufficient data, bigram-based modeling can lead to improved pitch accent detection. However, this difference is not statistically significant. This improvement on BURNC material indicates that the POS tag of a word, as well as its previous word, may dictate acoustic differences with respect to accenting behavior.

In summary, we find that approaches using unigram POS tags perform nearly as well as bigram POS tag representations, more or less regardless of the specific POS tag aggregation used. This leads us to hypothesize that, with approximately 10,000 data points, unigram syntactic representations are sufficient, if not preferred. When the size of the data set is increased to nearly 30,000 points, we start to observe improvements by the inclusion of bigram tags.

3.7.5 Evaluation of class-based modeling with corrected energy-based classifiers.

In Section 3.7.4, we observed improvements to pitch accent detection accuracy by incorporating syntactic information with acoustic models under a “class-based” modeling scenario. Based on the results of automatic part-of-speech (POS) tagging, we divided the training data into partitions based on tag classes. Then acoustic models are trained on the data points from each POS class separately. We find that this leads to improvements between 0.5% and 1.9% over Logistic Regression acoustic modeling without any syntactic information. We found that the best performing syntactic tag-set was a collapsing of the full Penn Treebank tag set into two broader tags. The broad tags are defined specifically for the accent detection task and can be understood as "likely to be accented" and "unlikely to be accented". Based on the training data, if 50% or more of training tokens with a given tag are accented, it is assigned the "likely to be accented" syntactic tag, while if less than 50% are accented, the "unlikely to be accented" tag is assigned. This is the *static threshold* word class tag set, defined in Section 3.7.1. In this section, we evaluate this class-based modeling scenario using the corrected energy-based classifier described in Section 3.6.

Evaluation of this syntactic class-based modeling on using this acoustic model can be found in Table 3.32. This cross-validation experiment uses identical fold assignments as the experiments described in Section 3.6, allowing for paired t-test analyses. We include the performance of the corrected classifier along with the performance based on the combination of unmodified energy-based predictions.

Corpus	Energy	Corrected	Change
BDC-read	80.79 ± 0.546	84.58 ± 0.407	3.79
BDC-spon	81.13 ± 0.592	83.63 ± 0.454	2.50
BURNC	84.19 ± 0.382	85.75 ± 0.148	1.56

Table 3.32: *Part-of-speech class-Based Corrected Energy Prediction Accuracy (%)*

By comparing the performance of class-based modeling with the standard corrected energy-based classifier reported in Table 3.6, we find that the corrected performance on all corpora shows small improvements that are not, however, statistically significant. On BDC-read the improvement is 0.20%; BDC-spon shows improved performance by 0.43 and BURNC material improves by 0.24%. When we examine the accuracy produced by the majority voting classifier based on uncorrected energy-based predictions, we find that this intermediate accuracy is improved by 1.01% on the BURNC data, an improvement which approaches significance with p=0.0117. On BDC-read and BDC-spontaneous, the uncorrected voting performances improve by 0.83% and 0.46%, respectively. These improvements are not statistically significant (p=0.0971 and p=0.359, respectively). Class-based modeling *may* improve both the energy-based classification as well as the ability to correct these energy-based predictions, though the observations of these improvements are not statistically significant.

3.7.6 Discussion of Syntactic Experiments

In this section, we explore six aggregations of POS tag annotations, in an attempt to incorporate syntactic information with acoustic features to improve automatic pitch accent detection. Three of these are based on syntactic class: *raw* POS tags (no aggregation), *broad class* (i.e. nouns, verbs) and *function/content*. The remaining three are based on the accent rate of POS tags. We aggregate POS classes into two clusters based on a threshold on the accent rate observed in the training data. The threshold is determined in three ways: *static thresholding* those over 50% and those under, *equal size* a threshold such that each cluster is approximately the same size, and *information gain* partitioning. We experiment on material from three genres of speech: non-professional read (BDC-read), spontaneous (BDC-spontaneous), and professional read (BURNC).

Experiments extending an acoustic feature vector with syntactic features are unable to consistently, significantly surpass the performance of acoustic-only modeling on any BDC data. However, the improvements on BURNC were fairly modest. We also run experiments training syntactic and acoustic models separately, and combining their predictions as a *post hoc* step. This technique was again modestly successful on BURNC BDC-spontaneous corpora, but unable to surpass acoustic modeling performance on the BDC-read data.

Finally, we present a modeling approach in which we train distinct acoustic models for data points from each POS tag class. This class-based modeling across corpora and POS tag aggregation represents a novel way of incorporating syntactic information along with acoustic modeling to detect pitch accent. We find this approach to yield the best pitch accent detection results, accuracies of 85.8% on BURNC, 84.1% on BDC-read, and 83.0% on BDC-spontaneous.

We then apply class-based modeling to the corrected energy-based classification technique described in Section 3.6. Using this more sophisticated model, we do not see any statistically significant improvement using this class-based approach to the incorporation of syntactic information.

We also investigate bigram and trigram contexts of these POS tag aggregations. We find that bigrams can generate modest improvements over unigram representations. Only on BURNC do these lead to an improvement over the best unigram performance. Using class-based modeling with *information gain* POS bigrams we observe accuracy of 85.9%, an insignificant 0.1% improvement over unigram *broad class* class-based modeling. Given the amount of data in the available corpora we find that trigram representations tend to show evidence of data sparsity problems. The increased context captured by the POS tag features is negated by the lack of data available to robustly model its relationship to accenting behavior.

Classifier combination, like that described in Equation 3.2, is a more flexible integration technique than class-based modeling. Class-based modeling requires an informed *a priori* partitioning of the data, like syntactic classes, to provide data to the acoustic modeling algorithm. However, classifier combination allows for independent development of multiple models, here, acoustic and syntactic models, as their results are combined in a *post hoc* phase. In the experiments described in this chapter, class-based modeling yields superior pitch accent detection than combination with a relatively simple syntactic model. It would be valuable to explore the use of combination techniques other than the argmax with a linear combination of model posteriors. The best way to integrate syntactic information with the correcting energy-based classifier remains an open question. HMM-based syntactic modeling as explored in [187], as well as a representation of information status like that developed by [81], could both be incorporated in a more sophisticated model to be applied in combination with an acoustic model.

3.8 Conclusion and Future Work

In this chapter we explore a number of techniques for automatic pitch accent detection. A summary of the results from select evaluations is presented in Table 3.33. Our investigation focuses mainly on acoustic based detection of pitch accent. We begin by exploring the main acoustic correlates of accenting – pitch, duration and energy. We find energy to be the most discriminative factor in the detection of pitch accent, confirming results reported by Silipo and Greenberg [182] and Kochanski, et al. [103]. In Section 3.3, we also examine the use of context in accent detection. Since accented words stand out from their surroundings, it follows that features that are able to represent the acoustic context surrounding a word would be able to demonstrate improved pitch accent detection performance. We confirm this hypothesis, finding detection accuracy to significantly rise with the incorporation of acoustic features that are normalized by their local acoustic context. On the BDC material, these improvements are observed on pitch, duration and energy domains. However, on BURNC material, only energy features show a sensitivity to this representation of context; feature sets using only pitch or duration do not show improved performance with the inclusion of contextual features.

As we have shown in Section 3.2 and summarized in Table 3.1, a great deal of previous work has used the BURNC material. This is quite helpful for comparing approaches using the same evaluation material. However, some approaches detect accents on the syllable level, while others determine if *words* bear accent or not. In Section 3.4 we address the impact of this decision. We find that word-based prediction yields improved performance regardless of whether the evaluation is carried out on the syllable or word level. Transferring syllable-based predictions to the word level is trivial – if a word contains a syllable that

Detection Approach	BDC-read	BDC-spon	BURNC
Logistic Regression Acoustic Modeling			
Pitch	77.46	75.78	73.50
Energy	80.08	80.50	82.50
Duration	75.72	77.23	79.73
All Features: No Context	81.16	80.41	83.35
All Features: w/ Context	83.45	82.88	85.01
Energy-based Voting Classifier	79.96	80.67	83.18
Corrected Energy-based Voting Classifier (CEBC)	84.38	83.20	85.51
Syntactic Detection			
Unigram: raw POS	74.7	74.5	80.5
Unigram: static threshold word class	75.0	74.6	80.5
Word class based Logistic Regression Acoustic Modeling			
Unigram raw POS tags	80.2	78.8	84.1
Unigram static tags	84.1	83.0	85.6
Bigram raw POS tags	75.7	73.8	79.2
Bigram static tags	83.5	82.9	85.7
Word class based CEBC Acoustic Modeling			
Unigram static tags	***84.58***	***83.63***	***85.75***

Table 3.33: A summary of selected pitch accent detection experiments. All evaluations use ten-fold cross-validation. The best performance on each corpus is indicated in bold.

is hypothesized to be accented, predict the word to be accented. Identifying the accented syllable within a word that is hypothesized to be accented is not so simple. We investigate a number of techniques for performing this word to syllable hypothesis transfer. However, none were able to achieve greater performance than hypothesizing that the lexically stressed syllable within the accented word bears the accent. In the BURNC material, abbreviations and compound words, like "school-based" are annotated as a single word, though they may bear multiple lexically stressed syllables. On this narrow set of tokens, we find that determining which syllable bears the accent by classifying the syllables within the accented word as accent-bearing or non-accent-bearing using a Logistic Regression classifier performs significantly better than the lexicon approach. As there are relatively few of these tokens, this improvement does not result in an overall significant reduction in classification error.

In Section 3.3, we observe energy to be the most useful acoustic predictor of pitch accent. Following the work on spectral balance, spectral tilt or high frequency emphasis by Sluijter, Van Heuven, Heldner and Fant [185, 186, 80, 78, 53], we examine the discriminative power of the energy contained in a variety of frequency subbands. Whereas previous work examined only four spectral regions, we identify 210 spectral regions and, using an analysis-by-classification approach, examining the use of the energy information from each region in automatically predicting accent. We find that, while there is a wide range of discriminative power in these filtered energy features, there is not a single "best" region. Moreover, we find that the predictive power of these features vary widely and inconsistently. On observing that more than 99% of all data points are correctly classified by at least one of these energy-based predictors, we turn our attention to using pitch and duration information – the other two acoustic accent correlates – to identify which predictor(s) contain the correct hypothesis. We construct a classifier combination technique where the energy-based predictions are corrected by pitch and duration based classifiers. Specifically, for each of the 210 energy-based pitch accent detectors we train a correcting classifier using pitch and duration features, that determines if a prediction should be trusted or not. If the correcting classifier predicts that the classification is incorrect, the prediction is inverted. Finally, the 210 corrected energy-based predictors are combined using a weighted majority voting decision.

This technique surpasses previous state-of-the-art word-based acoustic pitch accent detection performance. Under speaker independent evaluation, using leave-one-speaker-out cross-validation, the accuracy is 84.95%. When evaluated on the speech of the BURNC speaker for whom the most material is available with ten-fold cross-validation, f2b, the accuracy is 85.00%. The best previously published accuracies for these two tasks are 80.09% [187] and 82.8% [44], respectively.

There are a number of previously published techniques which use lexico-syntactic information in isolation for prosodic assignment, or in combination with acoustic features to significantly improve pitch accent detection performance. In Section 3.7, we investigate the use of part-of-speech tags for automatic pitch accent detection. We find that raw part-of-speech tags can be used to successfully predict accent location in isolation. However, when we combine these features with acoustic information, the impact, particular on BDC-read material, is rather minor. The experimental material was consistent and large enough to show improvement using bigram features, but when trigrams are used, data sparsity artifacts begin to emerge. We also investigate a number of ways to collapse part-of-speech tags into other word class sets. We explore collapsing the raw set of part of speech tags into broad syntactic categories (*NOUN*, *VERB*, etc.), as well as sets of *FUNCTION* and *CONTENT* words. In addition to these two, we also construct three data driven word class representations, where part-of-speech tags are grouped not by their syntactic function, but rather by the rate at which they are accented. These aggregations, in particular the *static threshold* aggregation, were able to predict pitch accent with modest accuracy, approximately 75% accuracy on BDC material, and 80.5% on BURNC data. Under this aggregation, tags whose tokens are accented more than 50% of the time are clustered together, as are those tags with less than 50% tokens bearing accent. Neither including these word class features with acoustic features, nor using a post-hot classifier combination approach, is able to significantly improve performance over acoustic based Logistic Regression pitch accent detection performance. However, if we construct acoustic models separately for each word class, we observe improved accent detection on all corpora over the baseline Logistic Regression acoustic classifier. Unfortunately, when we apply this class-based modeling approach to the correcting energy based classifier, the performance does not significantly differ.

The material presented in this chapter represents investigations into automatic pitch accent detection in three directions, and makes contributions to each. We investigated the use of filtered energy features for pitch accent detection, an investigation which led to the construction of a state-of-the-art pitch accent detection routine. The evaluation of syllable- versus word-based prediction indicates that word-based approaches tend to outperform syllable-based approaches, all else being equal. Moreover, this work presents some techniques for comparing word-based approaches to this task to syllable-based approaches. Finally, we look at the use of part-of-speech tag information for pitch accent detection. While the results of this work do not represent lexico-syntactic approaches to the task that rival the state-of-the-art, two novel ideas are explored. First, the data driven construction of word class aggregations is shown to be particularly helpful. Second, the word class based modeling represents a novel and potentially useful way to combine syntactic and acoustic information in a automatic pitch accent detection system.

In the future, it would be helpful to apply the lessons learned in these investigations to unsupervised or semi-supervised learning techniques. The resource requirements of manually labeling speech for intonation are quite high. The ability to leverage unannotated data to this task makes these learning techniques particularly attractive. Also, application of more sophisticated techniques using lexico-syntactic features for pitch accent detection still could be developed. Sridhar et al. [187] and others (e. g. [74]) have found success in constructing language models associating syntactic sequences with prosodic labels. It seems likely that the data driven word classes should be helpful in such a framework. On the acoustic side of the coin, the accuracy of the corrected energy-based classifier for pitch accent detection is already quite close to human performance. Because of this, it is likely that more significant gains can be made by focussing attention on syntactic and unsupervised techniques, rather than continuing development of this supervised acoustic technique.

3.8.1 Key Observations

- **Acoustic context is important.** Accents "stand out" from their acoustic surroundings. Therefore including acoustic context in a feature representation is critical in automatically detecting pitch accents.

- **Energy is a powerful an indicator of pitch accent.** The experiments presented in this chapter suggest that energy information is more discriminative to accent location than pitch information. However, energy and pitch information are not redundant; the best pitch accent detection approaches use features drawn from both streams.

- **Detection of accented words is preferred over detection of accented syllables.** While the acoustic realization of accents may be localized around lexically stressed syllables, the realization is not strictly contained within the syllable boundary; there is acoustic information that is discriminative to the presence or absence of pitch accent within a word outside the lexically stressed syllable.

- **Frequency filtered energy can be used by a voting classifier to reliably predict pitch accents.** Frequency-filtered energy, spectral tilt, spectral balance, high-frequency emphasis all capture the phenomenon that accented words show an increased amount of energy in high-frequency regions. While we do not identify a single representation that robustly detects this phenomenon, we find that frequency-filtered energy can be used in a classifier combination scenario to generate high accuracy pitch accent detection.

- **Hypotheses used in a voting classifier can be automatically corrected using a second classifier to improve overall performance.**

- **Merging part-of-speech (POS) tags into word-classes leads to improved pitch accent detection.** Moreover, we find word classes that are defined by the accent rate of the component tags to be superior to those that are syntactically defined (e. g. noun, verb, adjective).

- **Bigram word-class information may improve accent detection.** However, this appears to require data sets of approximately 30,000 tokens. On smaller data sets, bigrams do not significantly improve performance over unigrams.

Chapter 4

Phrase Boundary Detection

4.1 Introduction

Phrasing is the intonational mechanism that speakers use to break up speech into meaningful chunks of information [26]. Not only is phrasing useful for organizing the transmission of spoken information, it is physiologically necessary. Human speech is produced by vibrations produced by air passing through the vocal folds. Since human speakers do not have an endless supply of air, speech must be divided at a minimum when a speaker requires a breath. Rather than wait for this physiological necessity to dictate phrase boundary placement, human speakers initiate phrase boundaries at meaningful points, taking breaths at points where this disjuncture will not negatively impact the transmission of spoken information. But beyond the physiological necessity, speakers display acoustic indicators of phrasing even when not physiologically necessary, to serve syntactic or other organizational functions [151]. While physiologically influenced, intonational phrasing primarily serves a communicative, organizational function. For instance, intonational phrasing is frequently used to disambiguate and contribute to the intended meaning of an utterance.

Take for example the following utterance.

Bill doesn't drink because he's unhappy.

If the speaker inserts a prosodic phrase boundary, indicated by '—', after drink – "Bill doesn't drink — because he's unhappy" – the utterance carries the implication that Bill does not drink alcohol, and the reason for this is his unhappiness. Without the insertion of the phrase boundary, the implication is that Bill drinks alcohol, but for a reason other than unhappiness. This is an instance of the implication of a sentence having completely reversed polarity due to the presence of an intonational phrase boundary; the core proposition changes from BILL DRINKS, to BILL DOES NOT DRINK. This is an example of a broader function of phrasing. By dividing speech into units, phrasing strongly impacts interpretations of syntactic attachment. It is much more common for a modifier to attach to a token within a phrase, for example, than across a phrase boundary. In the case of the previous example, phrasing can convey "Bill doesn't drink" as a unit modified by the clause "because he's unhappy" leading to the interpretation that BILL DOES NOT DRINK. Without the phrase boundary, the negation, "doesn't" no longer modifies Bill's drinking, but rather the reason for his drinking, the clause, "because he's unhappy". Consider another example of the interaction between syntactic attachment and prosodic phrasing.

I saw the man with the telescope.

With no internal prosodic phrase boundary, the most likely interpretation would be that the man whom I saw had a telescope. The prepositional phrase "with the telescope" modifies "the man". On the other hand, if this utterance were produced with a phrase boundary following "man", "I saw the man — with the telescope", the desired interpretation is that I was looking through a telescope and saw a man. The presence of a prosodic phrase boundary leaves the prepositional phrase, "with the telescope" less likely to be a part of the noun phrase "the man with the telescope". Thus, "with the telescope" is more likely to attach to the head of the verb phrase “saw”.

The ToBI standard describes two levels of prosodic phrase boundaries corresponding to two levels of perceived disjuncture. The ToBI standard uses a phrase hierarchy where each intonational phrase is composed of one or more intermediate phrases, each of which contains at least one accented word. For the purposes of discussing phrase boundary detection, we distinguish between *intonational phrase boundaries*, and *intermediate phrase boundaries*. Intonational phrase boundaries refer to those points where an intonational phrase ends and another begins. While each intonational phrase boundary also represents the end of an intermediate phrase, the term “intermediate phrase boundary” will be used to refer to intonational-phrase-internal intermediate phrase boundaries. Intonational phrase boundaries are indicated by a break index of ‘4’ in the ToBI standard, and describe the greatest degree of disjuncture. Intonational phrase boundaries are frequently indicated by the presence of silence, and are differentiated from intermediate phrase boundaries by the presence of a final tonal event, called a boundary tone. Intermediate phrase boundaries, indicated by break indices of ‘3’ in the ToBI standard, represent a smaller degree of perceived disjuncture than intonational phrase boundaries.

Prosodic phrasing is described acoustically as the presence of “perceived disjuncture”. There are four major acoustic indicators of the presence of a phrase boundary: 1) the presence of silence, 2) pitch and energy reset, 3) pre-boundary lengthening and 4) changes in speaking rate across the phrase boundary. These all contribute to the perception of increased disjuncture at a word boundary. Silence is an obvious indicator of disjuncture; the signal stops for a perceptible amount of time. Pitch and energy reset describes the phenomenon where words are typically spoken louder and with increased pitch at the start of a phrase while pitch and intensity decrease over its duration. When describing pitch or fundamental frequency, this phenomenon is called “declination” [151]. This describes one aspect of the interaction between phrasing and pitch and intensity contour shapes. Pitch and intensity contours are influenced by more factors other than phrasing. Accenting (cf. Chapter 3) and phrase ending intonation (cf. Chapter 6) are both realized by modifications of the pitch and energy of the speech signal. Moreover, some phones are naturally produced with higher or lower pitch or intensity than others. This leads to changes of the pitch and intensity of speech due to segmental, as opposed to prosodic, variations.

As with all perceived phenomena there is some disagreement among humans concerning the presence of intonational phrase boundaries. Pitrelli et al. found ToBI labelers to agree at a rate of 93.4% concerning the presence of intonational phrase boundaries, and 89.8% on the presence of either intermediate or intonational phrase boundaries [156]. When labeling the BURNC material, labelers demonstrated 95% agreement on all break index values. Syrdal and McGory [196] found similarly high agreement (over 90%) regarding level '4' break indices – intonational phrase boundaries – though closer to 50% agreement on break indices of '3'. As we discuss in Section 4.3, accuracy can be a misleading measure of human agreement on phrase boundary placement. Since, roughly four in five word boundaries are not phrase boundaries, a phrase boundary labeler (human or automatic) which never predicts an phrase boundary will agree with an expert over 80% of the time. Thus, what looks like "high" accuracy is attainable without identifying any phrase boundaries correctly. F-measure is a measure used in information retrieval to evaluate how reliably uncommon events are discovered [216]. F-measure of a class i ($F_{i\beta}$) is calculated as the harmonic mean of "precision" and "recall'. Its formula can be found in Equation 4.3 where M_{ij} is a count of the number of times a token of class i is classified as class j. The calculation includes a tuning parameter, β, which can be used to bias the term towards Precision – what percentage of tokens labeled as the positive class are correct – or Recall – what percentage of positive class tokens are classified as positive by the classifier. A β value of 1 evenly weights these two measures; β values greater than 1 are biased towards Recall, while those less than 1 are biased toward Precision.

$$Precision_i = \frac{M_{ii}}{\sum_j M_{ji}} \tag{4.1}$$

$$Recall_i = \frac{M_{ii}}{\sum_j M_{ij}} \tag{4.2}$$

$$F_{i\beta} = \frac{(1+\beta^2) * Precision_i * Recall_i}{\beta^2 * Precision_i + Recall_i} \tag{4.3}$$

F-measure, F_1, is a more reliable measure of the performance of detecting rare phenomena [216]. Koehn et al. found the F_1 of the presence of intonational phrase boundaries between two expert labelers to be 0.930 [104].

This chapter addresses the use of acoustic and syntactic information in the automatic detection of prosodic phrase boundaries. The majority of these experiments concern the detection of intonational phrase boundaries. In Section 4.3, we investigate acoustic correlates of phrase boundaries. Syntactic indicators are explored in Section 4.4. In Section 4.5, we apply the techniques developed in the automatic detection of intonational phrase boundaries to the detection of intermediate phrase boundaries. We conclude and describe future work in Section 4.6.

4.2 Related Work

Intonational phrasing is critical in structuring spoken information. A significant amount of research has focused on techniques to identify intonational phrase boundaries from speech, as a way to access this structural information. Also, unexpected or erroneous phrasing decisions can severely degrade the naturalness and intelligibility of synthesized speech. Text-to-speech (TTS) systems typically include a prosodic assignment module which is responsible for performing lexical analysis and assigning either prosodic event locations or f0, energy and duration targets. Many prosodic assignment techniques can be used in the analysis of spoken intonation. If a hypothesized word stream is available, along with the speech signal, prosodic assignment can be used to generate syntactic hypotheses for phrase boundary locations. This, in effect, represents a prior on the likelihood of a phrase boundary occurring at a given word boundary based on the lexical content. The acoustic information carries the perceptual qualities that determine if a phrase boundary is present or not, while the lexical and syntactic information in the lexical stream can contribute by indicating how likely, in fluid, natural speech, a phrase boundary is to occur. In this section, we provide a survey of previous approaches to automatic phrase boundary detection, first discussing approaches which use acoustic information, and then present some of the prosodic assignment techniques that have been applied to this task.

In an early effort to detect intonational phrase boundaries, Wightman and Ostendorf [230] used decision trees with acoustic features as input to an HMM. Their approach performs a four-way classification, simultaneously detecting pitch accents and intonational phrase boundaries. This approach used decision tree output as input to an HMM classifier to predict the location of pitch accents and intonational phrase boundaries. The system operated at the syllable level and used features including the length of following silence, the presence of an audible breath, acoustic features such as the mean pitch and energy, and the ratio of these values across syllable boundaries. The system also used a feature capturing pre-boundary lengthening and changes in speaking rate. The difference in syllable duration of the mean of three syllables prior to a candidate boundary and three syllables following was used to capture changes in speaking rate. To detect preboundary lengthening, syllable rhyme durations were normalized by phone and speaker identity [231]. Analyzing the speech of a single BURNC speaker (f2b), this approach achieved phrase boundary detection accuracy of 94.15% with an F-measure (F_1) of 0.762.

Batliner et al. [11] in a similar study used a number of acoustic indicators of prosodic phrasing to predict phrase boundaries in German. However, the detected phrase boundaries were syntactically determined; a set of hand written rules described the relationship between syntactic structures and the location of ground-truth phrase boundaries. A three-way classification yielded an accuracy of 81%. The success of acoustic features to predict syntactic boundaries demonstrates the close association between syntactic and prosodic phrasing – at least in German. The acoustic features explored in this paper included length of subsequent pause, if any, and syllable and nucleus durations, normalized by phoneme class of syllable nucleus. To capture pitch and energy reset, they extracted the maximum energy within two syllables surrounding boundary as well as the location of this maximum. To model the reset associated with pitch declination, linear regression coefficients of the pitch contour over 2 and 4 syllables on either side of the boundary, were also calculated. This use of linear regression inspired the set of acoustic features to detect acoustic reset that we investigate in Section 4.3.1.

Chen et al. [38] combined a neural network syntacic-prosodic model and gaussian mixture model (GMM) acoustic model to detect phrase boundaries on BURNC material. Using leave-one-speaker-out cross-validation they achieved 93.07% accuracy. Evaluated in isolation, the syntactic models generated 90.09% accuracy, while the acoustics yielded only 68.15%. However, we again note one of the difficulties in the evaluation of syntactic features on the Boston University Radio News Corpus (BURNC). The BURNC consists of a small number of stories, repeated by many speakers. While this repetition allows the investigation of prosodic variation across speakers, care must be taken in the evaluation of lexical features. While not completely determined by lexical content, there is a strong correlation between syntactic and semantic information and prosodic phrasing and accenting. Therefore, to avoid modeling lexical, syntactic or semantic qualities unique to these repeated stories, it is critical to avoid the presence of repeated lexical material in training and evaluation data sets. This, unfortunately, makes speaker independent evaluation of lexico-syntactic features impossible on this corpus, as the material spoken by each speaker is also present in the corpus produced by another. Therefore, in this work, and others, it is difficult to evaluate the success of the syntactic-prosodic model in the detection of intonational phrase boundaries using a the leave-one-speaker-out validation approach. This problem undermines the evaluation of other studies using lexico-syntactic features in the prediction of prosodic events on BURNC material including [187, 189] and [5].

Yoon [236] predicted phrase boundaries on a subset of BURNC using memory-based learning (TiMBL). This algorithm was applied to a feature representation consisting of the numbers of phones, syllables, the presence of lexical stress, as well as part-of-speech (POS) tags, syntactic chunking information, named entity tags, and argument structure class – subject, object, predicate. This approach achieved an accuracy of 92.23% and F_1 of 0.861 in detecting intonational phrase boundaries. A 3-way classification – intonational, intermediate and no phrase boundary – accuracy of 88.06% was also reported with an intonational phrase boundary F_1 of 0.832 and an intermediate phrase boundary F_1 of 0.345. This represents the best reported performance on intonational and intermediate phrase boundary detection, though direct comparison to other approaches is difficult due to different training and test splits. While this work acknowledged that the evaluation material comprises multiple rendition of four news stories, it is unclear if stories in the test material are also present in the training data. This poses a potentially serious problem in the evaluation of this technique.

Many techniques of phrase detection operate on manually annotated or hypothesized word or syllable segmentations. However, some short frame based approaches have been explored. The advantage of these techniques is that by functioning similar to automatic speech recognition (ASR) systems, their processing can be more easily combined within a single module. Hirschberg and Nakatani [88] detected intonational phrases using at 10ms frames pitch values, a binary voicing feature, and rms energy including context of up to 27 frames, or 270ms, speaker normalization and delta values. Using CART trees to predict intontional phrase boundaries they achieved a mean F_1 of 0.697 and accuracy of 83.6% on the Boston Directions Corpus, where both the read and spontaneous material were analyzed as a single data set. Sridhar [187] also used a short (10ms) frame acoustic model in the detection of intonational phrase boundaries. Using f0 and rms energy values, along with delta and delta deltas, in a maximum entropy model, detection accuracy of 84.1% was reported on material from four BURNC speakers (f1, f2, m1, m2), and 83.53% on BDC material from both the read and spontaneous subcorpora. Under leave-one-speaker-out cross-validation, incorporating a syntactic model using part of speech tags and supertags with HMM acoustic models, accuracies were improved to 92.91% on BURNC, and 90.58% on BDC.

The next group of approaches to phrase boundary detection only using text-based information. Many of these have a similar flavor. A typical approach is to train syntactic-prosodic models to learn associations between text analysis and phrase boundary locations that have been identified from transcribed speech by hand. A wealth of research has been done on this topic, and the work described in this section represents only a portion of this. We highlight those papers that are most foundational to the task and those that specifically inform the work presented in this chapter.

In an early effort towards phrase boundary assignment for text-to-speech synthesis Bachenko and Fitzpatrick [7] defined hand-written rules for identifying candidate boundaries, and subsequently identifying the "salience" of these. The salience of a boundary was determined to be either strong – intonational – or weak – intermediate. While these rules operated over syntactic information they make no claims about a one-to-one correspondence between syntax and prosodic phrasing. "Prosody rules refer to syntactic structure, but they are not obliged to preserve it; independent principles of prosodic well-formedness, in particular, length calculations, may crease entirely different structures that appear at odds with the syntax" [7]. Using a very small evaluation corpus of 35 sentences, their system predicted spoken intonational phrase boundaries with 80% accuracy.

Veilleux's thesis on the interaction between prosody and syntax [217] began by examining single models to predict phrase break locations from syntactic information. Output from these models were then used as input to a "hierarchical modeling" technique. In effect this approach first predicted minor – intermediate – phrase boundaries, then uses these hypotheses to predict major – intontational – boundaries. Overall accuracy of nearly 93% was reported by applying this technique to a relatively small, less than 400 word, data set. This technique informs the approach examined in Section 4.5. This approach was also presented by Ostendorf and Veilleux [145]. Collapsing over both phrase types, 81% correct predictions with only 4% false predictions are reported. While the hierarchical framework presented can employ any statistical modeling technique, these results used a Markov process over decision tree outputs.

Decision trees are a very common approach to phrase detection. They are an attractive machine learning technique for this task as they train quickly, and their decision process is very comprehensible. Moreover, they represent a non-linear modeling technique and are well-suited to modeling nominal values – like part-of-speech, or syntactic constituent categories. Wang and Hirschberg [222, 223] applied CART trees to a feature vector containing parse tree and phrase positional features along with POS tags. This approach was able to achieve accuracy over 90%. Hirschberg and Prieto [90] extracted part-of-speech tags and morphological category information from a symmetric four-word window surrounding each word boundary. Using these nominal features along with numerical features representing the number of words and syllables in the utterance, the relative position of the current boundary, distance from punctuation and location within a noun phrase (NP), they found CART trees to be able to assign intonational phrase boundaries with 95% accuracy on English material and 94% in Spanish. Koehn et al. [104] extended this approach by including syntactic parse tree features, improving accuracy by 1.8% and F_1 of intonational phrase (IP) boundaries by 0.047. Using these and other syntactic parse features, along with supertags [9], Hirschberg and Rambow [91] use the Ripper, rule-induction algorithm to assign intonational and intermediate phrase boundary locations from text. On read Wall Street Journal text, they were able to achieve an F_1 of 0.701 on intonational phrase boundary detection with 89.8% overall accuracy. Collapsing intermediate and intonational phrases, the F_1 increases to 0.873 with an 86.1% accuracy. Syntactic parse tree features similar to those used in [104] and [91] were examined by Read and Cox [166]. These features treat a parse tree as a graph, representing numerical *syntactic distances* between two words by the graph distances between their corresponding parse tree nodes. We evaluate a set of syntactic distance features calculated using this representation in Section 4.4.

Most speech synthesis systems operate by generating speech output given an input utterance as text. McKeown and Pan [132] examined the task of prosodic assignment in a concept-to-speech (CTS) system. CTS systems have access to greater semantic information than a typical TTS system, and this information can thus applied to the task of prosodic assignment. By using features based on lexical content, concepts, syntactic functions, semantic boundaries, the length of the current "semantic constituent" as well as part-of-speech and Information Content (IC), a representation of the "semantic informativeness" of a word, they applied memory based learning to the task of assigning ToBI labels to CTS output. They achieved an accuracy of 78.69% in the assignment of all levels of break indices – a five-way classification task.

There has also been a significant amount of research in detecting phrase boundaries in languages other than English. Some correlates to phrasing are likely to be language independent. For example, the presence of an audible breath, or silence, most likely indicates a prosodic phrase boundary in all languages. Also, speakers typically insert prosodic phrase boundaries at the end of a sentence. This phenomenon is not unique to English. However, other acoustic and syntactic factors which indicate intonational phrase boundaries may differ across languages in both degree and dimension. While research in other languages can be used to identify areas for investigation or feature engineering techniques, there is no expectation that approaches that are successful in one language will *necessarily* be successful in another.

Heldner and Megyesi [79] exploited preboundary lengthening features, representations of speaking rate changes and a variety of word-class features and the presence of silence to detect intonational phrases in Swedish radio interview speech. Using discriminant analysis, they achieved a 3-way accuracy of 85.3%. In this study, lexical features performed better than acoustic information in the detection of "strong boundaries" – intonational phrase boundaries. While acoustic features were more helpful for detecting weak (intermediate) phrase boundaries. While, we believe germanic languages such as Swedish are similar to

SAE regarding prosodic variation, this conclusion has not been confirmed on SAE material.

On German and English material, Schmid and Atterer [177] applied a Hidden Markov Model over POS tags to the task. They reported an intonational phrase boundary F_1 of 0.778 on MARSEC data (English), 0.8023 on German read newspaper data and 0.853 on German Radio News Data using a syllable level of analysis. They found syllable based analysis to outperform word based analysis when measured by F_1. They model the part of speech of at a syllable level by inserting a DUMMY tag at word internal syllables. The fact that this syllable level representation performs better than a corresponding representation of words suggests that the modeling of syllable length helps detect phrase boundaries. These sequences of DUMMY POS tags corresponds to number of syllables per word.

Nakai et al. [137] used memory based classification with clusters of f0 contours to detect phrase boundaries in Japanese. Following the Fujisaki [59] superpositionlal intonation model, accent and phrase components were stored and clustered. Dynamic programming is used in a technique similar to dynamic time warping to generate predictions. Accuracy of approximately 75% of accent phrase boundaries were correctly detected in a phoneme balanced TTS data based with male speakers. This is impressive performance considering the technique uses only pitch contour information. The performance of this approach could likely be improved further with the inclusion of other observed correlates of prosodic phrasing such as energy, lengthening and syntactic information.

In this chapter, we do not explore the relationship between changes in speaking rate and phrase boundaries. However, this may be a useful indicator. In predicting phrase boundaries in Korean, Kim and Oh [101] found that modeling speaking rate could improve automatic boundary detection, reducing error by approximately 20% on a 3-level classification task. Wightman and Ostendorf [230] also used representations of speaking rate in their work on phrase boundary detection. The application of speaking rate information to the prosodic phrase boundary detection will be addressed in future work.

On Modern Greek material, Zervas and Maragoudakis, et al [239, 126] have applied

Naïve Bayes, Bayes Net, and CART classifiers to the task of phrase boundary assignment. Using surrounding part-of-speech and shallow chunking features tags Bayes Net classification achieved an intonational phrase F_1 of 0.796. Xydas [234] used manually enriched text markup to predict phrase boundaries in Modern Greek. Noun Phrases (NPs) in the text are marked with *intonational focus* features incorporating newness and "validation" factors: if the NP is the second argument of a verb, whether deixis is present and whether the NP contains a proper noun. These features along with with POS information classified the degree of disjuncture following each NP into four classes with 89.2% accuracy. This is an example of how additional lexical analysis can be used to improve prosodic phrase boundary prediction. While in this work the markup of additional information manually, the results show that extraction of this type of focus information can contribute to hypotheses of phrase boundary location.

On Chinese news data, Li et al. [120, 119] applied a two stage approach to phrase break assignment. An initial set of hypotheses were generated using Maximum Entropy (Logistic Regression) modeling using word identity, part of speech tags, and the number of syllables in a word. Then a prosodic phrase length distribution function was composed with these hypotheses to model the expected length of an intonational phrase. In isolation, the Maximum Entropy model generated an F_1 of 0.662; the length distribution function improves this to 0.697, compared to a human consistency F_1 of 0.750. This is a combination technique which could be applied to SAE speech as a way to merge individual phrase boundary hypotheses with phrase length information.

Also working on Chinese material, Hu et al. [95] hypothesizes that phrase breaks are more likely to occur between uncorrelated words and uses Mutual Information (MI) to capture this. They found a 2.4% improvement to a decision tree classifier by including mutual information measures. By applying an argmax function to a weighted combination of the likelihood of the POS model and MI likelihood (calculated at different windows surrounding the candidate) they achieved an intonational phrase boundary F_1 of 0.708.

The relationship between MI and phrasing has not yet been explored in SAE speech; the simplicity of the measure makes it an attractive feature to evaluate in the future.

Automatic detection of phrase boundaries has been shown to successfully contribute to the performance of spoken language processing tasks. As discussed in the examples above, the relationship between syntax and prosodic phrasing has been the subject of considerable research. There is evidence that syntactic ambiguity can be reduced by referring to intonational information [13, 218], and that intonational information can improve automatic parsing [146]. The speech synthesis community has been an important movitator of the investigation of the relationship between syntactic and prosodic phrasing. Prosodic structure plays a powerful role in the naturalness and intelligibility of synthesized speech [145, 184, 49]. Taking account of the syntactic and semantic content in intonation generation for synthesized speech can be used to improve the naturalness of the synthesized intonation [160]. From a recognition point of view, there is evidence that the inclusion of prosodic phrasing and other prosodic information can be used to improve automatic speech recognition performance [77, 191, 74]. We have also found that hypothesized intonational phrase boundaries can be used to improve the performance of automatic story/topic segmentation of broadcast news [171] and extractive speech summarization systems [131]. These two applications are discussed in greater detail in Chapter 7.

4.3 Acoustic Phrase Boundary Detection

There are three major acoustic indicators of the presence of a phrase boundary, as described in Section 4.1: 1) the presence of silence, 2) pitch and energy reset, and 3) pre-boundary lengthening. These all contribute to the perception of increased disjuncture at a word boundary. In this section, we explore representations of these phenomena and the use of these representations in supervised machine learning techniques to detect intonational phrase boundaries. These experiments are run on material from multiple speakers and are evaluated using either ten-fold or leave-one-speaker-out (speaker independent) cross-validation.

As an initial, baseline approach, we examine acoustic qualities of words to determine if there is any differentiation between those words that occur at the end of intonational phrases and those that do not. To evaluate any significant acoustic differences, we extract the minimum, maximum, mean, standard deviation of pitch and intensity within each word. We also calculate the z-score of the maximum of pitch and intensity within each word based on the mean and standard deviation of the feature within the word. These features are extracted over raw and z-score speaker normalized pitch and intensity contours, as well as corresponding slope contours. These comprise a feature vector of 20 acoustic features $\{5 \text{ aggregations}\} \times \{\text{raw v. normalized}\} \times \{\text{standard v. slope}\}$. We also include the duration, in seconds, of the word and the amount of silence immediately preceding or following it, bringing the number of features to 23. We evaluate these features using Logistic Regression, boosted Decision Trees (using AdaBoost), J48 Decision Trees, Sequential Minimal Optimization (SMO) of Support Vector Machines (SVM) with linear and radial basis function (RBF) kernels. This evaluation will be used to identify those classifiers that are best suited for this task. These evaluations are performed using ten-fold cross-validation. We evaluate each of these approaches using two measures: 1) Accuracy – the percentage of correct classifications and 2) F-measure of identifying phrase boundaries. Results of these baseline experiments can be found in Table 4.1.

We find that these relatively simple acoustic qualities are capable of detecting phrase boundaries with high accuracy and F-measure. The SVM classifiers perform significantly worse than the other explored techniques. The best performance on the BDC material is obtained using AdaBoost [57] with single split decision trees. However, the J48 decision trees and Logistic Regression perform only slightly worse; the F-measures do not significantly differ, though on BDC-read the accuracy of AdaBoost is significantly higher than that obtained via Logistic Regression. A t-test puts the significance of this 0.64% difference at p=0.000764.

Corpus	Majority Baseline	J48	Logistic	AdaBoost	SVM-l	SVM-rbf
BDC-read	87.08 / 0.0	94.97 ± 0.361 / 0.797 ± 0.0166	93.99 ± 0.361 / 0.737 ± 0.0123	95.61 ± 0.394 / 0.822 ± 0.0131	91.03 ± 0.459 / 0.578 ± 0.0153	87.18 ± 0.558 / 0.0163 ± 0.0256
BDC-spon	81.21 / 0.0	92.84 ± 0.344 / 0.807 ± 0.00959	92.60 ± 0.590 / 0.790 ± 0.0179	93.15 ± 0.525 / 0.811 ± 0.0127	88.13 ± 0.525 / 0.638 ± 0.0161	84.23 ± 0.836 / 0.318 ± 0.0232
BURNC	80.84 / 0.0	88.82 ± 0.279 / 0.653 ± 0.00484	85.07 ± 0.246 / 0.517 ± 0.0102	88.58 ± 0.377 / 0.601 ± 0.0136	84.40 ± 0.377 / 0.418 ± 0.00976	82.21 ± 0.394 / 0.134 ± 0.00891

Table 4.1: *Accuracy and F_1 of intonational phrase boundary detection using acoustic aggregations drawn from words. J48, Logistic Regression, AdaBoost, SVM with linear and RBF kernels are all evaluated.*

These features prove to be powerful indicators of the presence of an intonational phrase boundary. To more clearly understand the relative performance of each of these indicators, we examine separately a) the length of preceding and following silence, b) the duration of the word, and c-j) each of eight acoustic feature sets – {pitch v. intensity} × {raw v. normalized} × {standard v. slope}. We detect phrase boundaries using AdaBoost with each of these 10 feature sets, and report the accuracy, F-measure, precision and recall in 4.2. This analysis is performed using ten-fold cross-validation over the BURNC material.

Feature Set	Accuracy	F_1	Precision	Recall
Silence	88.22%	0.595	0.871	0.452
Duration	80.65%	0.225	0.491	0.265
Raw Pitch	80.87%	0.004	0.733	0.002
Raw Pitch Slope	80.84%	0.00	0.808	0.000
Norm Pitch	80.90%	0.008	0.800	0.004
Norm Pitch Slope	80.84%	0.00	0.808	0.000
Raw Intensity	80.84%	0.00	0.808	0.00
Raw Intensity Slope	81.06%	0.306	0.512	0.245
Norm Intensity	80.84%	0.00	0.808	0.00
Norm Intensity Slope	80.99%	0.349	0.507	0.283
All Pitch and Intensity	81.14%	0.233	0.528	0.149
All features	88.58%	0.601	0.907	0.449

Table 4.2: *Feature set analysis of intonational phrase boundary detection using AdaBoost and ten-fold cross-validation. Evaluated on BURNC.*

On the BURNC material, we find that silence is far and away the most predictive indicator of intonational phrase boundaries, yielding 88.22% accuracy. Moreover, we find that the acoustic aggregations are not well suited to differentiating phrase-final from phrase-internal words. Even using all of the pitch and intensity features, accuracy improves over that obtained using silence alone by 0.30%, a difference that is not statistically significant, and the F-measure is only 0.233. Similar relationships can be observed in the analysis of other corpora.

We find very little evidence that the pitch and intensity of a word indicates whether or not it occurs at the end of an intonational phrase. The literature indicates that the presence of a phrase boundary is attributable to a perception of disjuncture. These aggregations of pitch and intensity measure attributes of the word preceding the phrase boundary only. This approach classifies a word as intonational phrase ending or not. By taking a perspective where we identify which word *boundaries* are intonational phrase boundary, we can construct features which are more suited to representing the disjuncture associated with phrasing behavior.

From this perspective, we modify the previous feature set to reflect the changes in pitch and intensity aggregations on either side of a word boundary. We, therefore, construct a feature vector containing the difference of each of the pitch and intensity features extracted from the words preceding and following a candidate boundary. We will refer to these features as “Difference Features”. We evaluate these features separately on BURNC material using AdaBoost and ten-fold cross-validation. Results of this evaluation can be found in Table 4.3.

The performance of pitch features are not significantly impacted by the difference representation. They still are unable to predict phrase boundaries at a rate greater than the majority class baseline. However, the intensity features show significant improvements to accuracy when differences across word boundaries are used. In particular, the difference in raw or normalized intensity slope make up the most reliable feature sets. Examining the difference in mean slope, we find that phrase boundaries have a mean difference of

Feature Set	Accuracy	F_1	Precision	Recall
Raw Pitch	80.84% (-0.05)	0.00 (-0.004)	0.808 (+0.075)	0.000 (-0.002)
Raw Pitch Slope	80.84% (0.00)	0.00 (0.0)	0.808 (0.0)	0.000 (0.0)
Norm Pitch	80.84% (-0.06)	0.00 (-0.008)	0.808 (+0.008)	0.000 (-0.004)
Norm Pitch Slope	80.84% (0.00)	0.00 (0.0)	0.808 (0.0)	0.000 (0.0)
Raw Intensity	81.91% (+1.07)	0.140 (+0.140)	0.783 (-0.025)	0.077 (+0.077)
Raw I Slope	82.14% (+1.08)	0.337 (+0.031)	0.580 (-.068)	0.224 (-0.021)
Norm I	81.87% (+1.03)	0.140 (+0.140)	0.766 (-0.042)	0.077 (+0.077)
Norm I Slope	82.19% (+1.20)	0.318 (-0.031)	0.597 (+0.069)	0.217 (-0.066)
All	82.30% (+1.16)	0.340 (+0.107)	0.595 (+0.067)	0.238 (+0.089)

Table 4.3: *Difference Feature analysis of intonational phrase boundary detection using AdaBoost and ten-fold cross-validation. Evaluated on BURNC. Performance changes from corresponding Previous Word evaluations (cf. Table 4.1) are reported in parentheses.*

43.24dB/sec, while non-phrase word boundaries show a difference of -10.46dB/sec. This difference is significant with $p < 2.2 * 10^{-16}$. The overall mean is 0.76dB/sec. This analysis describes a phenomenon where the mean intensity slope of a word tends to be greater than the slope of the subsequent word – except at phrase boundaries, where the energy slope resets, with a dramatic increase. This reset in slope may contributes to a perception of disjuncture; a controlled perception study could evaluate to what degree human listeners respond to this acoustic quality.

It is somewhat surprising that the pitch features were not able to improve classification accuracy over the majority class baseline. A common acoustic indicator of phrase boundary location is pitch reset, where the pitch at the end of intonational phrases is lower than that at the start of the subsequent phrase. Note that this indicator does not occur at phrases ending with H% boundary tones (cf. Chapter 6). The difference features should capture this phenomenon. On the BURNC material, we find that the difference of mean normalized f0 is -0.0355 across phrase boundaries, and -0.0545 across word boundaries that are not also intonational phrase boundaries, a difference that approaches significance with p=0.0765. A negative difference value indicates that the value following the phrase boundary is *lower* than that preceding the phrase boundary. Positive values describe *increases* across phrase boundaries. Therefore, these differences in mean normalized f0 imply that there is a slight decrease of pitch at every word boundary, but this decrease is somewhat less at intonational phrase boundaries, on BURNC material. We previously hypothesized that the pitch reset commonly associated with intonational phrase boundaries would be realized by an *increase* of pitch across phrase boundaries. While we observe *less* of a reduction, examining the mean behavior, no increase in mean pitch realized across phrase boundaries. Broadcast news speech is particularly idiosyncratic. One of the qualities noticed by Bolinger [24] is that phrase-final words tend to be accented more than would in conversational or non-professional speech. This accenting behavior may increase the mean pitch at immediately preceding phrase boundaries, limiting the impact of pitch reset across these boundaries. Analyzing this feature on BDC-read material, we see some evidence that the phenomenon is genre dependent. On the non-professional read material, there is a mean difference of mean pitch of 0.219 at phrase boundaries, and -0.049 at non-phrase boundaries, a statistically significant difference ($p=1.51 * 10^{-8}$). The expected pitch reset is clearly observable on this material, as well as on the BDC-spontaneous corpora. On this set of non-professional spontaneous speech, the mean normalized pitch of words following intonational phrase boundaries is 0.0839 standard deviations greater than those immediately preceding boundaries. The mean difference between two words that are not divided by an intonational phrase boundary is -0.0318, a difference that a t-test determines to be significant with p=0.00250.

We evaluate the use of both the Baseline and Difference feature sets, on BDC-read, BDC-spontaneous and BURNC material. Recall that the Baseline features are extracted from a single word preceding a candidate boundary, while the Difference features are differences of features extracted from the two words surrounding each candidate boundary. This evaluation is performed using AdaBoost – the best performing classifier in the baseline acoustic experiments (cf. Table 4.1) – and ten-fold cross-validation. In the interest of space, confidence interval measures are omitted from the results reported in Table 4.4.

Across all corpora, we again observe the dominance of silence in the detection of intonational phrase boundaries. On the BDC-read corpus, a procedure using only the length of inter-word silence is able to detect phrase boundaries with 95.46% accuracy with an

Feature Set	BDC-read	BDC-spon	BURNC
Maj. Class Baseline	87.08% / 0.000	81.21% / 0.000	80.84% / 0.000
Silence	95.46% / 0.819	91.94% / 0.794	88.22% / 0.595
Duration	87.06% / 0.344	85.08% / 0.522	80.65% / 0.225
Baseline Pitch	87.76% / 0.289	81.96% / 0.245	80.86% / 0.003
Baseline Intensity	87.30% / 0.103	85.06% / 0.571	81.13% / 0.265
Difference Pitch	87.33% / 0.127	81.80% / 0.121	80.84% / 0.000
Difference Intensity	88.50% / 0.318	83.26% / 0.387	82.40% / 0.307
All Pitch and Intensity	90.42% / 0.549	85.90% / 0.596	81.14% / 0.233
All	95.65% / 0.883	93.13% / 0.810	88.89% / 0.647

Table 4.4: *Difference and Baseline Feature set analysis of intonational phrase boundary detection using AdaBoost and ten-fold cross-validation. Results include accuracy and* F_1.

0.819 F-measure. This is greater than the human agreement of 93.4% reported by Pitrelli, et al. [156]. The detection accuracy of a classifier using only this feature is less than the rate of human agreement reported in [156] on BDC-spontaneous and BURNC material. That said, the length of inter-word silence remains the dominant indicator of the presence of an intonational phrase boundary. We also find duration to be a moderately successful predictor of phrase boundary location. While the accuracy does not improve over the majority class baseline on BDC-read and BURNC material, the F-measure indicates the usefulness of this feature in the detection of phrase boundaries. On all corpora, the word that end intonational phrases are significantly longer than those that are not. The difference on average in the BDC-read material is 198ms, on BDC-spontaneous 235ms, and on BURNC 238ms.

This may be an effect of preboundary lengthening – phone, and specifically vowel, durations have increased length immediately preceding phrase boundaries. However, this may be explained in part to the accent rate of words preceding phrase boundaries. On all corpora we find that significantly more intonational phrase ending words are accented than those that are not. On BDC-read 65.55% of phrase-final words are accented compared to 35.43% of non-phrase ending words. On the BDC-spontaneous and BURNC material this difference is just as large, with 79.70% of phrase-final words accented in BURNC compared to 48.78% of non-phrase-final words. On BDC-spontaneous 77.89% of phrase ending words are accented, compared to 42.91% of non-phrase ending words. χ^2 tests confirm that these differences are significant with $p < 1 * 10^{-100}$. Accented words have a longer duration than non-accented words; phrase-final words have a longer duration than phrase-internal words. Using a factorial ANOVA to measure the impact of accent and phrase finality on word duration, on BURNC material, we find that both factors have a statistically significant effect on duration with $p < 2.2 * 10^{-16}$. Moreover, we find a significant interaction of the two factors, also with $p < 2.2 * 10^{-16}$. The mean duration of words under accent and phrase position conditions is presented in Table 4.5.

	Accented	Not Accented
Phrase-Final	533.6ms	432.0ms
Not Phrase-Final	381.9ms	172.5ms

Table 4.5: *The interaction of accent and phrase finality on word duration.*

The impact of duration on phrase boundary detection may also be an effect of the part-of-speech of common phrase ending words. Content words, for example, nouns and verbs, tend to have a longer duration than function words, like participles and prepositions. On the BURNC material, intonational phrases frequently end with a content word. 96.47% of intonational phrase boundaries are preceded by content words, while content words represent 57.77% of phrase-internal words. Using a factorial ANOVA, on BURNC data, we measure the impact that word class – Function versus Content – and phrase finality have on word duration. We find that content words are significantly longer than function words, and also that phrase-final words are significantly longer than phrase-internal words, both with $p < 2.2 * 10^{-16}$. Here, however, we find no evidence of an interaction between these two factors, $p = 0.8945$. Whereas accent and phrase position have a combined effect on word duration, it appears as though the effects of word class and phrase position are independent. The mean duration of words under word class and phrase position conditions is presented in Table 4.6. We examine the impact of part-of-speech based word-class information on intonational phrase boundary detection in greater detail in Section 4.4.

The acoustic features, particularly the intensity features, also demonstrate a capacity to

	Content	Function
Phrase-Final	520.4ms	308.6ms
Not Phrase-Final	363.5ms	153.0ms

Table 4.6: *The interaction of word class and phrase finality on word duration.*

be used to detect intonational phrase boundaries with greater than baseline performance. Note however, that the inclusion of Difference Features in the feature vector does not significantly change phrase boundary detection performance. This is evidence of the dominance of the silence feature in this classification task. While the Difference Features perform better than the acoustic aggregations extracted from the words in isolation, the impact of this difference is not able to improve performance when silence information is available. That said, pitch and energy reset, along with pre-boundary lengthening, are phenomena that can detect phrase boundaries even when there is no silence present. Intonational phrase boundaries occasionally occur without silence in the BDC material; 39.8% of boundaries are not indicated by silence on BDC-read, 40.8% on BDC-spon. The phenomena is more common in the BURNC, broadcast news, speech with 64.75% of boundaries not indicated by silence. Moreover these features may be used in the detection of *intermediate* phrase boundaries which tend not to have associated regions of silence. We evaluate these features for intermediate phrase boundary detection in Section 4.5. In Section 4.3.1 we examine different feature representations to capture the acoustic reset associated with phrase boundaries.

On all corpora, we find the best intonational phrase boundary detection accuracy to come from the combination of silence, duration, pitch and intensity features. On the two BDC subcorpora, the performance is at or above the rates of human agreement reported by Pitrelli, et al. [156]. Therefore, we will focus our analyses on BURNC material, where further improvement is necessary to reach this level of performance. In the following subsections, we examine techniques to improve performance through different representations of preboundary lengthening, and pitch and energy reset.

4.3.1 Representations of Pitch and Energy Reset

In the experiments described in Section 4.3, we found that the difference in pitch and intensity features on either side of a word boundary can be used to detect intonational phrase boundaries. In this section, we explore two other types of features to capture pitch and energy reset. In Section 4.3.1, we examine narrow regions around each word boundary as the locus of reset features. In Section 4.3.1, we explore ways to use linear regression best-fit lines to identify points of phrase indicating pitch and energy reset.

Narrow Window Reset Features

In Section 4.3, we extract Difference Features across word boundaries, where we calculate the difference of mean, maximum, minimum, zscore of maximum, and standard deviation of pitch and intensity contours from words on either side of each candidate boundary. In this section, we examine difference features calculated over a narrower region of analysis. The previous set of features calculate the difference in acoustic aggregations drawn from the full word. In this section we calculate the difference of mean pitch and intensity – both raw and speaker normalized – from shorter regions surrounding each word boundary. Specifically, we evaluate the regions within 200ms, 100ms, 50ms, and 20ms preceding and following a word boundary. As pitch and intensity contours are constructed with a 10ms frame rate, these correspond to a mean of (at most) 20, 10, 5 and 2 data points. Note, mean word length on these three corpora varies from 391ms to 514ms, thus, the longest, 200ms, region corresponds to roughly half a word. We evaluate these features as in Section 4.3 using AdaBoost with single split decision trees. Evaluations are performed using ten-fold cross-validation. Results of these experiments on BURNC material are reported in Table 4.7. In the interest of space, confidence intervals are omitted from this table.

We find no evidence of different degrees of pitch reset at intonational phrase boundaries using this representation, calculating the difference across a small window on either side of a candidate boundary. On the other hand, energy reset shows some discriminative impact. The

	20ms	50ms	100ms	200ms	All
Pitch	80.84% / 0.0	80.84% / 0.0	80.84% / 0.0	80.84% / 0.0	80.84% / 0.0
Energy	82.91% / 0.266	82.19% / 0.134	81.18% / 0.134	82.15% / 0.162	82.90% / 0.260
Both	83.06% / 0.277	82.19% / 0.134	82.19% / 0.134	82.14% / 0.181	83.04% / 0.267

Table 4.7: *Evaluation of pitch and energy reset features using AdaBoost on BURNC material. Widow sizes of 20ms, 50ms, 100ms and 200ms. Accuracy and F_1 are reported. Majority Class Baseline: 80.84%*

best performing single region to identify energy reset is the 20ms region. This is calculating as the difference of the mean of the last two intensity points of the word preceding a candidate boundary and the first two of the following word. In addition to being the best performing reset feature, the inclusion of other reset features does not improve detection accuracy or F_1 beyond that demonstrated the 20ms reset feature. This indicates that this extremely narrow region is a reliable region of analysis of intensity reset.

However, this feature, the difference of mean energy using a 20ms region of analysis, does not perform as well as the larger set of intensity Difference Features calculated over the full word explored in Section 4.3. Recall that these reset features examine only the difference in mean intensity on either side of a word boundary, while the intensity Difference Features compare the differences in maxima, means, minima and standard deviations of intensity and intensity slope from words preceding and following a candidate boundary. The Difference Features calculated from standard intensity features perform better than the 20ms intensity reset feature described here (cf. Table 4.3). It is the difference in *slope* features that yield higher F_1 values when evaluated over the full word. As a geometric correlate, the difference in slope features is equivalent to the inverse tangent of the angle between the two slope lines. In light of this observation, we calculate the narrow window mean reset features on the slope of intensity.

We find that the narrow window slope features significantly differ across intonational phrase boundaries when compared to their values across word boundaries that are not also intonational phrase boundaries. For example, the mean slope difference evaluated over a 20ms interval is 247.7 Hz/sec at intonational phrase boundaries, but 66Hz/sec at non-phased boundaries; a difference that is significant with $p < 2.2 * 10^{-16}$ determined by a t-test. However, when used as feature in isolation or combination in AdaBoost decision trees, these short window feature do not perform better than the majority class baseline. A Naive Bayes classifier with Gaussian modeling is able to correctly identify some phrase boundaries using these features with $F_1 = 0.352$. However, the accuracy of this classifier is 71.34%, significantly below the 80.84% majority class baseline. While narrow window representations of slope are significantly different at intonational phrase boundaries, these differences do not translate into improved automatic detection performance.

Regression based Reset Features

The second technique we use to represent pitch and energy reset relies on fitting line segments to the pitch and energy contour using linear regression. This approach is motivated by the assumption, at least for SAE declarative contours, that pitch and energy are relatively high at the start of an intonational phrase, and over the duration of the phrase decrease at a more or less constant rate. This assumption represents an idealized form of speech acoustics which is, obviously, not realized in practice; influences of accenting, intermediate phrase boundaries and segmental effects modify the pitch and intensity contours significantly away from this idealized form. However, a linear fit can capture an approximate trend over the course of a phrase. Changes in these trend lines may indicate points of disjuncture which correspond to prosodic phrase boundaries. To detect such changes we construct features from linear segments fit to pitch and energy contours on either side of a word boundary. From the parameterizations of these two linear segments we can construct features that can be used to identify intonational phrase boundaries. We should note that this approach is motivated by a hypothesis on the acoustic behavior of declarative contours. When a phrase contains a H% boundary tone or H- phrase accent, there is no expectation of a constant decline in pitch over the course of the phrase. On BURNC material, 56.16% of phrases end with a L-L% phrase accent and boundary tone combination, though only 49.00% of BDC-read phrases and 29.57% of BDC-spontaneous phrases do. The presence of high phrase ending tones, H- and or H%, will likely negatively impact the performance of classifiers using these regression features.

The window over which to generate a regression fit line can be defined in a variety of ways. Since the presence of silence can reliably identify many intonational phrase boundaries, these pitch and energy features are designed to detect those phrase boundaries that do not occur at silent regions. Therefore, we use silence to identify "breath groups", regions of speech between two silent regions. We fit linear segments to pitch and energy contours from a candidate word boundary to the nearest silent region. Figure 4.1 contains examples of these linear regression best-fit lines.

As pitch and energy contours are sensitive to changes due to accenting and segmental effects, we perform the linear regression over subsets of the pitch and energy points. We construct regression lines to *a*) the maxima, *b*) the minima, *c*) the mean values within each word, in addition to using *d*) every point. The intuition here is that one or more of these aggregations may be more reliable than others in reflecting the declension of pitch or energy over the course of a phrase.

Using coefficients (slope and intercept) and error of these best-fit lines we construct six features to represent pitch and energy reset.

- **Slope and Error** We include the slope and root mean squared error (RMSE) of the previous and following best-fit lines in the feature vector

- **Regression Reset** We evaluate both regression lines at the word boundary. Regression Reset is the difference in the linear regression estimates of the preceding and following best-fit lines at the word boundary. A visual representation of Regression Reset can be found in the top image of Figure 4.1. The Regression Reset is the length of the vertical black line.

- **Error Ratio** Our hypothesis is that we can fit a linear segment to the pitch and energy contours within an intonational phrase with relatively low error. If this hypothesis holds, the linear regression error of a contour containing a phrase boundary should be significantly greater than one that does not. We construct a third regression line over the full breath group contour. The sum of the errors of the preceding and following regression lines is a lower bound on the error of this third line. We construct an Error Ratio feature, calculated as the ratio of the squared error of the breath group best-fit line to the sum of the squared errors of the preceding and following best-fit lines. A large Error Ratio indicates that the preceding and following linear fits were significantly more successful than the full contour regression. This indicates that there are two distinct linear segments on either side of the candidate boundary that are not colinear. A visual example of this feature can be found in Figure 4.1. The top image contains a contour and regression lines of regions preceding and following a candidate word boundary. Note that there are no pitch points between seconds 1.67 and 2.08; the line segment in the lower example is an artifact of the linear interpolation between pitch points to aid visualization of the contour. The lower image contains the same contour with a single line segment fit over the full breath group.

We evaluate the use of these regression based features on BURNC material using AdaBoost and ten-fold cross validation. We evaluate the use of regression lines fit to each of the four aggregations, *a*) the maxima, *b*) the minima, *c*) the mean values within each word, *d*) every point.Table 4.8 contains results from these experiments.

	Maxima	Minima	Mean	All Points
Pitch	80.84% / 0.0	80.84% / 0.0	80.84% / 0.210	86.33% / 0.462
Energy	80.97% / 0.327	81.26% / 0.410	81.04% / 0.417	87.41% / 0.517
Both	80.84% / 0.328	81.29% / 0.375	81.06% / 0.417	87.41% / 0.517

Table 4.8: *Evaluation of pitch and energy regression features using AdaBoost on BURNC material using ten-fold cross-valudation. Accuracy and F_1 are reported. Majority Class Baseline: 80.84%*

These features use silence information in their calculation – regression lines are con-

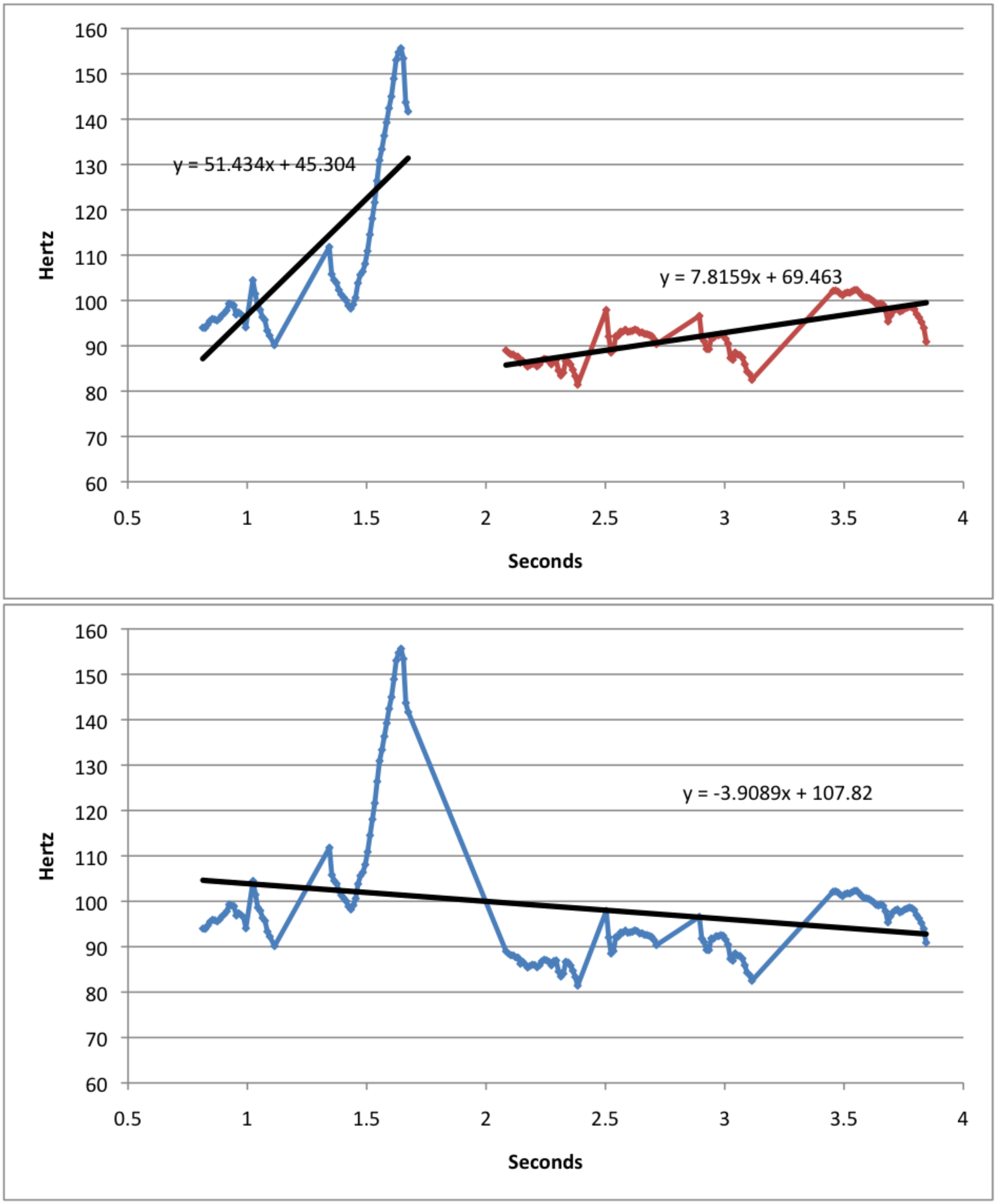


Figure 4.1: *Error Ratio example. The RMSE of the single fit line is 14.41, the RMSE with two fit lines is 7.74.*

structed from the region between the candidate word boundary and the nearest preceding and following silence. However, we find that the best performing results are identical to the decision obtained by simply identifying intonational phrase boundaries where silence is encountered. While this represents relatively high performance, the goal of exploring reset features is to capture indicators of intonational phrase boundaries where there is no silence present. Therefore, we omit word boundaries that coincide with silence from the data set, and repeat the analysis of the regression features. The regression features are unable to detect intonational phrase boundaries within breath groups with performance greater than the majority class baseline. However, t-tests reveal significant differences in the values of some of the regression based features. For example, the slope of energy preceding non-intonational phrase boundaries is 3.16, while preceding intonational phrase boundaries the mean value is -0.9932. That is, on average, the slope of intensity between a silent region and a word boundary is positive, roughly 3dB/sec. However, if the word boundary is an intonational phrase boundary, the slope is, on average, negative – approximately -1dB/sec. A t-test considers this difference to be significant with $p < 2.2 * 10-16$. The error ratio of pitch regression lines reveals a greater ratio at intonational phrase boundaries (1.9486) than other word boundaries (1.769), a significant difference of 0.179 ($p = 2.027 * 10^{-8}$). Also, the regression reset shows an increased reset at phrase boundaries, 33.92Hz, compared to non-phrase word boundaries 11.11Hz. This difference approaches significance with $p = 0.0867$. The use of these types of regression features is clearly limited. However, the results observed here suggest that pitch and energy trends may be useful for phrase boundary detection. Identifying the best way to extract this information remains an area of future study. These differences suggest that, while not themselves particularly discriminative, another representation of pitch and energy trend lines may be more successful in the detection of intonational phrase boundaries.

Evaluation of all Reset Features

In this section, we examine three approaches to capturing the pitch and energy features that indicate the presence of an intonational phrase boundary: 1) acoustic aggregations of words preceding phrase boundaries, 2) Difference Features calculated across word boundaries and 3) linear regression based features. In this subsection, we evaluate the use of each of these feature sets for phrase boundary detection on three corpora: BDC-read, BDC-spontaneous, and BURNC.

These feature sets are evaluated using AdaBoost with single split decision trees. Evaluations using ten-fold cross-validation can be found in Table 4.9. Results of speaker independent evaluations using leave-one-speaker-out cross-validation can be found in Table 4.10.

	BDC-read	BDC-spon	BURNC
Silence	95.46% / 0.819	91.94% / 0.794	88.22% / 0.594
Duration	87.06% / 0.344	85.08% / 0.522	80.65% / 0.344
All Pitch and Intensity	94.82% / 0.766	91.86% / 0.753	88.90% / 0.669
- Previous Word	90.46% / 0.538	85.54% / 0.591	81.21% / 0.387
- Difference Features	89.47% / 0.460	83.64% / 0.408	82.80% / 0.200
- Narrow Window	89.28% / 0.478	84.10% / 0.423	82.90% / 0.215
All Feature Sets	95.71% / 0.826	93.00% / 0.809	89.45% / 0.682

Table 4.9: *Evaluation of all acoustic feature sets on BDC-read, BDC-spontaneous and BURNC material. Evaluated using AdaBoost and ten-fold cross-validation.*

	BDC-read	BDC-spon	BURNC
Silence	95.44% / 0.819	92.01% / 0.795	87.74% / 0.559
Duration	85.46% / 0.367	84.77% / 0.486	79.73% / 0.273
All Pitch and Energy	94.64% / 0.764	90.61% / 0.705	88.91% / 0.661
- Previous Word	89.29% / 0.517	85.05% / 0.559	82.56% / 0.448
- Difference Features	88.63% / 0.417	83.70% / 0.432	82.81% / 0.383
- Narrow Window	88.80% / 0.419	83.87% / 0.385	82.32% / 0.186
All Feature Sets	94.53% / 0.763	92.39% / 0.786	88.98% / 0.653

Table 4.10: *Speaker Independent evaluation of all acoustic feature sets on BDC-read, BDC-spontaneous and BURNC material. Evaluated using AdaBoost and leave-one-speaker-out cross-validation.*

On all corpora, we find that the full set of pitch and energy reset features can be used to automatically detect intonational phrase boundaries with high accuracy and F_1. The pitch and energy reset features are able to detect phrase boundaries that occur without silence. Recall, in the BDC-read and BDC-spontaneous material, there are relatively few instances of intonational phrase boundaries occurring without silence (39.8% and 40.8%), and there are very few silent regions which do not indicate phrase boundaries (5.3% and 6.8%). However, on the BURNC material 64.8% of phrase boundaries occur without silence, though only 2.4% of silences are not indicative of intonational phrase boundaries. Therefore, on the BURNC material, features which do not rely on silence are particularly necessary. It is on this corpus that we see the full set of reset features outperforming the silence feature, and, moreover, combining with silence and duration information to improve overall automatic detection accuracy. On the BDC corpora, the improvements obtained by including duration and pitch and energy reset features with silence information are less dramatic.

When evaluated using ten-fold cross-validation, the inclusion of duration and pitch and energy reset features with the silence feature improves F_1 on BURNC by 0.083 (p= $0.1.584 * 10^{-6}$), on BDC-read by 0.005 ($p = 0.714$) and on BDC-spontaneous by 0.020 ($p = 0.111$). Note that a t-test reveals that this improvement is significant only on BURNC material. Under speaker independent evaluations none of these improvements demonstrated by the inclusion of pitch and energy reset features are statistically significant.

Despite their inability to significantly contribute to the overall intonational phrase boundary detection performance, pitch and energy reset features have the advantage of being able to detect phrase boundaries that are not indicated by silence. Due to the relative infrequency of these events and unreliability of these features, they fail to significantly improve performance. Despite this, they perform moderately well when evaluated without silence and duration features. As previously mentioned, there are effects that influence the pitch and intensity of an utterance in addition to phrasing. In the future we will work to account for these. For example if the effect of accenting, phrase accents, and boundary tones were subtracted or normalized out, pitch and energy reset features may be more reliable in their ability to detect intonational phrase boundaries.

4.3.2 Representations of Preboundary Lengthening

Contributing to the perceived disjuncture at intonational phrase boundaries is a phenomenon known as preboundary lengthening. Phones, and vowels in particular, immediately preceding phrase boundaries have increased duration. In this section, we examine the forced-alignment phone information contained in the BURNC corpus to observe the effect of pre-boundary lengthening. Time-aligned phone information is unavailable for BDC material.

We examine three segmental durations to measure the presence of any lengthening effect at phrase boundaries. We calculate the length of each word-final vowel, syllable, and syllable rhyme (the vowel and any following consonants). In addition, it is possible that pre-boundary lengthening influences phone durations earlier than the final phone. Therefore, we also investigate the mean vowel length and mean syllable length within each word. Phones have different inherent durations. To account for these inherent differences, we normalize vowel durations using z-score normalization based on vowel identity. In addition to this phone-identity normalization, we also normalize phone durations by speaker. This normalization has been used to detect preboundary lengthening by Wightman et al. [231]. Under this normalization, the mean and standard deviation of phone durations are calculated separately for each speaker.

Each of these representations of phone duration show a significant increase at intonational phrase boundaries compared to non-intonational phrase word boundaries. T-tests indicate that the differences for each feature are significant with $p < 2.2 * 10^{-16}$ For example, the mean final vowel length immediately preceding intonational phrase boundaries is 134ms, while at non-phrase ending word boundaries it is 78ms, only 58% as long. To evaluate the difference in performance of each of these segment duration features, we calculate Naïve Bayes [100] classification accuracy and F_1 of each evaluated in isolation using Gaussian models. Results of this classification as well as the mean values of each feature at phrase boundaries and non-phrase word boundaries are reported in Table 4.11. μ_{phrase} is the mean value of the feature extracted from words that immediately precede intonational phrase boudnaries. μ_{word} is the mean value extracted from words that do not precede intonational phrase boundaries. Naïve Bayes evaluations were performed using ten-fold cross-validation. Speaker identities are assumed to be known in the calculation of speaker normalized features.

	μ_{phrase}	μ_{word}	Accuracy	F_1
Final Syllable	0.296	0.182	82.45 ± 0.197	0.368 ± 0.00951
Mean Syllable	0.278	0.186	81.19 ± 0.180	0.277 ± 0.00869
Final Rhyme	0.229	0.125	84.64 ± 0.230	***0.476* ± *0.0108***
Final ID-Norm Rhyme	1.739	-0.279	***85.24* ± *0.246***	***0.472* ± *0.0108***
Final Spkr-Norm Rhyme	1.110	-0.053	83.02 ± 0.197	0.409 ± 0.00869
Final Vowel	0.134	0.078	83.23 ± 0.164	0.411 ± 0.00738
Final ID-Norm Vowel	1.112	-0.080	83.68 ± 0.279	0.412 ± 0.0118
Final Spkr-Norm Vowel	1.015	-0.101	82.82 ± 0.230	0.391 ± 0.00902
Mean Vowel	0.085	0.064	80.22 ± 0.230	0.187 ± 0.0118
Mean ID-Norm Vowel	0.437	-0.086	80.84 ± 0.197	0.218 ± 0.0123
Mean Spkr-Norm Vowel	0.287	-0.027	80.86 ± 0.164	0.198 ± 0.00935

Table 4.11: *Descriptive statistics and Naïve Bayes intonational phrase boundary detection performance using pre-boundary lengthening features.*

The raw and vowel-identity normalized final rhyme length are the duration features most indicative of intonational phrase-final pre-boundary lengthening. A histogram of the final rhyme lengths preceding intonational phrase boundaries and those preceding non-phrase-final word boundaries can be found in Figure 4.2. Increased final rhyme length is more predictive of intonational phrase boundaries than either the final or mean vowel length. This indicates that phrase ending consonants are also significantly longer than their word ending counterparts.

Whether measuring vowel (syllable nucleus) lengths or full syllable durations, analyzing the mean duration over the word is less predictive of phrase boundaries than examining only the final vowel or syllable. Moreover, the final rhyme duration is more predictive than the final syllable or final vowel duration. These observations indicate that the pre-boundary lengthening phenomenon is localized in the phrase-final vowel and subsequent consonants. The inclusion of segment duration information prior to the phrase-final vowel, whether by examining the full syllable, or by taking the mean vowel length over the word, yields worse performance.

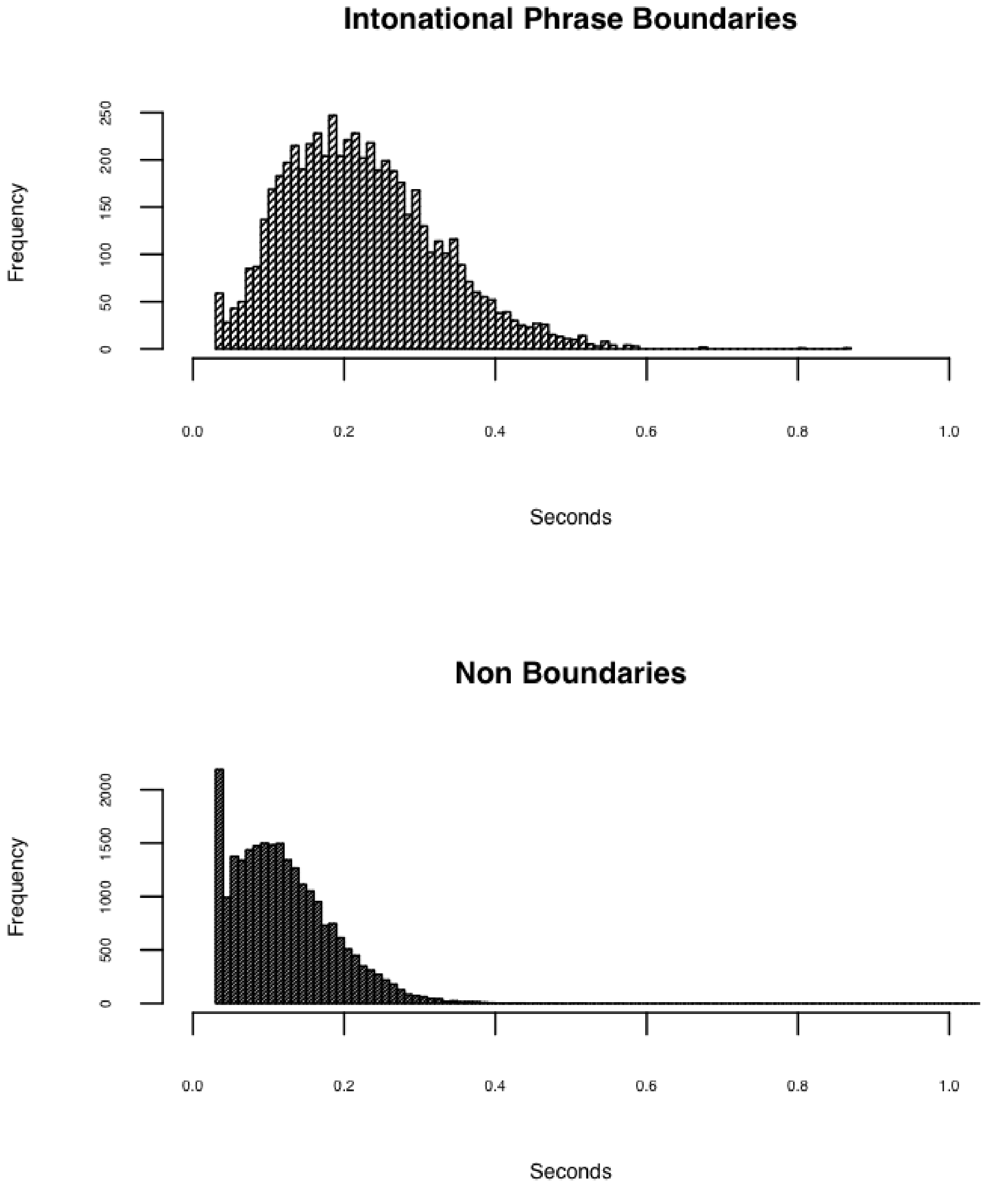


Figure 4.2: *Histograms of final rhyme durations preceding intonational phrase boundaries, and non-phrase ending word boundaries. Intonational Phrase Boundaries:* $\mu = 0.229$ $\sigma = 0.100$ *Non Boundaries:* $\mu = 0.125$ $\sigma = 0.0661$

The normalization of phone durations by phone identity slightly improves intonational phrase boundary detection performance. While F_1 using this normalization only yields a significant increase when applied to the mean vowel length, detection accuracy is significantly increased when the normalization is applied to both the final rhyme and final vowel. The speaker and identity normalization results in significantly *worse* performance than vowel-identity normalization when evaluating both final vowel and final rhyme durations.

In addition to these individual evaluations, we construct a feature vector including each of the 11 features, and evaluate AdaBoost with single split decision trees using ten-fold cross-validation. Under speaker independent evaluation, using leave-one-speaker-out cross-validation, none of the three speaker normalized features are included. In these evaluations the normalization parameters – mean and standard deviation of phone durations by phone identity – are calculated based on training data only. We also include these preboundary lengthening features with the acoustic features – silence, word duration and representations of pitch and energy reset – explored in Section 4.3. Results from these experiments can be found in Table 4.12. For reference, the performance using only acoustic features is also included. The combination of all preboundary lengthening features allows us to detect intonational phrase boundaries with an F_1 value of 0.576. Under speaker independent evaluation, the F_1 is reduced by 0.031, a difference that a t-test does not determine to be significant (p = 0.363). While we find final rhyme length to be the best single feature for capturing preboundary lengthening effects, the inclusion of other representations improves

Evaluation	Feature Set	Accuracy	F_1
10-fold	Lengthening Only	84.35 ± 0.574	0.576 ± 0.00951
	Acoustics Only	89.46 ± 0.377	0.682 ± 0.0144
	Lengthening + Acoustics	90.73 ± 0.180	0.736 ± 0.00754
Leave-one-speaker-out	Lengthening Only	84.91 ± 1.033	0.545 ± 0.0525
	Acoustics Only	88.98 ± 1.099	0.653 ± 0.0508
	Lengthening + Acoustics	90.00 ± 1.148	0.700 ± 0.041

Table 4.12: *AdaBoost intonational phrase boundary detection accuracy using pre-boundary lengthening features alone and with acoustic features.*

the F_1 of intonational phrase boundary detection. Moreover, we find that the use of both preboundary lengthening features and acoustic features yields improved AdaBoost detection performance. These features are relatively independent of speaker differences. While the performance is lower under leave-one-speaker-out cross-validation, the differences in accuracy and F_1 are statistically significant with t-tests with $p = 0.478$ and $p = 0.221$, respectively.

Preboundary lengthening is a powerful indicator of intonational phrase boundaries. Even without normalizing for phone or speaker identity, the length of word final vowels and syllable rhymes can be used to automatically detect intonational phrase boundaries. The combination of preboundary lengthening features predicts intonational phrase boundaries with 84.91% accuracy and F_1 of 0.545. These features are distinct from the silence and acoustic reset features investigated in other subsections of Section 4.3. Not only are the phenomena fundamentally different, but the performance of the combined feature set indicates that the predictive information they capture is non-redundant.

4.4 Lexico-Syntactic Phrase Boundary Detection

The use of lexical and syntactic information in phrase boundary detection has been heavily explored, as discussed in Section 4.2. While syntax does not *prescribe* appropriate prosodic phrasing, there is a relationship between the two. Study of the relationship between lexico-syntactic content of an utterance and valid prosody has been used in text-to-speech prosodic assignment decision making. In speech synthesis applications, the synthesizer determines a desired prosody for the synthesized utterance. Whether the assigned prosody takes the form of an intermediate symbolic representation like ToBI, or assignment of f0, energy and timing targets, the task of prosodic assignment only has access to the lexical content of the input text. The work described in this chapter concerns the detection of phrase boundaries where spoken, as well as lexical, material is available. While the task is distinct, the approaches and lessons learned from prosodic assignment can inform the use of lexico-syntactic information for phrase boundary detection.

In general, prosodic assignment for text-to-speech synthesis operates on single sentences or utterances – or at least has access to punctuation information. In the work presented in this chapter, we do not assume sentence boundary information or other punctuation to be available. This requires additional human labeling or an automatic system for sentence boundary detection or insertion of punctuation. While a reasonable assumption for prosodic assignment, inclusion of sentence boundary information in prosodic analysis of speech requires additional resources, and will likely introduce some amount of noise. While certainly useful in predicting phrase boundary locations, in this section we do not assume the presence of this information.

In this section, we examine the use of syntactic parse tree information and part-of-speech based word-class features in the detection of intonational phrase boundaries. The feature representations we examine are related to those used in [222] and [91]. We generate parse trees for each file in the BDC and BURNC corpora using an implementation of Charniak's maximum-entropy-inspired parser [36] trained on Switchboard data [64] and part-of-speech tags using the Stanford Tagger [213]. Prior to parsing, sentence boundaries are hypothesized for each file using MXTERMINATOR, a maximum entropy sentence segmenter trained on Wall Street Journal text [168]. Many of the syntactic features investigated in this section are also applied to the task of classifying phrase ending intonation in Section 6.5.2. We extract these features for each word boundary $[w_i, w_j]$, where w_i refers to the previous word and w_j, the following word.

- **Constituent Features** We include, as a nominal value, the identity of 1) the largest constituent containing w_i and not w_j, 2) the largest constituent containing w_j and not w_i and 3) the smallest constituent containing both w_i and w_j. A special None token is created for instances when these constituents do not exist. These features are used in [222, 223].

- **Positional Features** We extract the absolute and relative position of each word within its narrowest constituent, from both the start and end. Wang and Hirschberg found a strong influence of phrase length on the presence of a phrase boundary [222]. That is, phrase boundaries are more likely to occur if it has been a long time since the last phrase boundary. Rather than include manually annotated phrase boundary position information, we approximate this effect by including the number of words since the last silent region over 200ms was observed. Note that this is not strictly a lexical feature, but it serves as a crude proxy for sentence boundary information.

- **Parse Tree Distance Features** We calculate the degree of syntactic disjuncture by the parse tree distance between w_i and w_j. Treating a parse tree as a graph, the parse tree distance is the length of the path between the node representing w_i and w_j. This representation was used by Read and Cox [166]. However, this measure may penalize complex syntactic constructions too greatly. Therefore, we also normalize this distance by the parse tree depth of the deepest common ancestor of w_i and w_j. If two word boundaries have the same parse tree distance between them, the normalized measure will treat the words at a lower parse tree depth as less disjount than that pair higher in the tree. An illustrative example of this is presented in Figure 4.3. In both examples the distance between "stop", w_i, and "from", w_j is 4. In the first example, the depth of their closest common ancestor, the NP node representing "one stop from the train", is 2. Therefore the distance ratio is 4/2 = 2. In the second example, the depth of their closest common ancestor is 4. Therefore, the distance ratio is 4/4=1.

We evaluate these parse tree based features using AdaBoost with single split decision trees. Evaluation is performed using speaker independent, leave-one-speaker-out cross-validation, as well as ten-fold cross validation, which includes material from the same speaker in both training and testing folds. As noted in Section 4.2, the BURNC corpus includes many read stories that are repeated by multiple speakers. To avoid learning syntactic-prosodic idiosynrasies specific to this repeated material, we impose a restriction

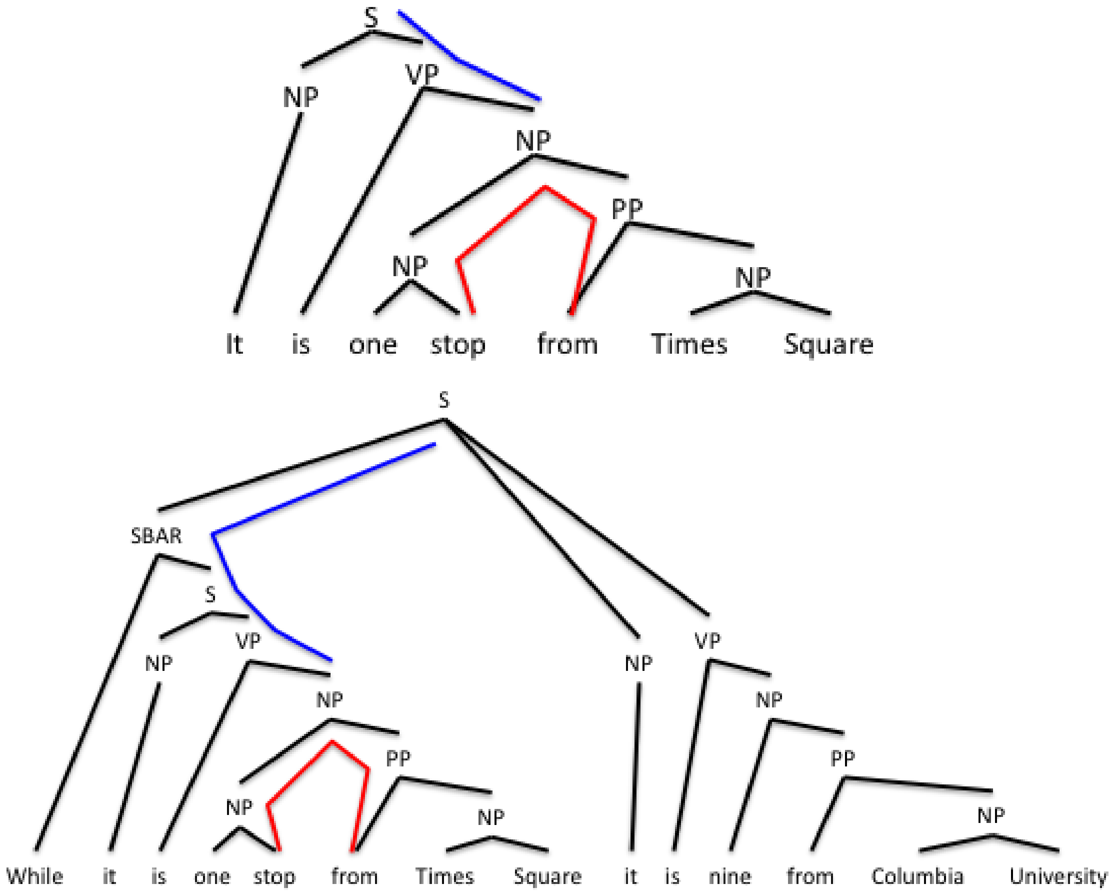


Figure 4.3: *A graphical example of the parse tree distance ratio feature. In both examples the distance between "stop" and "from" is 4; the distance ratio in the top example is 2, while in the bottom it is 1.*

on the ten-fold cross-validation that material from a story may not appear in both training and testing folds. This restriction makes leave-one-speaker-out cross-validation impractical – nearly all stories spoken by any test speaker will have also been read by one of the training speakers. Results of these evaluations can be found in Table 4.13, confidence interval values have been omitted in the interest of space.

On all corpora, we find parse tree based features to be able to detect intonational phrase boundaries with performance significantly greater than chance. The absolute improvement in accuracy falls between 3.8% and 3.24%, while the F_1 varies more widely. The BDC-spontaneous material demonstrates the worst performance of the three evaluated corpora. This suggests that in spontaneous speech the relationship between syntax and prosody may

Evaluation	BDC-read	BDC-spon	BURNC
Baseline	87.08% / 0.000	82.21% / 0.000	80.48% / 0.000
Ten-fold	91.88% / 0.635	85.45% / 0.444	84.08% / 0.521
Leave-one-speaker-out	91.39% / 0.606	85.47% / 0.458	NA

Table 4.13: *Evaluation of parse tree features using AdaBoost under ten-fold and leave-one-speaker-out cross-validation. Accuracy and F_1 are reported.*

be the loosest. By comparing the BDC-read and BDC-spontaneous results, we find evidence that the relationship between phrasing and syntax are closer in read speech than spontaneous utterances. The spoken material in these two corpora have the same domain, are spoken by the same speaker, and have almost the same lexical material. Genre and the presence of disfluencies are the two differences between these corpora. Spoken disfluencies introduce syntactic ambiguities which are difficult for automatic parsers to process. Parse errors on the spontaneous material may lead to the reduced performance on the BDC-spontaneous corpus. The variability in domain, speakers and genre makes comparison with the BURNC performance less informative, but such a comparison provides some evidence that phrasing in broadcast news speech may be more closely tied to syntax than spontaneous speech, but less than non-professionally read speech. We find no evidence of speaker dependency in these results. Neither F_1 nor accuracy significantly decreases under leave-one-speaker-out cross validation when compared to ten-fold cross-validation.

Each of the features examined shows significantly different distributions at intonational phrase boundaries and non-phrase word boundaries, on all corpora. The observations from the data are consistent with the hypotheses that directed the feature construction. Phrase boundaries fall closer to the end – and, naturally, further from the start – of syntactic constituents than across other word boundaries. The parse tree distance is significantly greater across intonational phrase boundaries than other boundaries – this is evident whether examining the raw distance or the distance ratio. Also, phrase boundaries fall further from silent regions than other word boundaries. On the BURNC and BDC-read material each of these numeric features show significant differences with p values less than 0.003. On the BDC-spontaneous material, the distance from silence feature shows a less dramatic difference, with phrase boundaries 7.05 words from silence, while non-phrase word boundaries are 6.64 words – a difference of only 0.41 words. This difference is significant with $p = 0.01261$. The same feature differs by 2.38 and 1.80 words on BDC-read and BURNC material with corresponding p-values less than $1 * 10^{-13}$.

The nominal constituent features are also helpful in automatically detecting intonational phrase boundaries. If words w_i and w_j lie in different sentences, they do not have a smallest covering syntactic constituent; in this case the value of this feature is None. Intonational phrase boundaries are quite common (98.7% of the time) when the smallest syntactic constituent covering w_i and w_j does not exist. On the other hand phrase boundaries are quite rare within prepositional phrases (PP); in only 1.74% of instances where the smallest covering constituent is a PP does $[w_i, w_j]$ represent an intonational phrase boundary. Intonational phrase boundaries are more common when the smallest covering constituent is a verb phrase (VPs) than a noun phrase (NPs) – with rates of 23.5% and 16.0% respectively.

Intonational phrase boundaries are relatively common, with a rate of 77.8%, when the largest constituent covering w_i but not w_j is an "unlike coordinated phrase" (UCP). UCPs describe instances of a coordinated phrase across syntactic categories. For example, "She flew [yesterday and on July 4th.]". A subset of formal context-free rules of the required to parse a UCP non-terminal in a Tree Adjoining Grammar can be found in Figure 4.4 [163]. Moreover, intonational phrase boundaries are more likely to occur at the end of VPs (53.4%) than NPs (45.3%). On the other hand, intonational phrase boundaries are more likely to fall at the start of an NP than a VP; 41.1% of noun phrases are immediately preceded by an intonational phrase boundary, while only 22.5% of verb phrases are. Moreover, 66.6% of constituents with a coordinating conjunctions as their head word are preceded by an intonational phrase and 50.6% of adjective phrases are. The expected rate of intonational phrase boundary placement is 19.27%. Each of these rates are significantly different than this expected rate of intonational phrase boundary placement as determined by a χ^2 test.

UCP → ADVP CC PP
UCP → PP CC ADVP
UCP → ADJP CC NP
UCP → NP CC ADJP

Figure 4.4: *A subset of UCP production rules.*

In addition to these parse tree features, we use part-of-speech based word-class features to predict phrase boundaries. The intuition here is that certain sequences of word-class tokens are more or less likely to precede or surround an intonational phrase boundary than others. To capture these likelihoods, we construct five part-of-speech (POS) multinomial models. We construct unigram, bigram, trigram, surrounding bigram and surrounding four-gram models. The description of these models can be found in Table 4.14, assuming a candidate phrase boundary falls between word i, and word $i + 1$, w_i is the word-class of word i, and b is a binary variable corresponding the the presence or absence of a intonational phrase boundary. These models are trained using raw part of speech tags, as well as two collapsed tag sets. The broad class tag set comprises six classes: 1) Noun 2) Verb, 3) Adjective, 4) Adverb, 5) Cardinal, and 6) Function. The function/content tag set categorizes all part of speech tags as Function or Content words. A more detailed description of these collapsed word-classes can be found in Section 3.7.1. These prosodic/syntactic "language models" operate only on the surface form of the syntactic structure and have no access to the deeper parse information. They are therefore able to capture a distinct, more linear, aspect of syntactic information than the parse tree features investigated previously. We evaluate each of the five model types with each of the three word-class tag sets, giving us a set of

Model Name	Formula
Unigram	$p(b \mid w_i)$
Bigram	$p(b \mid w_i, w_{i-1})$
Trigram	$p(b \mid w_i, w_{i-1}, w_{i-2})$
Surrounding Bigram	$p(b \mid w_i, w_{i+1})$
Surrounding Four-gram	$p(b \mid w_i, w_{i-1}, w_{i+1}, w_{i+2})$

Table 4.14: *Description of part-of-speech based word-class models.*

fifteen models.

The first approach we explore is to use an argmax classification over the likelihoods of each model separately. The unigram model with raw POS tags generates an accuracy of 81.56% and F_1 of 0.287 on BURNC, over the 80.48% majority class baseline. Interestingly, the argmax decision relies on only a single POS tag in its decision boundary, specifically, the POS tag of the word immediately preceding the candidate boundary. On the BURNC material, 69.2% of plural nouns (NNS) are followed by intonational phrase boundaries, while less than 50% of every other POS tag are. This indicates that surface syntactic information can be used to predict intonational phrase boundary locations. The performance of each of models is evaluated on BURNC. As mentioned previously, speaker independent evaluation of lexical features on this material is not feasible. The results of ten-fold cross-validation can be found in Table 4.15.

Model	RAW	BROAD CLASS	FUNCTION/CONTENT
Unigram	81.56% / 0.287	80.84% / 0.00	80.84% / 0.00
Bigram	81.63% / 0.288	80.73% / 0.149	80.84% / 0.00
Trigram	81.56% / 0.286	80.83% / 0.134	80.84% / 0.00
Surrounding Bigram	***86.12% / 0.547***	***84.20% / 0.553***	82.20% / 0.134
Surounding 4-gram	83.98% / 0.360	84.09% / 0.489	82.20% / 0.134

Table 4.15: *Evaluation of each word-class models on BURNC material under ten-fold cross-validation. Models trained on* RAW, BROAD CLASS *and* FUNCTION/CONTENT *word-classes are evaluated. Accuracy and F_1 are reported.*

First of all, we observe no improvement by collapsing part-of-speech tags into broader word-class categories. Such collapsing of categories should allow for the inclusion of longer context without suffering from data sparsity problems. However, the reduced granularity of the word-class information due to the collapsing of part-of-speech tags, in this case, significantly reduces performance. The RAW tags generate the highest accuracy and F_1 using most models. When applying the Surrounding 4-gram model to the RAW tag set, we do observe some data sparsity issues, which are avoided by the BROAD CLASS tag set. However, the Surrounding 4-gram model does not perform significantly better than the Surrounding Bigram model – suggesting that there is little discriminative information in the penultimate and second word after an intonational phrase boundary. We also find that the Trigram model does not perform significantly differently than the Bigram model, regardless of which tag set is used. These two results suggest that the part of speech tag information that is most relevant to phrase boundary detection is centered on the two words preceding and following the boundary. Since the addition of context does not significantly help automatic detection of phrase boundaries, the RAW tag set does not suffer from data sparsity issues. The Surrounding Bigram using RAW or BROAD CLASS tags generates the best results. The accuracy of the RAW model is significantly higher than the BROAD CLASS model (p=0.00138), while its F_1 is not significantly lower (p=0.649). Thus, we conclude the Surrounding Bigram model using RAW tags to be the preferred model.

To combine this part-of-speech information with the parse tree based syntactic features, we extend the feature vector with surrounding part-of-speech bigrams as well as likelihood of the Surrounding Bigram model. There is, of course, some degree of redundancy in the information contained in the POS based and parse tree based features. We include these features using both the RAW and BROAD CLASS tag set. We evaluate this extended feature vector on the BDC-read, BDC-spontaneous and BURNC material using J48 decision trees and ten-fold cross-validation. Evaluation of the performance of J48, SVM with linear kernel, Logistic Regression and AdaBoost in combining these features were performed, with J48 achieving the highest results. J48 results are reported for the Parse Feature sets as well as the combination of Parse and POS features.

Feature Set	BDC-read	BDC-spon	BURNC
Baseline	87.08% / 0.000	82.21% / 0.000	80.48% / 0.000
POS	88.11% / 0.227	81.45% / 0.303	86.12% / 0.547
Parse	92.87% / 0.662	85.81% / 0.505	86.23% / 0.591
POS + Parse	92.17% / 0.645	85.80% / 0.506	86.57% / 0.611

Table 4.16: *Evaluation of part of speech and parse tree based featues using AdaBoost under ten-fold and leave-one-speaker-out cross-validation. Accuracy and F_1 are reported.*

Despite the moderate success of part-of-speech word-class models to detect intonational phrase boundaries, they do not combine with parse features to improve overall performance.

None of the combined performance is significantly different from the performance using parse tree based features alone. These two syntactic representations likely capture redundant information. While part-of-speech information is not exactly contained in the parse tree features, they are certainly dependent on one another.

Lastly, we combine these POS and parse tree based features with the acoustic features explored in Section 4.3. Results of ten-fold cross-validation experiments using AdaBoost on BDC-read, BDC-spontaneous and BURNC material with a feature vector containing both acoustic and syntactic features can be found in Table 4.17. Speaker independent evaluations using leave-one-speaker-out cross-validation on the BDC corpora can be found in Table 4.18. Note, the Acoustic feature set includes preboundary lengthening features presented in Section 4.3.2 on the BURNC material, where time-aligned phone information is available.

Feature Set	BDC-read	BDC-spon	BURNC
Baseline	87.08% / 0.000	82.21% / 0.000	80.48% / 0.000
Syntactic	92.17% / 0.645	85.63% / 0.412	84.94 / 0.536
Acoustic	95.71% / 0.826	93.01% / 0.809	90.73% / 0.736
S+**A**	95.87% / 0.832	93.75% / 0.834	91.48 / 0.761

Table 4.17: *Evaluation of syntactic and acoustic features of intonational phrase boundary detection using AdaBoost under ten-fold cross-validation. Accuracy and F_1 are reported.*

Feature Set	BDC-read	BDC-spon
Baseline	87.08% / 0.000	82.21% / 0.000
Syntactic	92.08% / 0.635	85.48% / 0.452
Acoustic	94.53% / 0.823	92.39% / 0.786
S+**A**	95.75% / 0.824	92.55% / 0.786

Table 4.18: *Evaluation of syntactic and acoustic features of intonational phrase boundary detection using AdaBoost under leave-one-speaker-out cross-validation. Accuracy and F_1 are reported.*

Under ten-fold cross-validation, the combination of syntactic and acoustic features is able to improve intonational phrase detection performance on BDC-spontaneous and BURNC material. The improvements on both corpora approach significance. On BDC-spontaneous, the 0.74% improvement to accuracy approaches significance at the 0.01 level with p=0.044, and the 0.025 increase in F_1 has an associated p-value of 0.0178. On BURNC material, the improvement to accuracy is also 0.74% while F_1 is increased by 0.025. These differences have associated p-values of 0.00708 and 0.0158, respectively. Under speaker-independent evaluation, there is no significant improvement over the performance of the acoustic feature set.

The accuracy and F_1 obtained by the combination of syntactic and acoustic features using AdaBoost and single split decision trees represent the best intonational phrase boundary described in this chapter. The accuracies on BDC-read and BDC-spontaneous material are above the 93.4% rate of human agreement reported by Pitrelli, et al. [156]. The 91.48% accuracy on BURNC material approaches this rate. However, all results are significantly below the interannotator F_1 of 0.930 reported by Koehn et al. [104]. We see no evidence that this performance is speaker dependent – the accuracy and F_1 on BDC material do not significantly differ under ten-fold or leave-one-speaker-out cross-validation.

The results presented here on BDC-read and BDC-spon represent state-of-the-art performance on this material. On BURNC material, previous work has reported accuracies greater than those reported here, accuracy of 91.48% and F_1 of 0.761 [236, 187]. However, there is a question about the consistency of these results. The BURNC corpus contains repeated lexical material. Care must be taken when evaluating the use of text based features on this material, lest identical lexical material appear in training and testing data. Many number of papers published on the use of lexical and syntactic features in predicting phrase boundaries on BURNC data do not address this issue including [236, 187].

As described in Section 4.2, previous results have reported similar accuracies and in some cases slightly higher F_1 values on other corpora. Schmid and Atterer [177] reported an F_1 of 0.778 in predicting phrase boundaries on MARSEC data, a corpus of standard southern British English. Hirschberg and Prieto [90] reported 95% accuracy on a corpus of newswire text.

4.5 Detection of Intermediate Phrase Boundaries

In the ToBI framework, each intonational phrase is composed of one or more intermediate phrase. Intermediate phrase boundaries are marked by points of less disjuncture than full, intonational phrase boundaries. They are not typically associated with silence. However, pitch and energy reset as well as preboundary lengthening are common markers of intermediate phrase boundaries. There is significantly less agreement regarding the location of intermediate phrase boundaries compared to the agreement on intonational phrase boundary location. Humans agree with respect to the presence of intermediate and intonational phrase boundaries with 89.8% accuracy, compared to the 93.4% agreement on intonational phrase boundaries [156]. Human agreement on the presence of intonational phrase-internal intermediate phrases is closer to 50% [196].

The indicators of intermediate phrase boundaries are similar to those of intonational phrase boundaries, but differ in terms of degree. The acoustic reset is less dramatic, and the preboundary lengthening is less dramatic. In this section we apply the acoustic and syntactic features described in Sections 4.3 and 4.4 to intermediate phrase boundary detection. We omit intermediate phrase boundaries that occur at the end of an intonational phrase from the material used in developing the technique for detecting intermediate phrase boundaries. That is, rather than classify intermediate phrase boundaries from intonational phrase boundaries, we omit all intonational phrase boundaries from both training and evaluation material. This oracular scenario represents the best case scenario for intermediate phrase boundaries. In addition to this evaluation, we evaluate the performance of the intermediate phrase boundary detector using automatically detected intonational phrase boundaries, rather than manual annotations. From the point of view of the hierarchy of prosodic phrasing, this is a top-down approach. Intonational phrases, major phrasal units, are identified first, and their locations constrain the search for intermediate phrase boundaries. This can be viewed in contrast to the hierarchical approach taken by Veilleux and Ostendorf [217, 145], where intonational phrase boundaries are detected using information from hypothesized intermediate phrases.

Neither of the top-down or joint hierarchical approach is *a priori* superior to the other. We opt for the top-down approach due to the success in detecting intonational phrase boundaries. By detecting intonational phrase boundaries first, we can use these high accuracy predictions to aid the prediction of intermediate phrase boundaries – a more difficult prediction task.

We evaluate the use of J48 Decision Trees, Logistic Regression, AdaBoost, SVM with linear and radial basis functions to detect intonational phrase boundaries. We evaluate these algorithms using a feature vector comprised of acoustic and syntactic features using ten-fold cross validation. We perform these evaluations on BDC-read, BDC-spontaneous and BURNC corpora. Results of these decisions are reported in Tables 4.19, 4.20 and 4.21. Recall, intonational phrase boundaries are manually omitted from the training and testing material when performing these evaluations. Thus the results presented in these Tables assume oracular intonational phrase boundary information.

Classifier	Accuracy	F_1
Baseline	91.99%	0.000
J48	92.11% ± 0.410	0.459 ± 0.0279
AdaBoost	92.20% ± 0.0804	0.102 ± 0.0459
Logistic	93.54% ± 0.230	***0.543 ±0.0120***
SVM-linear	***93.94%*** ± ***0.295***	0.509 ± 0.0213
SVM-rbf	92.18% ± 0.0672	0.059 ± 0.0164

Table 4.19: *Evaluation of intermediate phrase boundary detection using acoustic and syntactic features on BDC-read material using oracular intonational phrase boundary information. Accuracy and F_1 are reported.*

Classifier	Accuracy	F_1
Baseline	83.78%	0.000
J48	90.66% ± 0.443	0.502 ± 0.0295
AdaBoost	91.73% ± 0.459	0.527 ± 0.0279
Logistic	91.65% ± 0.295	***0.541*** ± ***0.0213***
SVM-linear	***91.79%*** ± ***0.246***	0.472 ± 0.0180
SVM-rbf	89.70 ± 0.108	0.0830 ± 0.0180

Table 4.20: *Evaluation of intermediate phrase boundary detection using acoustic and syntactic features on BDC-spontaneous material using oracular intonational phrase boundary information. Accuracy and F_1 are reported.*

Classifier	Accuracy	F_1
Baseline	88.72	0.000
J48	87.89% ± 0.410	0.370 ± 0.0110
AdaBoost	88.63% ± 0.410	0.254 ± 0.0262
Logistic	89.50% ± 0.230	***0.394*** ± ***0.021***
SVM-linear	88.72% ± 0.443	0.000 ± 0
SVM-rbf	88.72% ± 0.344	0.000 ± 0

Table 4.21: *Evaluation of intermediate phrase boundary detection using acoustic and syntactic features on BURNC material using oracular intonational phrase boundary information. Accuracy and F_1 are reported.*

On all corpora, the best F_1 in the detection of intermediate phrase boundaries is obtained by Logistic Regression classification. On BURNC the best F_1 is 0.394; on BDC-read this is 0.543, and on BDC-spontaneous it is 0.541. The automatic detection accuracy is greater than the rate of human agreement on both BDC corpora, 93.54 on BDC-read and 91.65% on BDC-spontaneous. However, the F_1 of intermediate phrase boundary detection, between 0.543 and 0.394, is significantly lower than that achieved on the intonational phrase boundary detection task, between 0.834 and 0.761. This task is significantly more difficult. Intonational phrase-internal intermediate phrase boundaries, by definition, fall at points of less severe disjuncture than intonational phrase boundaries. These boundaries are not typically indicated by silence, the most powerful indicator of intonational phrase boundaries.

To determine the influence of syntactic and acoustic features on the detection of intermediate phrase boundaries, we evaluate the two feature sets separately. These evaluations are all performed using Logistic Regression with ten-fold cross-validation. Results of these evaluations on BDC-read, BDC-spontaneous and BURNC material can be found in Table 4.22. We find that acoustic features are significantly more discriminative of intermedi-

Corpus	**A**coustic	**S**yntactic	**A+S**
BDC-read	92.83% / 0.442	92.38% / 0.212	93.54% / 0.543
BDC-spon	91.29% / 0.484	89.73% / 0.161	91.65% / 0.541
BURNC	88.68% / 0.257	88.49% / 0.036	89.50% / 0.394

Table 4.22: *Evaluation of intermediate phrase boundary detection using acoustic and syntactic features on all corpora. Accuracy and F_1 are reported.*

ate phrase boundaries than syntactic features. However, the combination of both feature sets yields significant improvements to F_1. While performing better than chance, the syntactic feature set shows only a minor relationship between parse tree and part-of-speech tag information and intermediate phrase boundaries. There is a significant relationship between relative constituent position and intermediate phrase boundaries. On BDC-read material, intermediate phrase boundaries fall on average at 0.44, approximately the middle of a syntactic constituent, while word boundaries that are not phrase boundaries have an average relative position of 0.16. This difference is significant with $p = 3.33 * 10^{-16}$. We also find that the parse tree distance and distance ratio between consecutive words show a significant difference at intermediate phrase boundaries. On average, words surrounding intermediate phrase boundaries have a mean parse tree distance of 4.2 corresponding to a ratio of 0.87 while consecutive words that do not surround prosodic phrase boundaries have an average parse tree distance of 3.6, a ratio of 0.51. These differences are significant with p=$4.44 * 10^{-16}$ and p=$9.99 * 10^{-16}$, respectively.

We also observe significant differences in a variety of representations of pitch and energy reset relative to the presence on intermediate phrase boundaries in BDC-read material. We find that the mean difference in maximum pitch across intermediate boundaries is 12.42Hz, while across word boundaries that are not phrase boundaries this difference is -1.99Hz, a difference that a t-test determines to be significant with p=$2.227 * 10^{-8}$. The regression based reset shows a mean value of 3.66Hz at intermediate phrase boundary but 0.77Hz at other word boundaries (p=0.002). The slope of the regression fit line preceding intermediate phrase boundaries has a slope of -17.53Hz/sec compared to a slope of -0.096Hz/sec preceding other word boundaries (p=$7.85 * 10^{-5}$). The mean energy across intermediate phrase boundaries differs by -3.10dB, while at other word boundaries this value shows almost no difference, with a mean value of -0.08dB (p=$2.22 * 10^{-16}$). This energy reduction following intermediate phrase boundaries is unexpected given the hypothesis of energy reset at phrase boundaries. However, this finding is most likely due to the fact that intermediate phrase-final words are accented 71.8% of the time, while words following intonational-phrase-internal intermediate phrase boundaries are rarely accented, with a rate of only 33.4%. This influence of accenting behavior, and, possibly, the vocal effort required to produce the phrase accent, are likely causes of this unexpected reset behavior. We see a similar effect in the regression based energy reset, where a mean reset of -1.63dB is observed at intermediate phrase boundaries, and a mean reset of -0.086dB at other word boundaries (p=$1.63 * 10^{-12}$). These acoustic reset features are able to detect intermediate phrase boundaries with accuracy that approaches or exceeds the rate of human agreement reported by Syrdal [196], approximately 50%, on all three corpora. When an evaluating the interannotator agreement on a subset of the BURNC, Ostendorf et al. find that the labelers agreed on break indices with 95% accuracy [143]. However, this evaluation does not isolate the agreement on intermediate phrase boundary locations indicated by a break index of '3'. The inclusion of syntactic features does not significantly improve accuracy of automatic intermediate phrase boundary detection except on BURNC material where the difference approaches significance with $p = 0.0278$. However, the F_1 is improved significantly on BDC-read material with p=0.000751, BURNC material with p=$3.88 * 10^{-5}$, and approaches significance with p=0.028 on BDC-spontaneous.

With some success in the automatic detection of both intonational and intermediate phrase boundaries, we evaluate the combination of these two detection systems. Under this evaluation, we first detect intonational phrase boundaries, and then detect which of the tokens not predicted to be intonational phrase boundaries are intermediate phrase boundaries. This two-stage classification system uses the best performing phrase boundary detectors on each of these tasks. We find the best intonational phrase boundary detection to be achieved with AdaBoost on single split decision trees, while Logistic Regression yields the best performing intermediate phrase boundary detection. We evaluate this two-stage phrase boundary detection scheme using ten-fold cross-validation on BDC-read, BDC-spontaneous and BURNC material. Contingency tables and results of these evaluations are reported in Tables 4.23, 4.24 and 4.25. We also evaluate the speaker dependency of this approach by perform leave-one-speaker-out cross-validation on the BDC corpora. These contingency tables and results can be found in Tables 4.26 and 4.27.

	Actual Class		
Hypothesis	Intonational	Intermediate	None
Intonational	1112	104	57
Intermediate	182	243	186
None	104	409	8432

Table 4.23: *Contingency Table for two-stage phrase boundary detection, based on ten-fold cross-validation on BDC-read. Overall Accuracy=90.38% ± 0.262. Intonational Phrase F_1: 0.832 ± 0.0166 Intermediate Phrase F_1: 0.356 ± 0.0205*

	Actual Class		
Hypothesis	Intonational	Intermediate	None
Intonational	1766	274	82
Intermediate	240	242	180
None	178	487	8176

Table 4.24: *Contingency Table for two-stage phrase boundary detection, based on ten-fold cross-validation on BDC-spontaneous. Overall Accuracy=87.60% ± 0.5412. Intonational Phrase F_1: 0.820 ± 0.0122 Intermediate Phrase F_1: 0.290 ± 0.0241*

	Actual Class		
Hypothesis	Intonational	Intermediate	None
Intonational	3915	468	489
Intermediate	448	436	434
None	1303	1793	20291

Table 4.25: *Contingency Table for two-stage phrase boundary detection, based on ten-fold cross-validation on BURNC. Overall Accuracy=83.31% ± 0.361. Intonational Phrase F_1: 0.743 ± 0.00984 Intermediate Phrase F_1: 0.217 ± 0.0129*

On all corpora, the performance in detecting full, intonational phrase boundaries is dramatically higher than the intermediate phrase boundary detection. Using this top-down classification technique, on the BDC material, it is relatively uncommon for intonational phrase boundaries to be misclassified as non-phrase boundaries, and vice versa. However, intermediate phrase boundaries are misclassified as either of the other classes, and both non-phrase boundaries and intonational phrase boundaries are frequently misclassified as

	Actual Class		
Hypothesis	Intonational	Intermediate	None
Intonational	1108	112	57
Intermediate	124	113	197
None	167	531	8422

Table 4.26: *Contingency Table for two-stage phrase boundary detection, based on leave-one-speaker-out cross-validation on BDC-read. Overall Accuracy=89.03% ± 0.656. Intonational Phrase F_1:0.828 ± 0.0428 Intermediate Phrase F_1: 0.190 ± 0.0807*

	Actual Class		
Hypothesis	Intonational	Intermediate	None
Intonational	1664	250	67
Intermediate	255	219	525
None	266	534	7847

Table 4.27: *Contingency Table for two-stage phrase boundary detection, based on leave-one-speaker-out cross-validation on BDC-spontaneous. Overall Accuracy=83.68% ± 3.641. Intonational Phrase F_1: 0.806 ± 0.0385 Intermediate Phrase F_1: 0.210 ± 0.0672*

intermediate phrase boundaries. This is not surprising. Human annotators agree with accuracy of approximately 50% regarding the presence of intermediate phrase boundaries [196] and there is anecdotal evidence that level ‘3’ break indices (intermediate phrase boundaries) are commonly confused with both ‘2’s (non-phrase boundaries) and ‘4’s (intonational phrase boundaries) [156]. We see an increased error rate overall on the BURNC evaluation. In particular we see an increase in “missed” intonational phrase boundaries – where an intonational phrase boundary is misclassified as a non-phrase boundary.

Overall, however, this two stage classification technique yields high three-way classification accuracy, from 83.31% on BURNC to 87.6% BDC-spontaneous up to 90.38% on BDC-read. The speaker-independent evaluation is somewhat lower at 89.03% on BDC-read and 83.68% on BDC-spontaneous.

While Yoon [236] has reported higher three-way classification results on BURNC – 88.06% accuracy with intonational phrase boundary F_1 of 0.832 and intermediate phrase boundary F_1 of 0.345 – these results are flawed by failure to exclude story repetitions from test data. Yoon’s results were obtained using TiMBL, a memory based learning algorithm, with feature including word identity, position from the end of the current sentence, part of speech and grammatical relation. These features are likely to be sufficient to uniquely identify a word in a story. The memory based learning algorithm, therefore, essentially applied the phrasing from one reading of a news story to in the training material to the reading of the same story by another reader in the evaluation data. Thus, these results do not give a reliable estimate of how well the approach would generalize to unseen material; at best it indicates how well the approach performs when different readers read the same material.

Particular improvements could be made to the intermediate phrase boundary detection performance. Within this top-down hierarchical detection structure, the intermediate phrase boundary detector could benefit from features derived from the hypothesized intonational phrase boundaries. Also, it is worth investigation to determine if a top-down, hierarchical (cf. [217]) or a three-way classification (cf. [236]) is a more reliable phrase boundary detection scenario. Alternately, a hybrid, iterative approach may be helpful, where hypotheses from intonational and intermediate phrase boundary detectors are combined to identify the best prosodic phrasing results.

4.6 Conclusion and Future Work

In this chapter, we present a number of approaches to prosodic phrase boundary detection. We divide these experiments into acoustic (cf. Section 4.3) and syntactic (cf. Section 4.4) approaches. A summary of the most significant results from this chapter is presented in Table 4.28.

In the acoustic investigations, we confirmed the importance of silence in the detection of intonational phrase boundaries and presented a number of representations of acoustic reset. The use of linear regression fit lines to generate reset features is a novel contribution of this work. While linear regression has been used for this task by Batliner et al. [11], its

		Detection Approach	BDC-read	BDC-spon	BURNC
Intonational Phrase	AdaBoost	Silence	95.46% / 0.819	91.94% / 0.794	88.22% / 0.595
		Pitch and Intensity Prev. Word	90.46% / 0.538	85.54% / 0.591	81.21% / 0.387
		Pitch and Intensity Difference	89.47% / 0.460	83.64% / 0.408	82.80% / 0.200
		All Pitch and Intensity	94.82% / 0.766	91.86% / 0.753	88.90% / 0.669
		All Acoustic Features	95.65% / ***0.883***	93.13% / 0.810	88.89% / 0.647
		All Acoustic Features w/ Preboundary Lengthening	NA	NA	90.73% / 0.736
		Parse Tree Features	92.87% / 0.662	85.81% / 0.505	86.23% / 0.591
		Raw POS bigrams	87.39% / 0.245	81.76% / 0.202	81.63% / 0.288
		Raw POS surrounding bigrams	90.02% / 0.670	84.63% / 0.600	86.12% / 0.547
		All Syntactic Features	92.17% / 0.645	85.80% / 0.506	86.57% / 0.611
		All Features	***95.87%*** / 0.832	***93.75% / 0.834***	***91.48% / 0.761***
Intermediate Phrase	Logistic	Acoustic Features w/ oracular IP[a]	92.83% / 0.442	91.29% / 0.484	88.68% / 0.257
		Syntactic Features w/ oracular IP[a]	92.38% / 0.212	89.73% / 0.161	88.49% / 0.036
		All Features w/ oracular IP[a]	***93.54% / 0.543***	***91.65% / 0.541***	***89.50% / 0.394***
		All Features w/ hyp. IP[b]	90.38% / 0.356	87.60% / 0.290	83.31% / 0.217

[a] Accuracy of Intermediate Phrase boundary detection. These results assume 100% Intonational Phrase boundary detection accuracy.

[b] Overall accuracy of Intonational and Intermediate Phrase boundary detection.

Table 4.28: A summary of significant phrase boundary detection experiments. All evaluations use ten-fold cross-validation. Classification performance is reported using Accuracy and F-measure. The best intermediate and intonational phrase boundary detection performance on each corpus on are presented in bold.

use in this chapter is significantly more sophisticated. One of the results that we observe in our experiments that has not been discussed in much detail elsewhere is the phenomena of energy reset at phrase boundaries. The pitch declination phenomenon has been well documented and studied from many perspectives. The energy correlate of "declination" has been largely unexplored. However, we find representations of reset that involve energy features perform better than similar representations of pitch declination and reset. Moreover, we find that reliable reset features can be extracted from a very narrow window surrounding a candidate boundary. Specifically we find that examination of only 20ms (two 10ms frames) of energy information was able to detect intonational phrase boundaries with performance better than subword windows. That said, reset features calculated over the whole word are more reliable than these narrow windowed features. Changes in speaking rate are believed to be a perceptual correlate to the perception of phrase boundaries. Speaking rate and speech rhythm are acoustic qualities that have not been explored in this thesis and which remain an area for future research.

These contributions along with the use of many confirmed acoustic indicators of the presence of intonational phrase boundaries leads to state-of-the-art performance on BDC material, and competitive performance on the BURNC corpus. On BDC-read under speaker-independent evaluation, we detect intonational phrase boundaries with 95.87% accuracy and a corresponding F_1 of 0.832. On BDC-spontaneous material, the best reported accuracy is 92.55% with a 0.786 F_1. Evaluating the performance on BURNC material, the highest observed speaker-independent accuracy is 90.00% with an F_1 of 0.700 using only acoustic features.

Yoon reported best previous results in the automatic detection of intonational phrase boundaries on BURNC – 92.23% accuracy and an F_1 of 0.861 [236]. However, this and other evaluations [189, 187, 4] *may* be fundamentally flawed. While not exactly one-to-one, there is a strong relationship between lexical content and prosodic phrasing. The BURNC corpus contains multiple speakers reading the same news stories. Therefore, when using text based features such as syntactic information, POS tags, or word identity, it is critical to ensure that no news story appears simultaneously in training and evaluation material. Otherwise, the evaluation measures how consistently two or more readers of broadcast news produce the same material, rather than the ability to generalize about the phrasing of unseen material.

In the work described in this chapter, the inclusion of syntactic indicators of phrase boundaries only modestly improves detection performance. Most approaches to syntactic-prosodic modeling use a standard *N*-gram modeling approach. We find that a syntactic model that incorporates not only part of speech tags from words *preceding* candidate boundaries, but also *following* words is significantly better at detecting phrase boundary locations. This is a potentially helpful contribution toward future research on prosodic assignment.

Towards detecting intermediate phrase boundaries, in addition to full, intonational phrase boundaries, we examine a top-down detection approach. Under this approach, intonational phrase boundaries are detected first, and intermediate phrase boundary detection is then performed on the remaining tokens. In the future, we intend to compare the performance of this approach to the hierarchical approach described by Veilleux and Ostendorf [217, 145], as well as simultaneous detection of intermediate and intonational phrase boundaries under a 3-way classification approach. These approaches take advantage of the relationship between prosodic events, in a way that is not thoroughly explored in this thesis. In addition to the hierarchical structure of prosodic phrasing, we also have observed that there is a relationship between accent placement and intermediate phrase boundaries. Accented words tend to precede, but not follow intonational phrase-internal intermediate phrase boundaries. The integration of these disparate detection and classification tasks is an area of future research.

4.6.1 Key Observations

- **Silence is the single most powerful indicator of intonational phrase boundaries.**

  While instances of intonational phrase boundaries that do not coincide with silence, it is fairly uncommon for silence to be present within an intonational phrase.

- **Energy is more reliable for detecting phrase boundaries than pitch is.** Phrasing is indicated by perceived disjuncture, commonly signaled by acoustic reset. Features representing a change in intensity are more discriminative to the presence of phrase boundaries than similar features capturing changes in pitch.

- **Part-of-speech tags that surround a candidate boundary are consistently superior to traditional n-gram features.** Modeling the likelihood of a phrase boundary falling between two part-of-speech tags (surrounding bigram) is dramatically better than modeling the likelihood that a phrase boundary falls after a sequence of two part-of-speech tags.

- **Detection of intermediate phrases is more difficult than intonational phrases.** Intermediate phrases are indicated by a smaller degree of disjuncture than intonational phrases. This difference in degree may lead to intermediate phrasing being a fundamentally subtler phenomenon than intonational phrasing. This difference may also lead to reduced human annotation consistency which in turn would limit automatic detection performance.

Chapter 5

Pitch Accent Type Classification

5.1 Introduction

In Standard American English (SAE), words can be accented in different ways. The most common realization is with an increase in pitch up to a point high in the speaker's pitch range. The ToBI standard describes this as an H* accent. During the discussion of the identification of pitch accent, we describe accenting as a phenomenon that draws the listener's attention to a concept (cf. Chapter 3). The use of different pitch accent types has the effect of drawing a listener's attention in different ways. For example, sharp rising accents (L+H*) commonly have the effect of indicating contrast or increased salience [22, 178, 134]. Low toned accents are more often used to accent concepts that are already in the discourse space, representing 'given' information [31].

The ToBI prosodic annotation framework describes intonation as a series of high (H) and low (L) tones which are associated with prosodic events [183]. The annotation standard defines five pitch accent types. These include the basic H*, L* accents. H* accents have an associated high tone and L* accents are produced with a low pitch. In addition to these, there are three complex tones, L+H*, L*+H and H+!H*. L+H* accents are characterized by a low tone followed quickly by a high tone, while L*+H are roughly defined as a scooped accent which begins at an accented low tone, followed by an unaccented high tone. In these two complex tones, L+H* and L*+H, the starred tone is aligned with the lexically stressed syllable of the word. The fifth accent type, H+!H*, is characterized by a sharp pitch drop from a previously unaccented high tone. In addition to these, high tones are produced in a compressed pitch range, i.e. compressed from a previous high tone. These "downstepped" high tones are indicated as !H, and each accent produced with a high tone, can be produced with a downstepped high. The downstepped phenomenon is also called "catathesis". This leads to the additional accent types, !H*, L+!H*, and L*+!H.

Pitch accent types are difficult to classify automatically for two reasons. First, the distinctions are subtle. Even human agreement on this task is fairly low. While the location of pitch accent is fairly high, agreement on pitch accent type varies significantly. Pitrelli, et al. [156] report human agreement of only 64.1% on accent classification in the ToBI framework. If downstepped variants of accents are collapsed with their non-downstapped forms this agreement improves to 76.1%. Second, pitch accents are overwhelmingly H* in most labeled corpus, including the BDC and BURNC material used in this thesis. This skewed class distribution leads to a very high baseline, at or above the rate of human agreement. The skewed distribution of accent types also also leads to a dearth of training data available for the minority classes.

There is agreement that accent shape and alignment impacts the interpretation and communicative effect of accenting a word. However, the existence of discrete pitch accent types is mildly controversial. Taylor [202] lays out the argument against the categorical point of view in the presentation of his Tilt model, a continuous parameterization of intonational events. Arguing against the segmental phonology view implicit in the ToBI and other intonational theories, he draws a parallel with phone perception. The argument for the existence of a discrete segmental phonology space is supported by the presence of minimal pairs, such as /b/ and /p/. The voicing of the /b/ in "bill" can be incrementally reduced such that the phone eventually becomes perceived as a /p/ in "pill". There is an acoustic continuum on which these two phones lie. However, at no point does the perception of the word evoke a combined semantic representation that is partially "bill"-like and partially "pill"-like. There is a perceptual boundary prior to the interpretation of the phone or word. Taylor goes on to claim that there is no evidence of such a boundary between pitch accent types. However, Ladd and Morton [111] found evidence of abrupt, categorical, shifts in the *interpretation* of "normal" accent peaks and "emphatic" accent peaks. Subjects, however, showed no evidence of categorical perception in rating *how* emphatic accents were. The perception of degree of emphasis appeared continuous, while the interpretation, categorical. The Ladd and Morton study is a perceptual experiment, evaluating how human subjects interpret pitch accents. Pierrehumbert and Steele [155] found evidence that subjects *produce* categorical accent types. In this study, subjects heard rise-fall accents where the alignment between f0 peak and vowel onset was varied in small increments. Subjects were then asked to imitate the contour which they heard. Analysis of the subject's imitated productions revealed that subjects produced two categories of accenting. While this does not reveal whether subjects *perceived* accent as a binary classification, it does provide evidence that "tonal alignment" is a binary distinction in SAE intonation. These two studies provide some empirical evidence of the existence of accent categories. Evidence of the type that Taylor claimed did not exist. A claim could be made that "contrastive-ness" or incredulity could be perceived on a continuum with more L+H*-like or L*+H-like accents indicating greater degrees of contrast or incredulity, respectively. However, the hypotheses of Pierrehumbert and Hirschberg [154] require a categorical representation for their application. They hypothesize that different accent types determine how the accented word is interpreted in discourse. For example, words bearing H* accents are generally 'new' to the discourse and should be instantiated in the listeners mutual belief space. L* accents make entities that are already in the mutual belief space more salient or in the case of questioning, inhibit information from being added to the mutual belief space. There is no representation of degree in this semantic representation. That is, there is no sense that a term that is accented with an accent half-way between a H* and L* might have the effect of being partially introduced to the mutual belief space of the hearer.

In this chapter we present experiments that attempt to automatically classify pitch accent type. In Section 5.5, we present a descriptive analysis of the acoustic qualities of pitch accent types, focusing on the differences between the two most common accent types, H* and L+H*. In Section 5.6, we investigate the use of acoustic features capturing pitch contour shape information. In Section 3.4, we find that word-level pitch accent detection consistently outperforms syllable-level approaches, all else being equal. While this implies that there is acoustic material informative to the presence or absence of accent in a word outside the lexically stressed syllable, this additional acoustic material likely introduces noise to pitch accent **type** classification. To evaluate this assumption, we explore different regions of analysis for pitch accent type classification in Section 5.6.3. Training data sampling can be used to overcome skewed class distributions in supervised learning. In Section 5.6.4, we evaluate the impact of some commonly used sampling techniques on pitch accent type classification. Phrase ending intonation (cf. Chapter 6) can modify the shape of a pitch accent. The two phenomena both affect the shape of the pitch contour, and when the accent is phrase-final, their influence may overlap. We examine the influence the presence of phrase ending intonation has on automatic pitch accent type classification in Section 5.6.5. We make some observations about the interaction of part-of-speech based word class information and pitch accent type in Section 5.6.7. We conclude and describe future work in Section 5.7.

5.2 Related Work

In this chapter, we classify pitch accents into five categories based on the ToBI pitch accent inventory: H*, L+H*, L*, L*+H, H+!H*. In general, we do not distinguish downstepped tokens from their non-downstepped variants – though we do briefly address the classification of H* versus !H* in Section 5.6.2. This classification task, formulated as such, has not been performed previously in the literature. However, other work has addressed the classification of pitch accent type in other ways.

One technique that has been used in a number of research efforts is to simultaneously detect and classify pitch accent. This is commonly done by representing pitch accent detection and classfication as a four-way classification task, where each word, syllable or accent may be classified as UNACCENTED, HIGH, LOW, or DOWNSTEPPED. This accent inventory does not allow a system to distinguish H* from L+H* accent types. The complex accent type, L+H*, is associated with increased salience and contrast. Two factors that are potentially valuable for SLP tasks. This technique was explored by Ross and Ostendorf [173] in investigating prosodic assignment techniques. That is, the input features explored in this classification were exclusively lexically based, capturing part of speech information, positional, and structural information, as well as a representation of information status. When evaluated on one BURNC speaker (f2b) a decision tree trained was able to classify accent types with 72.4% accuracy over the 71.8% majority class baseline. While the classifier performed a four-way classification, this performance reflects the classifier accuracy in differentiating accent types on syllables that, in fact, bear accent. Incorporating acoustic information, Ostendorf and Ross [144] address the same classification task using stochastic models. This modeling structure predicts the joint probability of pitch accent and boundary classes, simultaneously combining information from acoustic and lexical streams. The approach described the combination of acoustic – pitch range, pitch contour and duration – and phonotactic models for classification. Isolating the three-way, HIGH, LOW, or DOWNSTEPPED, classification performance, and assuming the presence on intermediate phrase boundaries, the technique classifies pitch accent with 83.4% accuracy, over a baseline of 77.0%. This performance, is again, achieved on a the speech from a single BURNC speaker (f2b).

This four-way simultaneous classification and detection approach was also used by Sun [194]. In this work, boosting and bagging CART decision trees were used for prosodic analysis. Similar to [173, 144] this approach was evaluated on a single BURNC speaker (f2b). Using only acoustic features, boosted CART trees were able to classify accented words with 74.3% accuracy, while incorporation of text features improved the performance to 79.5%. Levow also used this four-way classification for pitch accent detection and classification under supervised [115], and unsupervised and semi-supervised learning approaches [116]. Using SVMs with context sensitive acoustic features, four-way classification accuracy of 81.3% accuracy at the syllable level is achieved. Using unsupervised spectral clustering, 78.4% accuracy is reported, while using the semi-supervised technique, Laplacian SVMs 81.5% accuracy is achieved. While this work is also evaluated on the same BURNC speaker, f2b, the results do not isolate the accent classification from detection performance. This makes is impossible to directly compare the performances of these techniques with previous techniques from a strictly pitch accent classification point of view.

Read and Cox [165] also used this four-way detection and classification approach for evaluating three prosodic assignment techniques: dynamic programming (memory based learning), ngrams (language modeling), decision trees. They used part of speech and syntactic parse tree features. These included representations of structural and identity information such as parse tree distance between tokens, surface position, and lexical stress of surrounding syllables, as well as vowel and syntactic constituent identity. This work implicitly acknowledged the impact of the skewed pitch accent class distribution on the evaluation of classification techniques. The authors evaluated their four-way classification performance using the *Balanced Error Rate*. This measure is the average recall over each class. In Equation 5.1, C is the number of classes, and M_ij is the contingency matrix representing the number of times a token of class i was recognized as class j.

$$BER = \frac{1}{C}\sum_i \frac{(\sum_j M_{ij}) - M_{ii}}{\sum_j M_{ij}} \tag{5.1}$$

By equally weighting the contribution of each class BER increases the contribution of minority class performance. This may be justifiable in the case of such a skewed class distribution. We address this issue of pitch accent type classification evaluation in more detail in Section 5.6.1.

Ananthakrishnan and Narayanan [5] used RFC (Rise/Fall/Connection) [200] and Tilt [202] parameters along with word and part of speech language modeling to classify pitch accents as H*, !H*, L+H* or L*. It is unclear how other pitch accent types are addressed in this work. The inclusion of the complex tone, L+H* makes this evaluation most similar to the approach we take in this chapter. The major difference is the presence of the downstepped high accent (!H*); the other two complex tones, L*+H and H+!H*, are infrequently observed. The best performing approach reported in this work used only RFC parameters in a multi-layer perceptron model (MLP). When evaluated on six BURNC speakers using leave-one-speaker-out cross-validation, accuracy of 56.4% over a majority class baseline of 54.0% was obtained.

There has been a fair amount of research on classifying pitch accent types in languages other than English. While the accent type inventories in other languages may differ from those found in SAE, the approaches explored may be successfully applied to accent type classification in SAE speech. Fach [52] described work on the classification of accents in German as high or low using Hidden Markov Models (HMMs). A forward looking series of pitch (f0) estimates were used as input to the HMMs. Evaluated on 60 sentences spoken by a single female speaker, with a lookahead of three 10ms pitch frames, 81% accent type classification accuracy was reported. Obviously the language difference makes comparison with other results impossible, but this work suggests that HMM modeling, or other sequential models may be able to capture contour shapes with some success. Greek ToBI [6] annotation uses a similar inventory of pitch accent types as English, though they are used with much more equal rates. The majority class on her Modern Greek corpus is L*+H, with 30% of words bearing this accent type; compare this to the observation that nearly 80% of words in the BDC corpora are accented with H* and !H* accents. Xydas et al. [234] used the wagon [203] decision tree learning algorithm to simultaneously predict accent location and type on a corpus of 516 Modern Greek utterances describing museum exhibits. Using part of speech and structural features, punctuation and phrasing information, they were able to correctly classify 71.9% of pitch accent types.

The approaches described above relied on manually labeled annotation of accent types. There are intonation standards which describe accent types using continuously parameterized functions, among these the Tilt [202] and Fujisaki [59] models. Some research has been done to identify an inventory of categorical accent types directly from pitch contour data. This approach involves identifying accents and then automatically clustering the pitch contours into a taxonomy in an unsupervised manner. Oliver [142] clustered accents into three classes using Self-Organizing Maps in Polish, while Iwano [98] clustered Japanese accents into 11 contour types. Both of these approaches used accent type clustering as an intermediate representation for performing another task. Iwano found that syntactic boundary detection can be improved by the use of accent clusters. Oliver hypothesized that accent clusters could be used for prosodic assignment for speech synthesis applications. However, evaluation of such an application was not reported.

5.3 Examples of Pitch Accent Types

In this section, we present examples of each pitch accent type in the ToBI standard: H*, L*, L+H*, L*+H, H+!H*, and downstepped variants, !H*, L+!H* and L*+!H*. In each of the figures below, we will present a relatively clear example, and an example which is somewhat more difficult to identify or confusable with another accent type. All examples are drawn from the BURNC data and are spoken by speaker f2b. In all examples, the accent bearing syllable is highlighted, and the vertical bar marked the location at which the accent is annotated. In the images, the blue line represents the pitch of the speech signal. The green line represents the intensity.

The examples of H* accents can be found in Figures 5.1 and 5.2. In the clear example we can that the accent falls on the vocalic region with both the highest pitch, and greatest intensity. The sibilant /s/ has higher intensity, but no pitch. In contrast, the confusable example has a pitch peak that is approximately the same height as the surrounding syllables, and is spoken with consistent intensity. This word may have been identified as being accented due to its duration, which is slightly lengthened. It is possible that this accented is annotated as an H* accent as much as a default than anything else. It is not produced in the low region of the speaker's pitch range, preventing it from being a L*, but it does not have an associated pitch excursion.

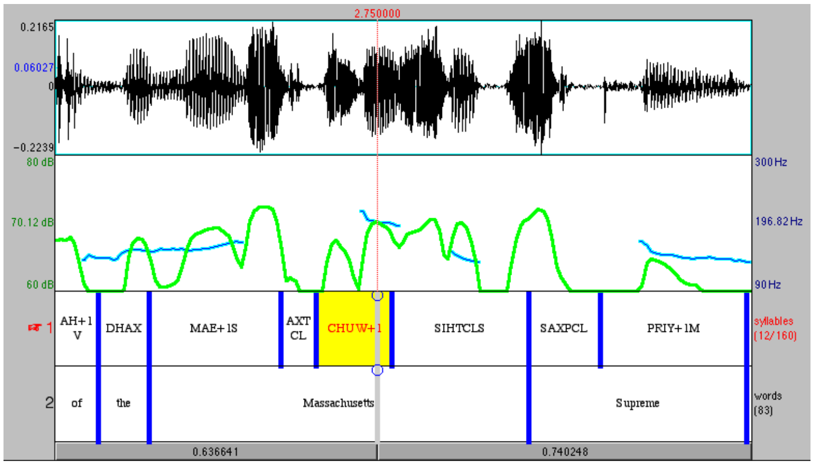


Figure 5.1: *Clear Example of H* accent.*

Downstepped H* accents, marked as !H*, appear similar to H* accents, but are produced in a compressed pitch range. Downstepped accents are only produced following a previous high tone. The previous high tone defines the initial pitch range. The compressed pitch range which identifies an accent as downstepped is defined relative to the previous high tone. In Figure 5.3 an example of a !H* accent can be observed. The red vertical line represents the downstepped accent. The previous H* accent lies within the syllable shaded grey. Due to their compressed pitch range, it can be difficult for labelers to distinguish !H* accents from L* accents. This can be seen in the example in Figure 5.4. In this example, the pitch peak of the !H* accent reaches only 104Hz – a height that is towards the bottom of the speaker's pitch range.

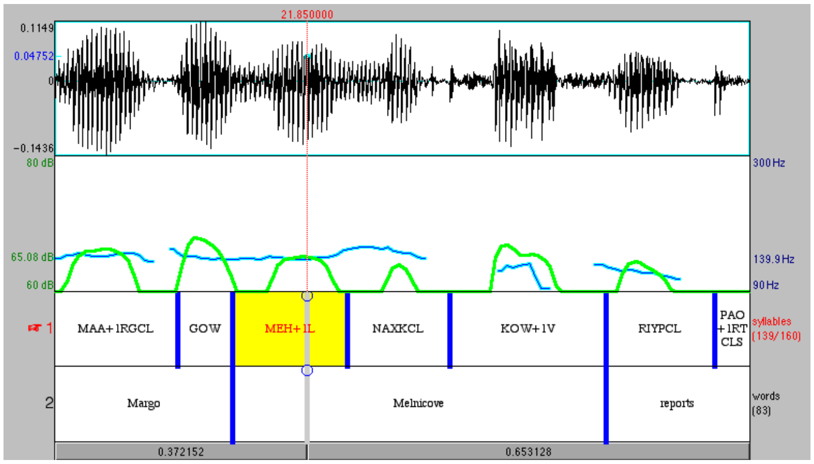


Figure 5.2: *Confusable Example of H* accent.*

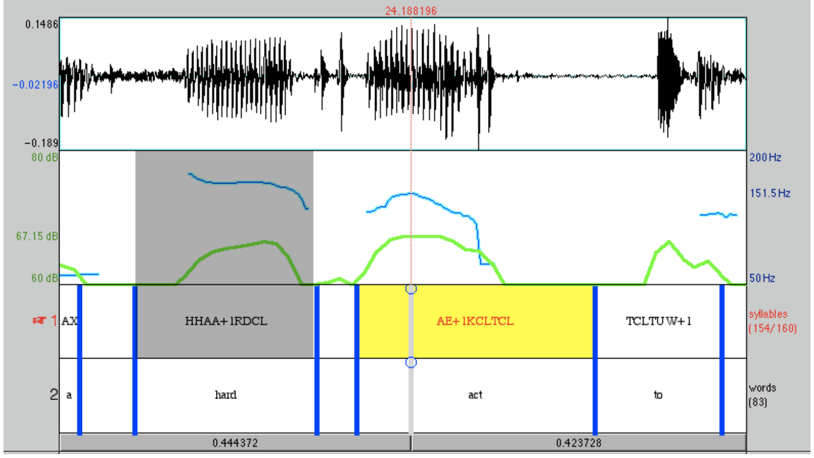


Figure 5.3: *Clear Example of !H* accent.*

L* examples can be found in Figures 5.5 and 5.6. In the clear example, the accented syllable is longer than surrounding syllables, and has a clear intensity peak. These make it clear that the syllable is accent-bearing. However, its pitch is quite low in the speaker's pitch range. The biggest problem with L* accents, isn't being able to differentiate them from H* accents, but rather, to identify that they are accented at all. This ambiguity can be observed in the confusable example. The accented syllable is quite short, and does not have an associated intensity excursion. While its pitch is low, there is very little acoustic indication that this syllable is accent bearing.

Two examples of L+H* accents can be found in Figures 5.7 and 5.8. The ToBI Anno-

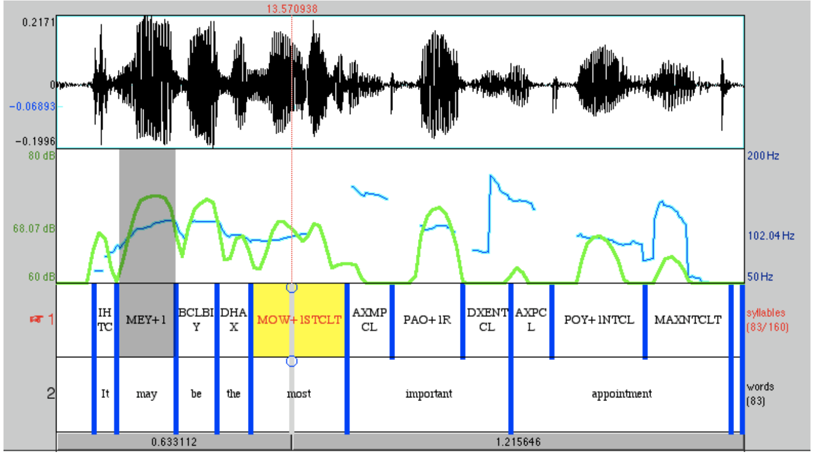


Figure 5.4: *Confusable Example of !H* accent.*

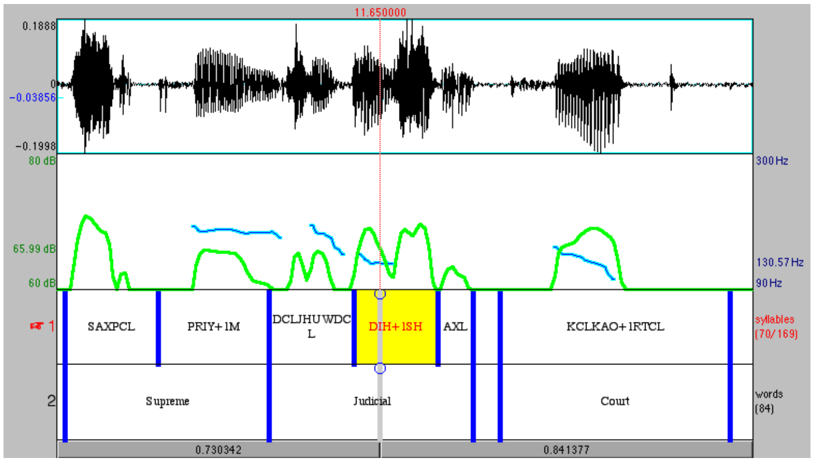


Figure 5.5: *Clear Example of L* accent*

tation Conventions [85] describe this accent type as a “rising peak accent – a high peak target immediately preceded by relatively sharp rise from a valley in the lowest part of the speaker’s pitch range.” In the clear example, this rise is clear. Also observable, and common to L+H* accents, is a pitch peak that falls after the vowel onset. In L+H* examples, the pitch peak is often later in the syllable than in H* accents. We examine the differences between H* and L+H* accents in greater detail in Section 5.5.1. The confusable example, on the other hand, has a relatively shallow rise to its pitch peak. This rise does not begin in “a valley in the lowest part of the...pitch range” – even ignoring the pitch halving error in the previous syllable. Also, the pitch peak shows no displacement. This L+H* accent is

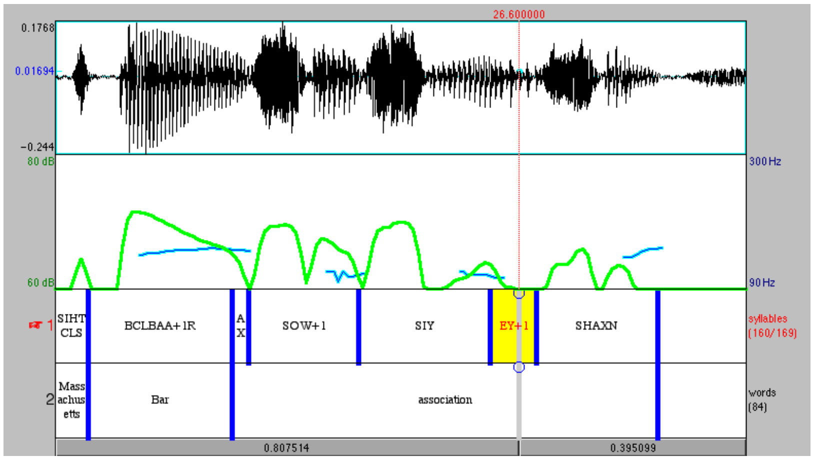


Figure 5.6: *Confusable Example of L* accent.*

difficult to distinguish from a standard H* accent.

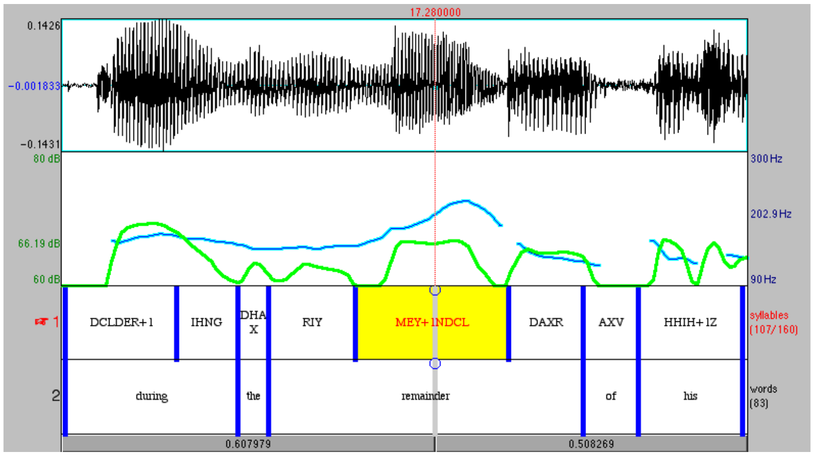


Figure 5.7: *Clear Example of L+H* accent.*

L+H* accents can be produced in a compressed pitch range. These accents are annotated as L+!H*. They share the right displaced pitch peak and relatively sharp pitch rise characteristics with the L+H* accent. However, the pitch peak reaches a maximum at a significantly lower value than a previous high tone. An example of this accent type appears in Figure 5.9. In this case, “reaches” is accented with an H* accent and has a maximum pitch of approximately 240Hz. The maximum pitch within “mandatory” is only 193Hz, a significant compression. In the BURNC and BDC material, L+!H* accents are very infrequently used. The similarities that may make an L+H* accent confusable with an H* accent may make

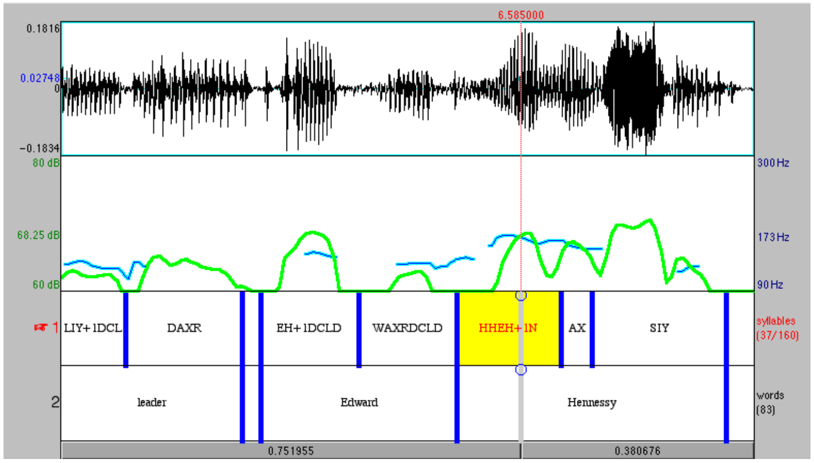


Figure 5.8: *Confusable Example of L+H* accent.*

an L+!H* accent appear similar to a !H* accent. For example, in Figure 5.10, the L+!H* accent does not display the characteristic pitch rise. Rather the second syllable in “sweeping” is produced low in the speakers pitch range followed by the relatively higher “court”. To describe the increase up to the peak in the second syllable, this accent is annotated as L+!H*; the downstep diacritic (!) is used to indicate that the pitch peak is significantly lower than that of the previous H* accent on the “sweeping”.

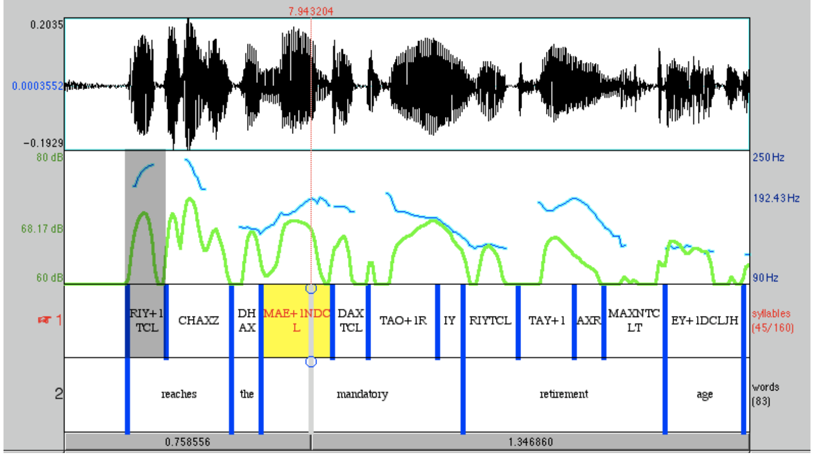


Figure 5.9: *Clear Example of L+!H* accent.*

L*+H accents are quite uncommon in the BDC and BURNC material (cf. Section 5.4). The ToBI Annotation Conventions describe this accent type as a “scooped accent – a low tone target on the accented syllable immediately followed by relatively sharp rise...”.

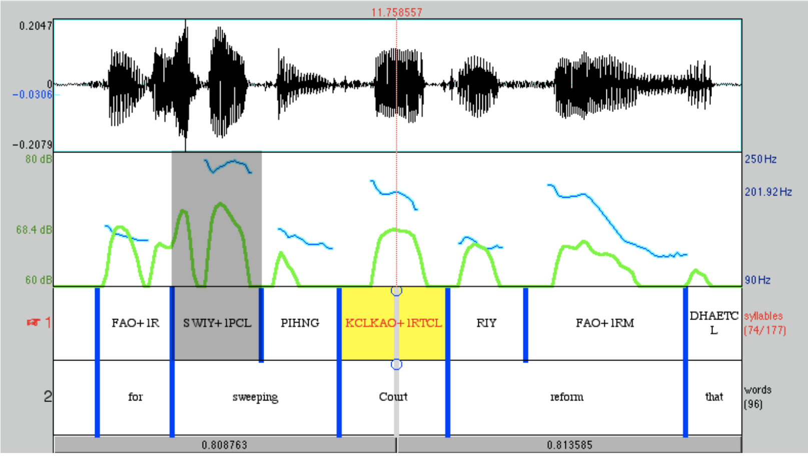


Figure 5.10: *Confusable Example of L+!H* accent.*

The clear example (cf. 5.11) shows this pitch rise within the accented syllable. The H tone following the accented L typically occurs on unaccented speech following the accent. Because of this phenomenon, it is possible for the peak of an L*+H accent to occur in a subsequent unaccented syllable. Associating this pitch peak to the complex L*+H accent is a source of difficulty in identifying L*+H accents and differentiating them from L* accents. The confusable example (cf. 5.12) has a pitch peak within the accented syllable, making it quite identifiable. However, the pitch peak in the confusable example does not occur until two syllables following the accented "one". Moreover, the slope to this peak is fairly shallow. Due to the intensity of the syllable containing the pitch peak, this pitch peak may be associated with an H* or L+H* accent, as opposed to the H tone of an L*+H on a previous syllable.

While L*+H accents are uncommon, L*+!H accents are more so. There are only four instances of this accent type in the BURNC material, and six across both the BDC corpora. The clear example shows all of the characteristics of a typical L*+H accent. The pitch is low at the vowel onset and the pitch contour shows a "scoop" up to a high tone following. In this case the following high tone is in a significantly compressed pitch range compared to the previous H* accent. In the confusable example, the scoop up to a high tone is difficult to notice. This would suggest that this accent is a L* accent. It is unclear why the labeler

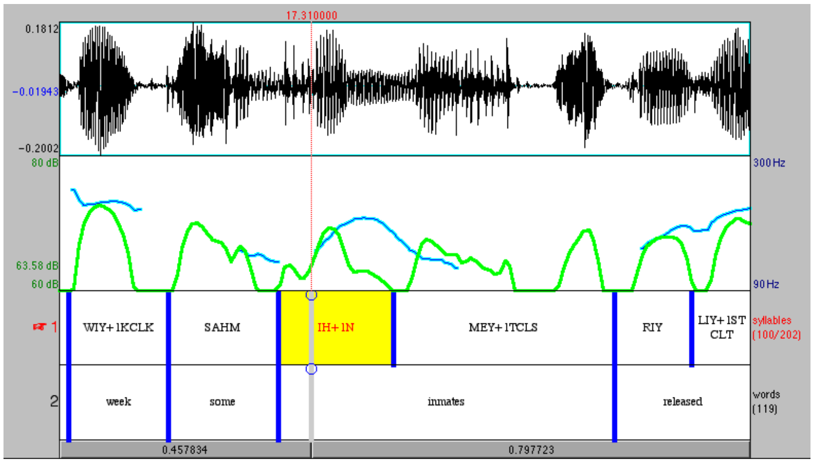


Figure 5.11: *Clear Example of L*+H accent.*

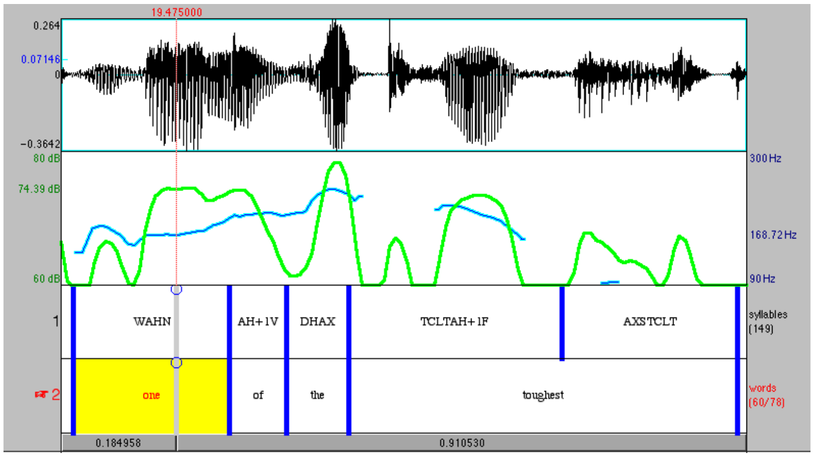


Figure 5.12: *Confusable Example of L*+H accent.*

considered this accent to be L*+!H. Our best hypothesis is that the relatively modest pitch movement within the accented syllable was considered sufficient to be annotated as a high tone produced in an extremely compressed pitch range.

The fifth main pitch accent type is also quite infrequent in BDC and BURNC material. The H+!H* accent is an accent that falls on a syllable with downstepped pitch, where the previous high pitch cannot be described by a previous H tonal element, on a previous pitch accent, a phrase accent or boundary tone. These accents tend to have a sharply falling pitch over the duration of the accented syllable. The accent includes a high tone and a fall aligned with the stressable syllable of the accented word. Examples of this accent can be found

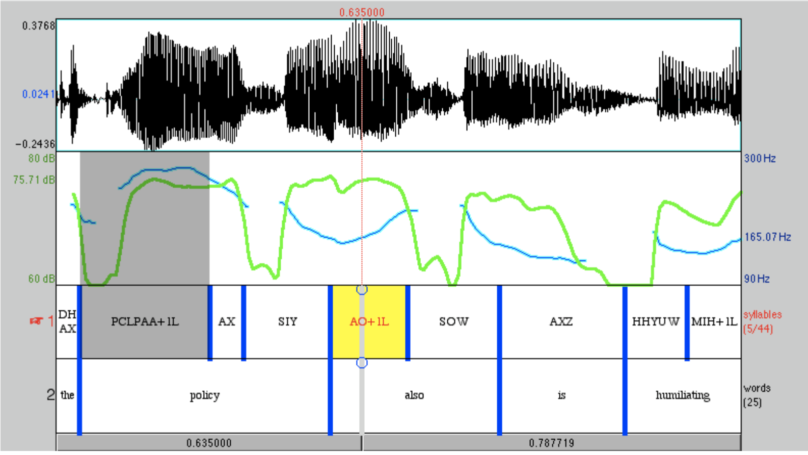


Figure 5.13: *Clear Example of L*+!H accent.*

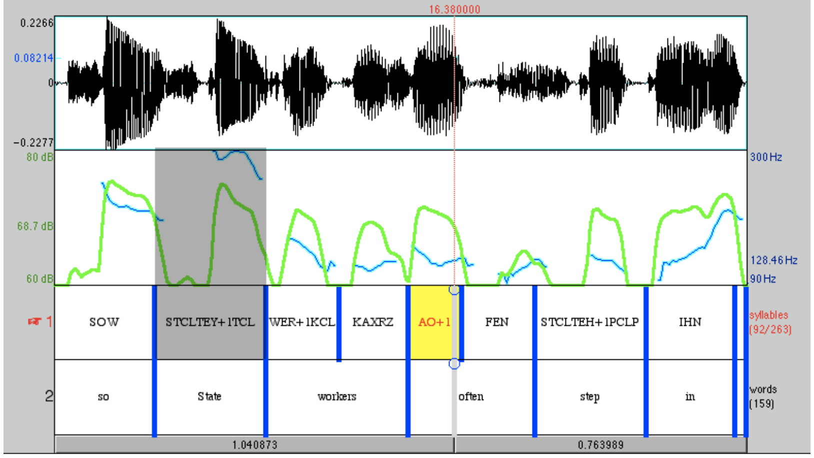


Figure 5.14: *Confusable Example of L*+!H accent.*

in Figures 5.15 and 5.16. The characteristic pitch fall is observable in the clear example. H* pitch accents have pitch peaks within the energy excursion, while this H+!H* token has a peak prior to the start of the lexically stressed syllable, and continues to fall into the following syllable. In the confusable example, the previous high tone is only slightly higher than the downstepped pitch in the accented syllable. Moreover, there is no downward slope of the pitch, the pitch is relatively flat throughout the duration of the accented syllable. This makes this token is easily confusable with an H* accent.

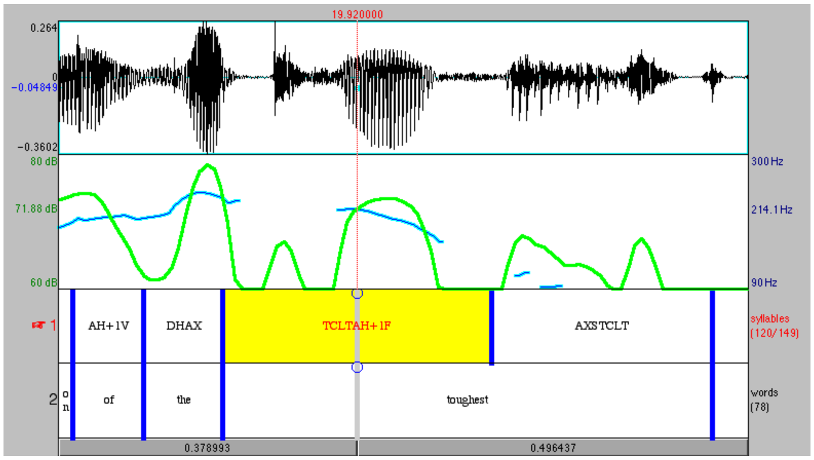


Figure 5.15: *Clear Example of H+!H* accent.*

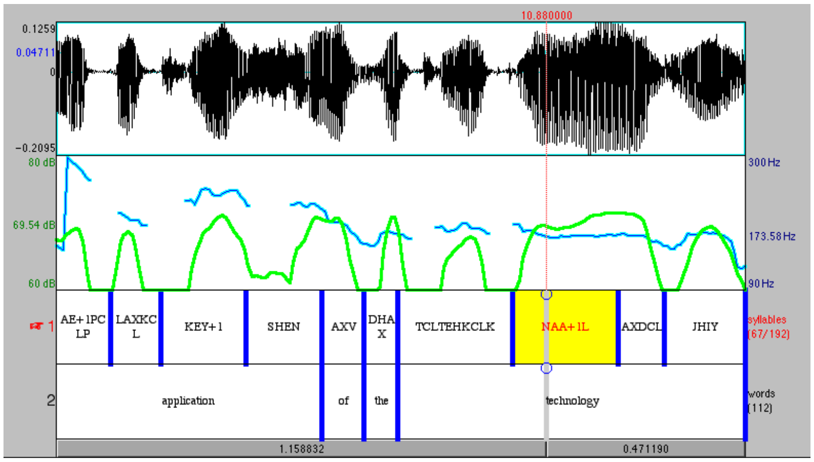


Figure 5.16: *Confusable Example of H+!H* accent.*

5.4 Materials

In the following sections, we describe pitch accent type classification experiments and analysis on BDC-read, BDC-spon and BURNC data. Chapter 2 contains thorough descriptions of these three corpora. The pitch accent type distributions within each of these corpora can be found in Table 5.1. The ToBI standard allows for the annotation of a pitch accent at X*? to indicate an uncertain tone. For the purposes of these experiments we omit these tokens from our analysis. On the BDC material, these uncertain accents are uncommon; 2% of accents in both the BDC-read and BDC-spontaneous corpora are labeled with X*?.

Corpus	H*	!H*	L+H*	L+!H*
BDC-read	48.2% (2158)	30.0% (1344)	12.4% (555)	1.3% (59)
BDC-spon	59.2% (3322)	25.4% (1425)	5.8% (325)	0.5% (30)
BURNC	54.0% (7713)	16.0% (2284)	17.0% (2433)	4.6% (658)

Corpus	L*	L*+H	L*+!H	H+!H*
BDC-read	6.0% (267)	1.3% (59)	0.0% (2)	0.7% (32)
BDC-spon	7.7% (432)	0.6% (34)	0.1% (4)	0.7% (41)
BURNC	3.7% (524)	0.3% (44)	0.0% (4)	4.4% (624)

Table 5.1: *Distribution of pitch accent types*

On the other hand, 12% of BURNC accents are annotated with this uncertain tone, X*?. In these experiments we collapse pitch accent types with their downstepped variants. The distributions of pitch accent types following this collapsing are presented in Table 5.2. The pitch accent type distribution of the BDC-read and BDC-spon corpora are largely similar. The read subcorpus has a higher proportion of L+!H* accents and a lower rate of H*. We find the BURNC data to have a significantly higher rate of L+H* and H+!H* and a lower rate of L* and H*. It has been noticed that newscaster speech, such as that comprising the BURNC data, is particularly idiosyncratic [24]. The difference in pitch accent distributions between these corpora may be further evidence of this.

Corpus	H*	L+H*	L*	L*+H	H+!H*
BDC-read	78.24% (3502)	13.72% (614)	5.97% (267)	1.36% (61)	0.71% (32)
BDC-spon	84.57% (4747)	6.32% (355)	7.70% (432)	0.68% (38)	0.73% (41)
BURNC	69.99% (9997)	21.64% (3091)	3.67% (524)	0.34% (48)	4.37% (624)

Table 5.2: *Distribution of pitch accent types with collapsed downstepped variants*

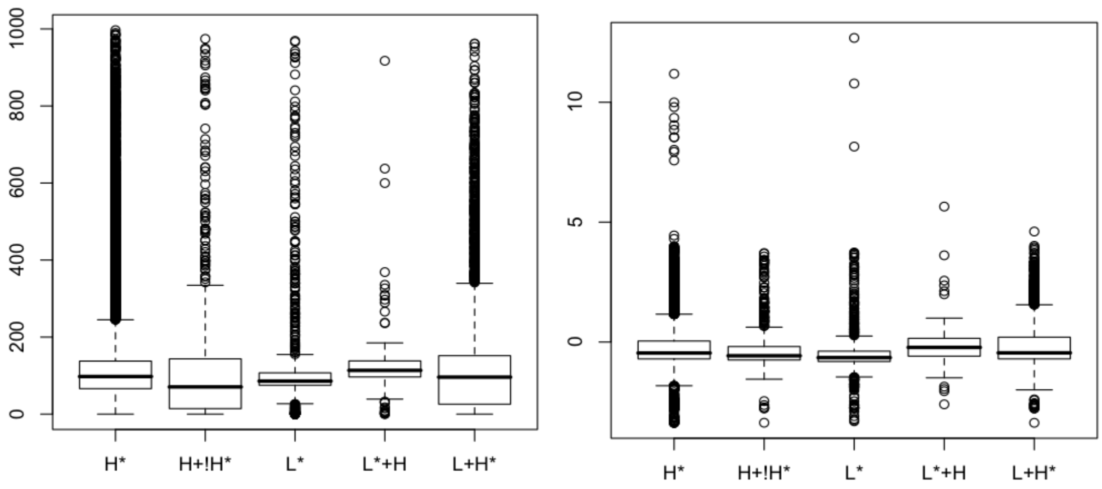
In the experiments presented in the following sections, we collapse downstepped and non-downstepped variants of the same pitch accent type. For example, H* and !H* are collapsed to a single class labeled as H*. Downstepped contours have the same canonical shape as their non-downstepped counterparts, but are realized in a compressed pitch range. While the shape and timing of pitch contours is a local phenomenon, occurring only for the duration of the prosodic event, measures of pitch ranges necessarily require an analysis of greater context. This quality makes the task of determining if a pitch accent it downstepped or not, quite different from determining which type of pitch accent it is. In Section 5.6.2, we present the results of a classification experiment distinguishing H* from !H* accents. We find with aggregations of acoustic information we can differentiate these two with over 80% accuracy.

5.5 Descriptive Analysis

In this section, here is to describe the acoustic characteristics that distinguish one pitch accent type from another. In this section, we analyze many of the acoustic features used in the classification experiments presented in Section 5.6. Unless otherwise stated, in these analyses we will collapse pitch accent types with their downstepped variants.

We examine the minimum, maximum, mean and standard deviation of pitch and energy extracted over the whole word, as well as the word's duration. In addition to extracting these over the raw pitch and energy contours, we speaker normalize these contours using z-score normalization. We also analyze these features extracted from the slope of the raw and normalized tracks. Along with these aggregations we extract tilt and skew coefficients of both the raw pitch and energy tracks. Lastly we examine the relative locations of the maxima of these contours, and immediately preceding and following slopes to these maxima. Based on the results in Section 5.6.3 we examine these features extracted from energy-peak defined regions as well as the whole word. In these analyses, we examine all accents in our three corpora, BDC-read, BDC-spontaneous and BURNC, without distinguishing between corpora.

We analyze the relationship between of pitch accent type on the acoustic features using ANOVA. All but one of the features described show a statistically significant ($p < 0.001$) interaction with pitch accent type when calculated over the full word.. Minimum pitch within the word does not demonstrate a significant influence of pitch accent type, yet the speaker normalized minimum pitch does, with $p < 2.2 * 10^{-16}$. Plots of raw and speaker normalized minimum pitch by pitch accent type can be seen in Figure 5.17.

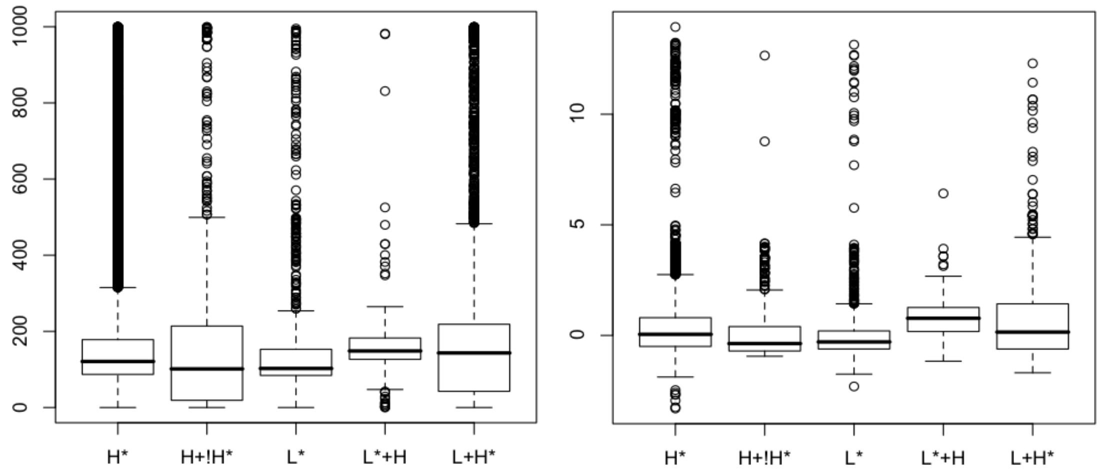


(a) minimum pitch (Hz)

(b) minimum normalized pitch (stdev)

Figure 5.17: *Minimum raw and speaker normalized pitch by pitch accent type.*

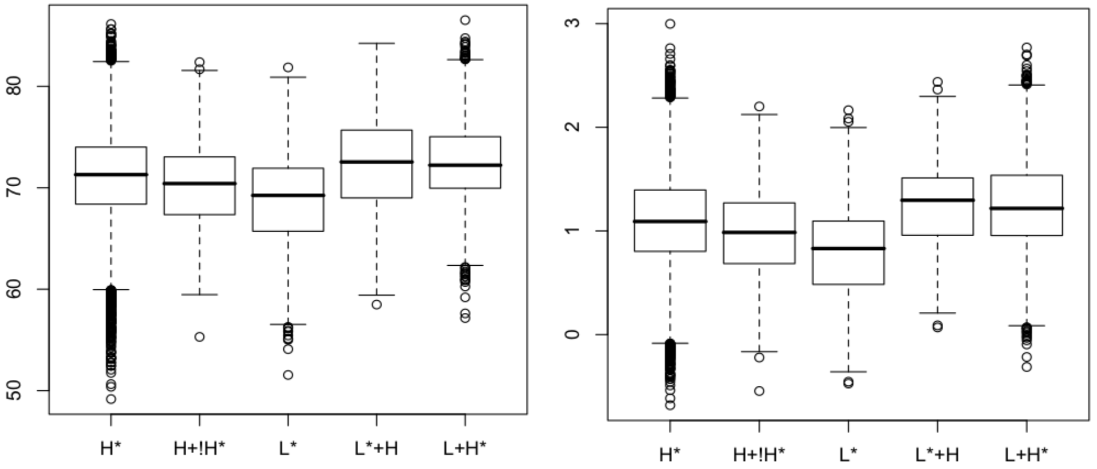
Examining these plots and associated means, we see that L* tokens have lower than average minimum pitch, while the complex tones, L+H* and L*+H have slightly above average values. The speaker normalization eliminates much of the noise generated by speaker differences. This leads to narrower distributions which are more likely to lead to significant differences across classes. We see similar relationships in maximum f0 and maximum intensity. Plots of these values can be observed in Figures 5.18 and 5.19.



(a) maximum pitch (Hz)

(b) maximum normalized pitch (stdev)

Figure 5.18: Maximum raw and speaker normalized pitch by pitch accent type.

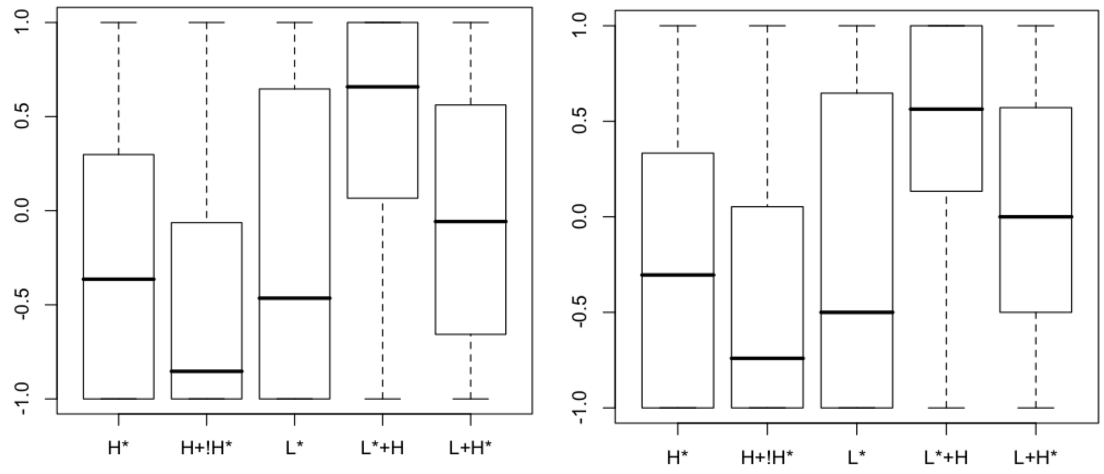


(a) maximum intensity (dB) (b) maximum normalized intensity (stdev)

Figure 5.19: Maximum raw and speaker normalized intensity by pitch accent type.

The speaker normalization of maximum pitch again serves to narrow the distributions, but in this case, as in the case of maximum intensity, we see the differences in means highlighted by the normalization. The maximum pitch and intensity of the low-to-high complex tones is higher than the overall mean, while that of L* accents is lower. The H+!H* tones lie somewhere between the H* and L* averages. The figures show that all of these examples contain a substantial number of outliers. However, the significance of these differences suggests that the failure of the classification experiments is more due to the skewed class distribution and dearth of minority class training instances.

Recall that Taylor's Tilt coefficients were developed to be a parametric representation of prosodic events such as pitch accents. The statistical significance of both $tilt_{amp}$ and $tilt_{dur}$ supports the claim that Tilt coefficients are efficient parameterizations of contour shapes in general, and of ToBI pitch accent types in particular. The plots of $tilt_{amp}$ and $tilt_{dur}$ can be seen in Figure 5.20. As expected, the H+!H* has negative tilt values, indicating and early peak with a greater decline than rise. Conversely L+H* and L*+H have later peaks and greater rises than falls, as indicated by their positive tilt values. With respect to Tilt values, L* and H* accents look very similar, limiting their usefulness to pitch accent classification.



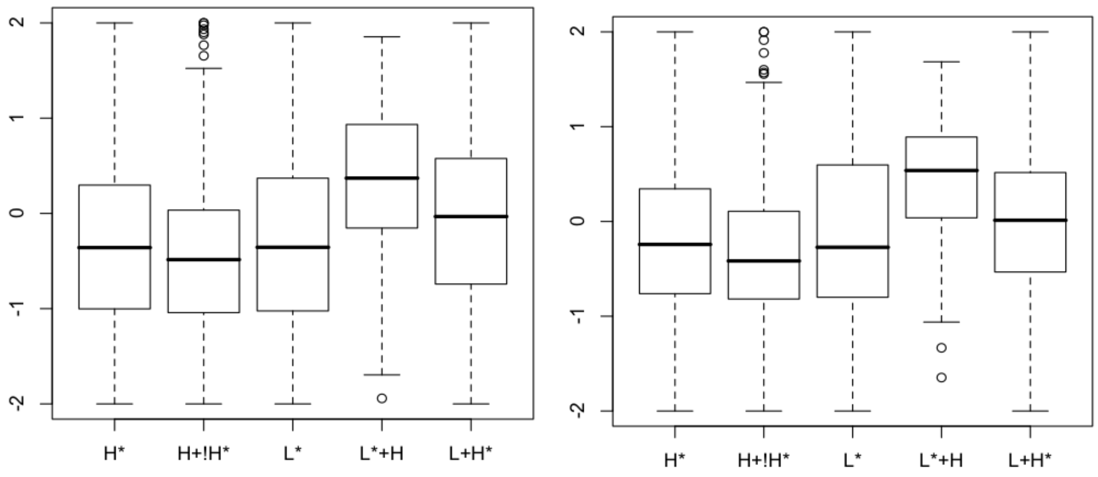
(a) $tilt_{amp}$

(b) $tilt_{dur}$

Figure 5.20: *Tilt coefficient values by pitch accent type.*

In Section 5.6.3, we proposed an extension to tilt coefficients by which we simultaneously parameterize the pitch and intensity contours and calculate the differences between their tilt values. These *skew* parameters also show significant influence of pitch accent type. Plots of these $skew_{amp}$ and $skew_{dur}$ are presented in Figure 5.21. Skew coefficients capture the degree to which the pitch and intensity contours over an accent are similarly shaped. These plots show that the H*, H+!H* and L* contours have similarly shaped contours, with the intensity contour skewing slightly earlier than the pitch contour, and with a slightly greater fall than rise. However, L*+H contours show a marked difference from this, with later and “more rising” pitch peaks. Our expectation was that L+H* accents have displaced pitch peaks, falling after the intensity peak. We find, on average, however, that these two are almost identically timed ($skew_{dur} = -.013$). However, we do find that, on average, L+H* accents have a *more* right displaced pitch peak than H* accents, which tend to have their pitch peaks *before* the associated intensity peak, with mean $skew_{dur} = -.19$.

In Section 5.6.3, we extract acoustic features from regions within a word defined by the location of maximum energy. These features lead to improved classification of minority class pitch accent types. Here we compare the influence of pitch accent type on these



(a) *$skew_{amp}$*

(b) *$skew_{dur}$*

Figure 5.21: Skew coefficient values by pitch accent type.

features, drawing contrast with the same acoustic property extracted from the whole word.

First, we find that every acoustic property demonstrates a statistically significant influence of pitch accent type with $p < .01$, including minimum raw pitch. Therefore, in this analysis, it is more informative to identify those features which are more discriminative when extracted from the energy-peak region than when extracted from the full duration of an accented word. A majority of features show the same or more significant influences of pitch accent type when extracted from the energy-peak region than from the full word. We therefore hypothesize that classifiers trained on features extracted from energy-peak regions will perform better than those extracted from full words. There are, however, some features for which this is not the case.

The slope of the pitch leading to the pitch maximum and the slope of the pitch leading away from this maximum show greater influence when extracted from the full word. Pitch values that lie outside the energy-peak region are omitted from the computation of pitch slope. This result suggests that these pitch points may be helpful in distinguishing pitch accent types by the slope of the pitch leading to and falling from the pitch maxima. Additionally, we find that the aggregations of the slope of the intensity contour are more influenced by pitch accent type when extracted from the full word. The mean slope over a word for example shows significantly narrower distributions when extracted from the full word than from the energy-peak region. The narrow distribution could merely be an artifact of the energy-peak region having fewer data points to aggregate than the full word. All else being equal, the presence of fewer data points would lead to a larger variance in the values extracted from energy-peak regions which in turn limits the observed influence of pitch accent type on these values.

When we identify the energy-peak region within a word, there are four special cases that may arise. Recall that the basic method for identifying this region is to locate the point of maximum energy within a word, then identify the local minima surrounding this extremal point. However, the local minima may fall in previous or following words. We refer to these cases as early valleys and late valleys. These are fairly common with 33% of accents having an early valley, and 19.9% having a late valley. Moreover, the maximum energy value may be the first or last value in the word, which may indicate that it is part of a peak that reaches a maximum in the previous – an early peak – or following word – a late peak. While we do not allow the energy-peak region to have a maximal value in the previous word, we do allow late peaks to be the peak of an energy-peak region. Early peaks are quite uncommon, occurring in only 0.44% of accents. Late peaks are still rare, though more common than early peaks, occurring in 1.43% of accents. This segmentation strategy approximately identifies syllable boundaries. These special cases occur only when the energy peak containing the point of maximum energy in the word span a word boundary. The rate of late peaks and late valleys only describe a phenomenon when the last syllable contains the point of maximum energy in the word. Similarly early peaks and early valleys only occur when the first syllable in the word contains the point of maximum energy. Therefore, the analysis of the rate of special cases contains an effect of the syllable structure of the accented word. We use a chi-square test to examine if these special cases occur equally in all accent types. We find that early valleys significantly differ with p=$3.478 * 10^{-17}$. L* accents have more early valleys than expected, while L+H* have fewer. This suggests that L+H* accents are more likely to start their energy excursion within an accented word. This is somewhat expected as the pitch movement is later timed than H* accents – the major contributor to the expected values. It is unclear why L* accents tend to begin earlier than H* accents, but this may contribute to the difficulty in detecting and classifying these accents. However, we also find that L* accents have more early peaks than expected. There is a significant interaction between early peak and pitch accent type with p=$6.18 * 10^{-8}$. These two results suggest that L* accents are identified by early energy peaks, relative to the accented word. Further investigation into the shape and location of L* accents is required to fully explain this result. Additionally, H+!H* have more early peaks than expected, and fewer late valleys. Late valleys also have a significant interaction with pitch accent type with p=$7.88 * 10^{-10}$. This again, is expected, given the shape of H+!H* accents. They are characterized by a sharp decline from a high pitch point. It is not surprising that this decline continues into a subsequent word. Moreover, the high pitch point which marks the start of the sharp decline need not occur within the accented syllable (or word). This early high pitch may have a concurrent energy excursion falling before the boundary of the accented word. We find no interaction between the presence of a late peak and pitch accent type.

In our classification experiments, we analyze the corpora – BDC-read, BDC-spon and BURNC – separately. In our discussion above, we find each acoustic property to demonstrate some influence of pitch accent type. Here we examine what if any influence of genre exists on pitch accent type. It is important to note from the onset of this analysis, that pitch accent types are a source of significant human disagreement. According to the Pitrelli et al. study there is only 76.1% human agreement with respect to pitch accent type [156]. This is *lower* than the majority class baseline on each BDC subcorpora, suggesting that if one of the labelers only used the H* accent, the agreement would *improve*. While this is, of course, relevant to the pitch accent type classification performance reported in Section 5.6 , it also has an impact on this analysis. While the BDC-read and BDC-spontaneous subcorpora were annotated by the same set of labelers, the BURNC data was labeled by a distinct set of experts. While a labeler may be internally consistent, the low interannotator agreement may lead to different accent type annotations across corpora. Before we make claims too broadly about the impact of genre on the associations between pitch accent type and acoustic features, it is necessary to acknowledge that these differences are between BDC and BURNC material likely due to labeler and speaker idiosyncrasies than a result of genre differences.

In the following analyses we examine the differences with respect to features extracted from the energy-peak region. These features demonstrate a greater influence of pitch accent type. The slope of raw or normalized pitch leading into and trailing away from the pitch maximum only demonstrates a statistically significant effect on the BDC subcorpora. In the BDC corpora, the H+!H* and L+H* corpora both show greater than expected pitch declines following pitch maxima, while on the BURNC data these are in line with the mean. This could be an influence of a more controlled speaking style associated with broadcast news.

The maximum slope of intensity is a particularly anomalous feature. On BURNC data, L*+H accents have a below average maximum slope. On BDC-spontaneous the influence is still significant, though L*+H accents have an *above* average maximum slope, while on BDC-read data this feature shows no significant effect of pitch accent type. There are very few L*+H tokens to draw conclusions from and the mean slope over an energy-peak does not have an obvious perceptual correlate. Positive values of this feature correspond to syllables that get louder over their duration. However, the perceptual salience of this seems rather minimal. It is particularly surprising to see significant and opposite effects of pitch accent type on this acoustic feature.

Also, the duration of the energy-peak region, while having an effect on BDC-spontaneous and BURNC shows no effect on BDC-read. The effect on BDC-spontaneous is observed by H+!H* accents having longer than average durations, while in the BURNC data L+H* accents have longer durations. The effect that is observed across all data reflects both of these, with both H+!H* and L+H* accents with above average duration. It is unsurprising for either of these pitch accent types to have prolonged durations. They are used in emphatic accenting situations to draw contrast or indicate finality. This emphasis may lead to longer than normal accents on these accent types.

The minimum slopes of pitch and intensity have no effect when analyzed on BURNC data but do when examining the BDC subcorpora. This is another acoustic feature with an unclear perceptual correlate. It indicates rapidly falling pitch or intensity. On the BDC data, the H+!H* accents as expected showed below average falling pitch – minima pitch slope. However, on the BURNC data these accents are somewhat in line with the mean, but L+H* accents have a below average minimum pitch slope – presumably corresponding to a rapid fall following the accent. The H+!H* accent is relatively uncommon in the examined material. This difference is likely due to labeler differences in the identification of this accent. It is also possible that H+!H* accents are produced with a less severe downward pitch slope by non-professional speakers, while L+H* accents have peaks that rise more sharply in professional speech.

We are encouraged that the extracted acoustic features do, in fact, show an influence of pitch accent type. The observed differences of genre do not clearly indicate differences in how speakers produce accents in different speaking styles. The read and spontaneous subcorpora of BDC showed largely the same relationships between acoustic features and pitch accent type. This is encouraging; it indicates that speakers and labelers are generally consistent in their production and annotation of pitch accent types regardless of genre. As we observe in Section 5.6, despite the statistically significant influence of pitch accent type on these acoustic features, the classification performance using these remains low. The statistical significance in an ANOVA test can be due to an observed effect of a particular pitch accent type on an acoustic feature. If this type is only one of the four minority classes, this discriminative difference may occur in only a fraction of the minority class instances. While indicative of a significant effect, the presence of difference in minority class tokens may explain why the significance we observe in these analyses does not directly translate to improved classification performance.

5.5.1 H* v. L+H* in BURNC

In Section 5.5, we examined which acoustic features show an influence of pitch accent type on their values. In this section, we narrow the analysis and examine the differences between the most common pitch accent type, H*, and the most common minority type, L+H*, commonly used to indicate contrast or emphasis. We examine only those accents in the BURNC data to eliminate any noise introduced by labeler idiosyncrasy and genre differences. On BURNC, and both BDC corpora, the combination of H* and L+H* accents represent over 90% of all pitch accents. Moreover, Syrdal and McGory find that confusion between H* and L+H* to be the most common source of interannotator disagreement regarding pitch accent type, accounting for one fourth of all pitch accent confusions [196]. Therefore, being able to distinguish between these two types is critically important to high accuracy pitch accent type classification. Also, indicators of emphasis or contrast are likely to be useful for downstream spoken language processing tasks. For example, utterances which contain L+H* accents likely include particularly salient information. Isolating these utterances may be able to improve extractive summarization of speech material.

With only two classes, we use a t-test to compare the means of each acoustic feature on H* and L+H* accents. The features are extracted over the energy-peak region (cf. Section 5.6.3) within each accent-bearing word.

Narrowing the analysis to only two pitch accent types, there are fewer acoustic features that show a significant influence of pitch accent type. The significant mean-differences we find, reveal a description of L+H* accents in line with the ToBI conventions. Hirschberg and Beckman describe the L+H* accent as “a high peak target on the accented syllable which is immediately preceded by relatively sharp rise from a valley in the lowest part of the speaker’s pitch range” [85]. In addition to the characteristic sharp pitch rise, L+H* accents are believed to have a right-displaced, or later in the time domain, pitch peak in comparison to H* accents.

We first find that the maximum, minimum and mean pitch in H* and L+H* accents do not significantly differ, while the standard deviation does. This suggests that while the range is approximately equivalent L+H* accents have greater pitch movement within this range. When we look at the slope of the pitch contour, we find that L+H* accents have a significantly greater maximum slope ($p = 3.074 * 10^{-4}$) – likely corresponding to the "sharp rise" – as well as a significantly greater mean slope over the energy-peak region ($p = 9.30 * 10^{-6}$). We observe positive mean slopes in L+H* accents, indicating a larger pitch at the end of the region than at the beginning, and negative mean slopes in H* accents. The negative slope in the H* accent is probably due to the natural declination of pitch [152], whose effect in the presence of the rise of the rise of the L+H* accent, is overwhelmed. We find the same set of features to show significant differences at the $p < .001$ level when analyzing raw and speaker normalized pitch contours.

When examining raw intensity aggregations, we find no significant differences in the loudness of H* and L+H* accents. However, when speaker normalized, we find that L+H* accents have greater maximum, mean and minimum intensity than H* accents all with $p < 1.0 * 10^{-6}$. This intensity effect supports the hypothesis that L+H* accents are more emphatic than H* accents, insofar as speakers tend to produce them more loudly. We also find that the maximum slope of intensity significantly differs across these two classes, with H* accents having a greater maximum slope. This effect is observed in both the raw and normalized contours. Since the energy-peak region consists of one intensity rise and one fall, this indicates that H* accents get louder faster than L+H* accents.

We also identify the location of the maximum of pitch and intensity within the region. We find that the intensity peak of L+H* accents falls 14ms later than that of H* accents ($p = 4.39 * 10^{-7}$), while the pitch peak falls 25ms later ($p < 2.23 * 10^{-16}$). This is evidence of the right-displacement of pitch peaks in L+H* accents, as defined by the ToBI standard. The right-diplaced intensity peak in combination with the observation about the slower intensity rise of L+H* describes a late slow increase in intensity. Moreover, we find that L+H* accents have energy-peak regions that are, on average, 20ms longer than H* energy-peak regions. This may be an artifact of the slow rise of intensity, or further support to the notion that L+H* accents are more emphatic than H* accents.

Tilt coefficients, when parameterizing both the raw and speaker normalized pitch contour, show statistically significant differences between these two pitch accent types. Both the $tilt_{amp}$ and $tilt_{dur}$ coefficients are significantly greater in L+H* accents. We find the average $tilt_{dur}$ value of H* accents to be -0.16 while L+H* has a mean value of 0.034, indicating that H* pitch peaks occur in the early half of the energy-peak region, while L+H* peaks fall just after the mid-point. As claimed by Taylor [202], we find very similar mean values of $tilt_{amp}$ for these two classes: -0.22 for H* and -0.025 for L+H*. H* accents have a greater pitch fall after their pitch peak, while the rise and fall of L+H* accents is nearly equal. When parameterizing the intensity contour using the tilt system we find that there is no significant difference between the $tilt_{amp}$ values – the amplitude rise and falls are roughly equivalent, and slightly negative, -0.017 for H* and -0.030 for L+H*. However, the $tilt_{dur}$ coefficients significantly differ with $p = 2.06 * 10^{-8}$ with H* having a mean value of -0.11, compared to -0.06 on L+H* accents. On both accent types, the energy peak falls before the mid-point of the energy-peak region, but is significantly earlier on H* accents. This is similar to the finding that the intensity peak location is significantly earlier under H* accents. Tilt coefficients were designed to be a concise parameterization of contour shape. The observed differences in Tilt values calculated from H* and L+H* accents indicate that this parameterization successfully captures some of the differences in accent productions.

We propose a combination of pitch and intensity Tilt coefficients called *skew* in Section 5.6.3. Both the $skew_{dur}$ and $skew_{amp}$ coefficients significantly differ in L+H* and H* accents with $p < 2.2 * 10^{-16}$. The mean $skew_{dur}$ value for H* accents is -0.05, while for L+H* it is 0.10. This indicates that in H* accents, the pitch peak falls slightly before the energy peak while in L+H* accents it falls somewhat after. The mean $skew_{amp}$ coefficient is -0.204 on

H* accents while 0.005 on L+H*. This indicates a more negative pitch tilt than energy in H* accents, while L+H* accents have approximately identically shaped pitch and energy contours – with a slight right displacement of the pitch contour. Therefore, *skew* coefficients represent a succinct parameter to describe relationship between pitch and intensity contours associated with the one of the key differences between H* and L+H* accents – the presence of a right-displace pitch peak.

In general, the analysis of acoustic features in H* and L+H* accents supports the differences between these two accent types prescribed by the ToBI convention annotation guidelines. L+H* accents have a sharper rise as indicated by the significance of the maximum pitch slope. They also have a right-displaced pitch peak compared to H* accents. However, this displacement is not very large. We find that the pitch peak of H* accents tends to fall *before* the energy peak, while expectedly on L+H* accents it falls after. We also find some evidence of L+H* accents being more *acoustically* emphatic than H* accents. They are produced with greater intensity – relative to the normal intensity of the speaker – and they are significantly longer in duration. We hesitate to hypothesize that these differences in acoustic qualities directly lead to differences in pragmatic effect. However, loudness and duration are particularly salient acoustic cues. While the ToBI conventions differentiate H* from L+H* accents by pitch contour shape and alignment, the observed increases in loudness and duration may contribute to the perception that L+H* accents indicate greater emphasis than H* accents.

5.6 Pitch Accent Type Classification Experiments

In this section, we describe a series of experiments which attempt to automatically classify pitch accent types using acoustic information. The data sets for training and evaluating these experiments include only accented words. This work can be seen as evaluating a *post hoc* process following pitch accent **detection**. This is a similar approach to that taken by

[5]. There is an assumption made in these experiments, that automatically detected pitch accents will have the same distribution as the manual annotations. This is an optimistic assumption. For example, the Corrected Energy Based Classifier (cf. Section 3.6) evaluated using ten-fold cross-validation on BDC-read correctly detects 97.45% of complex pitch accents (L+H*, L*+H, H+!H*), 81.2% of H* accents, but only 60.67% of L* accents. In the experiments in this section, however, we will assume the presence of an oracular pitch accent detection module, focussing our attention, not on accommodating errorful input from a pitch accent detector, but rather, on differentiating pitch accent types.

5.6.1 Basic acoustic aggregations

In this first set of experiments, our goal is to evaluate the performance of acoustic aggregations on the pitch accent classification task. For these experiments we use a feature vector containing aggregations of acoustic information extracted from each word. We extract pitch and energy contours from the speech tokens using Praat [20]. To account for speaker differences in these acoustic qualities, we normalize the pitch and energy contours for each speaker using z-score normalization. The slope contours of these normalized and unmodified contours are also constructed. Since the energy contour contains data points at every 10ms frame, the slope is equivalent to the delta of the contour. The pitch track, however, contains gaps at silent, unvocalized or noisy regions. Therefore, the calculation of slope captures the variable differences in time between consecutive pitch points. Delta calculations ignore the inconsistent distance in time between consecutive point points. From each these eight contours ({pitch, energy} × {raw, normalized} × {raw, slope}), we extract the minimum, maximum, mean, and standard deviation over each accented word. We also extract the z-score of the minimum and maximum value within the word, relative to its mean and standard deviation. In addition to these pitch and energy aggregations, we include in the feature vector the duration of the word in seconds, as well as the length of any previous or following pause.

Results of experiments using this feature set and a variety of classifiers on BDC-read, BDC-spon and BURNC appear in Table 5.3. These results treat pitch accent types and their downstepped variants as a single category. The Combined Error Rate (CER) is defined later in this section, and is included in these results for future comparisons.

Corpus	Baseline	J48	SVM
BDC-read	78.24	74.93 ± 1.279 / 0.371 ± 0.0267	78.19 ± 0.935 / 0.500 ± 0.0195
BDC-spon	84.57	82.24 ± 0.968 / 0.402 ± 0.0318	84.57 ± 1.033 / 0.500 ± 0.0205
BURNC	69.99	63.27 ± 0.689 / 0.466 ± 0.0161	70.00 ± 0.558 / 0.500 ± 0.0112

Table 5.3: *Accuracy (%) and Combined Error Rate of pitch accent type classification with collapsed downstepped variants.*

The accuracy rates indicate that these aggregations are not discriminative with respect to pitch accent type; none achieve accuracy higher than the majority class baseline. When the majority class baseline is so high due to a skewed class distribution, accuracy does not give a complete picture of the performance of a classifier. In this task, even the most modest baseline is ˜70%. However, not all accuracies of 70% are equal. A classifier that predicts all pitch accent types to be H* is behaving very differently from a classifier which is 70% correct in the classification of each class, with equally distributed errors. The impact of these distinct 70% accurate classifiers is very task dependent. However, a task that seeks to use pitch accent type information in a decision making process will not benefit from a module that only predicts H*, despite the fact that it is 70% accurate. This classifier, despite its accuracy, delivers no information to a downstream task.

To get a closer look at the performance of these classifiers, we examine the contingency matrices of the J48 decision tree classifier and the SVM with linear kernel classifier trained on BDC-read. Tables 5.4 and 5.5 contain the contingency tables of the J48 and SVM classifiers evaluated using 10-fold cross-validation on BDC-read material.

The SVM classification, we find, performs as well as the majority class baseline because it performs almost identically to the majority class rule. The J48 classifiers, while less accurate than the SVM classifiers, are able to correctly identify some minority class instances. This is not to suggest that high accuracy is unimportant, merely that it doesn't fully describe

	Actual Class				
Corpus	H*	L+H*	L*	L*+H	H+!H*
H*	3115	336	218	32	29
L+H*	252	257	5	24	0
L*	105	2	40	0	2
L*+H	25	18	1	5	0
H+!H*	5	1	3	0	1

Table 5.4: *J48 Pitch Accent Type Contingency Matrix. Evaluated on BDC-read material.*

	Actual Class				
Corpus	H*	L+H*	L*	L*+H	H+!H*
H*	3498	612	267	61	32
L+H*	4	2	0	0	0
L*	0	0	0	0	0
L*+H	0	0	0	0	0
H+!H*	0	0	0	0	0

Table 5.5: *SVM Pitch Accent Type Contingency Matrix. Evaluated on BDC-read material.*

the classifier performance. Read and Cox [165] implicitly noticed this issue as well. They reported their performance in terms of "balanced error rate" (BER), which equally weights a term related to the recall of each class (cf. Section 5.2 and Equation 5.1). This measure weights the importance of all classes equally.

We propose to use a different measure for the evaluation of pitch accent type classification. We desire a measure which is sensitive to minority class performance without artificially inflating its value. The Type I error rate measures the false positive rate for a given class. That is, for a given class, i, at what rate are token incorrectly classified as this class (cf. Equation 5.2).

$$\text{false positive rate of class } i = p(FP_i) = \frac{\text{number of false positives}}{\text{number of negative instances}} \tag{5.2}$$

When used to evaluate a classifier this term represents the likelihood of a Type I (false positive) error being generated by the classification process. Typically false positive rates are used in evaluations of binary classifiers where there is a clear notion of positive and negative examples of some phenomena. We extend this to this multi-class classification problem by calculating the expected Type I error rate over all classes, and weighting the contribution for each class by the likelihood that a token is a member of that class. The formula for this measure can be found in Equation 5.3, where $p(C_i)$ is the likelihood of a point being a member of class i and $p(FP_i)$ is the false positive rate for class i, that is treating class i as the "positive" class.

$$\text{Weighted Type-I Error Rate} = p(FP) = \sum_i p(C_i)p(FP_i) \tag{5.3}$$

Under this measure, every "miss" – incorrect classification – contributes to the error measure exactly once. Every incorrect classification is a false positive for one and only one class. However, the impact of each incorrect classification is weighted by the relative size of the predicted class. Incorrectly predicting the majority class is more significant than predicting a minority class. This measure, thus, severely penalizes over-predicting the majority class. However, it somewhat under-represents the cost of majority class false negatives. To address this, we couple this measure with the Weighted Type II Error Rate (cf. Equation 5.5 by taking the average of the two. The Type II error rate measures the false negative rate for a given class – the rate at which tokens of this class are incorrectly classified as another class (cf. Equation 5.4).

$$\text{false negative rate of class } i = p(FN_i) = \frac{\text{number of false negatives}}{\text{number of positive instances}} \tag{5.4}$$

$$\text{Weighted Type-II Error Rate} = p(FN) = \sum_i p(C_i)p(FN_i) \tag{5.5}$$

The combination of these measures results in an increased penalty for errors – false negatives and false positives – of the majority class while being more sensitive to minority class performance than accuracy. We believe this measure more faithfully measures the desirable behavior of pitch accent classification techniques than accuracy or BER. Throughout this chapter, we will continue to report accuracy for comparison to other work, but will prefer the Combined Error Rate (CER) (cf. Equation 5.6) for our discussion of classification success.

$$\text{Combined Error Rate} = CER = \frac{p(FP) + p(FN)}{2} \tag{5.6}$$

Note that both accuracy and Combined Error Rates can be described in terms of percentages. However, to avoid confusion, accuracy will be reported as a percentage (%) while Combined Error Rate will be reported as a decimal.

If we return to the contingency tables reported in Tables 5.4 and 5.5, we see that the SVM classification has a higher accuracy than the J48, 78.19% versus 76.36%. However, the J48 classification generates a lower Combined Error Rate [1] – 0.371 vs. 0.500. The Combined Error Rate is able to capture the improved classification of minority classes (and increased majority class precision), where accuracy fails to capture the useful aspects of the performance of the J48 classifier.

5.6.2 The value of context-normalized aggregations

We find consistently in the pitch accent detection experiments, inclusion of contextual information to improve detection performance (cf. Section 3.3 and [115]). We observe that normalizing acoustic features by surrounding context improves pitch accent detection. Moreover, we find that there is discriminative acoustic information in unstressed syllables within accented words that can improve detection of accents. In particular, this information was helpful in the detection of complex pitch accents. In this section, we examine if the context-based normalization that improved pitch accent detection is valuable for pitch accent type classification.

The context normalized features we include are z-score normalizations of the word mean and maximum of a contour within a word based on acoustic information drawn from each of

[1] Recall that lower error rates, and *CER* in particular, indicate better performance.

a set of context regions. The set of contours are the same as those analyzed in Section 5.6.1, namely, pitch and energy, and their slope, both raw and speaker normalized. We define six regions: 1) the previous word, 2) the following word, 3) the two surrounding words, 4) the two previous words, 5) the two following words, 6) the four surrounding words. In Table 5.6, we report the results of classification experiments using J48, and SVM with linear kernel classifiers using the set of pitch accent types with collapsed downstepped types.

Corpus	Baseline	J48	SVM
BDC-read	78.24 / 0.5	73.97 ± 1.214 / 0.371 ± 0.0271	80.90 ± 1.033 / 0.409 ± 0.0246
BDC-spon	84.57 / 0.5	82.88 ± 0.935 / 0.399 ± 0.00248	84.57 ± 0.869 / 0.500 ± 0.0172
BURNC	69.99 / 0.5	61.15 ± 0.558 / 0.459 ± 0.0110	70.00 ± 0.672 / 0.500 ± 0.0136

Table 5.6: *Accuracy (%) and CER of pitch accent type classification with collapsed downstepped variants using context-based features*

Decision tree classification, using the J48 algorithm, is consistently worse with the inclusion of context-normalized aggregations. However, SVM classification shows modest improvement on the BDC-read corpus. This data point keeps us from concluding that context-normalization is *completely* useless for pitch accent type classification. However, the inclusion of this material does not, in most cases, significantly improve performance. The broader acoustic context is not helpful in the classification of pitch accent types.

Classifying downstepped accents

In Section 5.4, we make the claim that distinguishing downstepped accents from canonical variants is a rather distinct task from classifying broad pitch accent types. The latter is a classification of contour shape, while identifying downstepped accents can be accomplished by recognizing whether or not the accent is produced in a compressed pitch range. In this section we briefly evaluate this claim.

The context-normalized features that we examined in Section 5.6.2 capture the acoustic context surrounding the accented word. From this context we should be able to determine the approximate pitch range in which an accent was produced. In addition to accents produced in a compressed pitch range having lower raw pitch values, they are likely to be lower relative to the preceding acoustic material when compared to non-downstepped variants.

Using the same feature set described in Section 5.4 we run an experiment classifying H* accents from !H* variants. We do not repeat this experiment distinguishing L+H* from L+!H* or L*+H from L*+!H* due to the dearth of available training data for these classes. The accuracy and F-measure of detecting !H* accents from ten-fold cross-validation experiments using J48 and SVM classification are reported in Table 5.7. The majority class baseline on all corpora is H*.

Corpus	Baseline	J48	SVM
BDC-read	61.62 (0.0)	79.75 ± 0.820 (0.739)	83.27 ± 0.836 (0.782)
BDC-spon	69.98 (0.0)	78.55 ± 0.820 (0.636)	81.21 ± 0.886 (0.666)
BURNC	77.15 (0.0)	74.11 ± 0.787 (0.336)	77.09 ± 0.787 (0.000)

Table 5.7: *Accuracy (%) and !H* F-measure of classifying H* from !H* accents.*

We find a major corpus effect in these experiments. On the BDC data, the acoustic aggregations and context features are quite capable of distinguishing H* from !H* accents. However, the performance is significantly worse on the BURNC data. This could be a genre effect, professional speakers delivering broadcast news may produce more ambiguous downstepped contours than lay speakers participating in a direction-giving task. However, the ToBI convention allows some degree of labeler discretion in determining how compressed a speaker's pitch range must be in order for an accent to be considered downstepped. It is possible that BDC labelers apply the downstepped annotation more conservatively, making automatic disambiguation of H* and !H* easier. If we compare the maximum pitch values within words that are accented with H* and !H*, we can see some evidence of this. On the BURNC data, these do not differ significantly, and moreover, the mean maximum pitch in downstepped tokens is 221.8Hz, compared to 220.94Hz in H* accents. If we compare the speaker normalized values, the !H* tokens have a mean maximum normalized pitch of .239 standard deviations compared to .245 for H* accents. This difference is also not statistically significant as evaluated by Student's t-test. In contrast, the normalized maximum pitch values from the BDC-read corpus are .724 for H*, .411 for !H*, on the BDC-spon data. These values are .628 for H* and .471 for !H*. Both of these differences are statistically significant with $p < .005$.

The accuracy in excess of 81% and an f-measure greater than 0.65 found on the BDC-read material is quite successful, though there is obviously some room for improvement. A unique characteristic of the !H* accent is that it must occur in an intermediate phrase that previously contains a non-downstepped high (H) tone, whether in H*, L+H*, L*+H or H+!H* accent. Therefore, a !H* accent can never be the first accent in an intermediate phrase. If a task requires information about downstepped accents, incorporation of this constraint should be able to improve on classification performance.

5.6.3 Shape modeling

The results in Sections 5.6.1 and 5.6.2 show the limitations of acoustic aggregations for pitch accent type classification. These features are quite successful in distinguishing accented from unaccented words. However, the pitch and energy contour shapes that correspond to accent types are not exclusively differentiated by excursions in either the pitch or energy domain, rather pitch accent types are distinguished by their pitch contour shape. In this section, we examine features that attempt to capture narrower phenomena within each accented word by modeling the contour shape.

Extrema-based shape features

The first set of shape features we examine are based on the maxima of pitch and energy within each accented word. We identify the maximum within the word and include in the feature vector, the position in seconds from the start of the word, as well as the position relative to the word duration. We also calculate the slope from the preceding local minimum to the maximum, and the slope from the maximum to the following local minimum. These are calculated for both the pitch and energy contours. We also include the distance between the pitch and energy maxima. The canonical examples of the ToBI accent types, L+H* has a delayed pitch peak, where the pitch maximum follows the energy maximum. This phenomenon is still more pronounced in L*+H accents, where the pitch peak may occur as late as the following syllable.

We perform a set of classification experiments using only these shape features. The results of these experiments can be found in Table 5.8. We evaluate the performance of J48 decision trees and SVM with linear kernels using ten-fold cross-validation. We find that these features do not yield a dramatic reduction in Combined Error Rate from the majority class baseline. However, on the BDC corpora they demonstrate a statistically significant reduction in the *CER* from the baseline when evaluated using a decision tree classifier.

Corpus	Baseline	J48	SVM
BDC-read	78.24 / 0.5	73.24 ± 1.328 / 0.445 ± 0.0441	78.24 ± 1.296 / 0.5 ± 0.0257
BDC-spon	84.57 / 0.5	81.74 ± 0.836 / 0.443 ± 0.0276	84.57 ± 1.000 / 0.5 ± 0.0202
BURNC	69.99 / 0.5	65.74 ± 0.361 / 0.483 ± 0.00886	70.00 ± 0.689 / 0.5 ± 0.0136

Table 5.8: *Accuracy (%) and CER of pitch accent type classification with collapsed downstepped variants using extrema based features*

Tilt coefficients

Tilt coefficients describe the shape of a pitch contour within a prosodic event by four terms, which are combined into a single *tilt* term [201, 202].. The pitch contour is divided into two regions, the rise – the region preceding the maximum pitch – and the fall – the region following. One of these regions may be empty, if the first or last pitch point is the maximum in the event. Two Tilt coefficients are determined by the location of the pitch maximum in the time domain. These duration terms are dur_{rise} and dur_{fall}, and are equal to the duration from the start of the event to the time of the pitch maximum and from the maximum to the end of the event, respectively. The remaining two coefficients describe the amplitude of the pitch maximum. The rise of the pitch contour leading to the maximum is calculated as the difference between the amplitude of the maximum and the minimum pitch value within the event, prior to the maximum. This term is called amp_{rise}. Correspondingly, amp_{fall} describes the pitch movement following the maximum, which is the difference between the maximum value to the minimum pitch value within the event following the maximum location. A graphical representation of these terms can be seen in Figure 5.22.

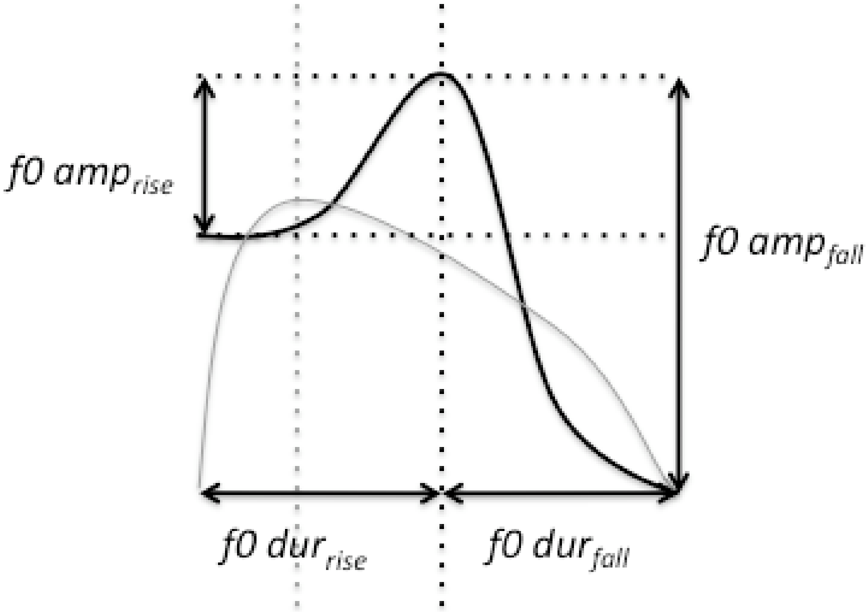


Figure 5.22: *An example of Tilt coefficients based on pitch contour.*

These four terms are combined to describe the shape of the event. The "tilt" of the amplitude is computed as follows.

$$tilt_{amp} = \frac{|amp_{rise}| - |amp_{fall}|}{|amp_{rise}| + |amp_{fall}|} \tag{5.7}$$

If $tilt_{amp} = 0$, the event has an equal rise and fall on either side of the pitch maximum. A positive value indicates a larger rise than fall, while a negative value describes a larger fall than rise. As an example, the contour in Figure 5.22 has a negative $tilt_{amp}$ value.

The duration tilt is defined as

$$tilt_{dur} = \frac{dur_{rise} - dur_{fall}}{dur_{rise} + dur_{fall}} \tag{5.8}$$

If $tilt_{dur} = 0$, the pitch maximum is exactly in the center of the event. A positive $tilt_{dur}$ term indicates a maximum in the latter half of the event, while a negative value indicates an early maximum, i.e. in the first half of the event. The contour in Figure 5.22 has a $tilt_{dur}$ value close to zero.

According to [201], "empirical evidence has shown that [$tilt_{amp}$ and $tilt_{dur}$] are highly correlated." This allows the two to be combined, with their average used to describe the overall *tilt* of an event.

$$tilt = \frac{tilt_{dur} + tilt_{amp}}{2} \tag{5.9}$$

We hypothesize, however, that pitch accent types are differentiated not only by their pitch contour, but also the relationship between the pitch contour shape and the concurrent energy contour. Both the timing of pitch peaks relative to vowel onset, and their alignment with the word's stressable syllable have been noted as a distinguishing characteristic of pitch accent types. Vowel onset times can be detected in the energy contour as vocalic regions typically have greater energy than non-vocalic regions. The Tilt coefficients, while designed to describe pitch shapes, are equally applicable to any sequence of time-value pairs, like an energy contour. Figure 5.23 contains an example of the calculation of Tilt coefficients from the energy contour concurrent with the pitch contour displayed in Figure 5.22.

In addition to the independent pitch and energy Tilt coefficients, we examine a combination of these two, to represent the difference in their shapes. Since the Tilt coefficients are unitless ratios, the energy- and pitch-derived terms can be combined without normalization terms. We call these differences *skew* coefficients.

The *skew* in duration tilt is represented by

$$skew_{dur} = tilt^{f0}_{dur} - tilt^{I}_{dur} \tag{5.10}$$

If this value is zero, the pitch and energy maxima are concurrent. If it is positive then the f0 maximum occurs after the energy maximum, if negative, vice versa.

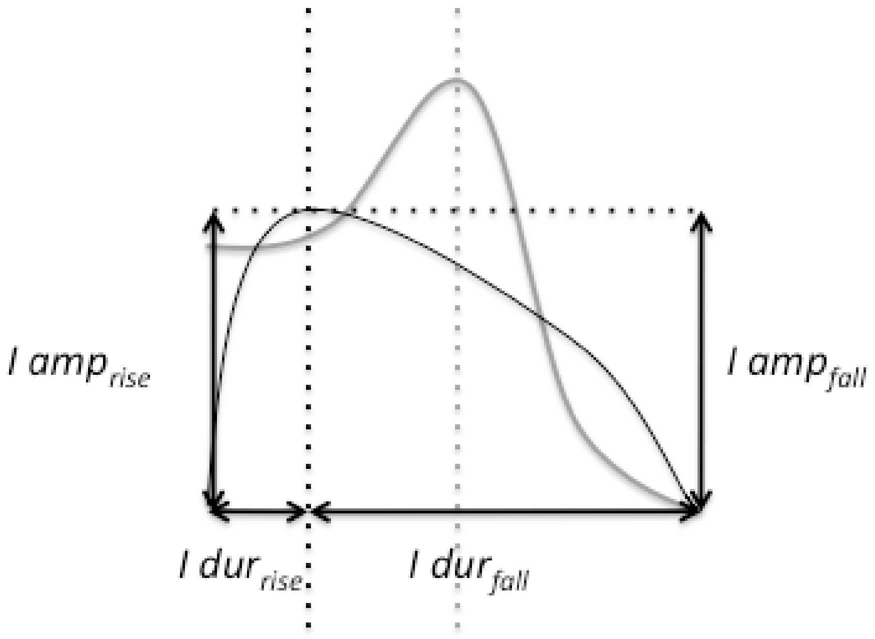


Figure 5.23: *An example of Tilt coefficients based on a concurrent energy contour.*

The *skew* in amplitude is, similarly, defined as

$$skew_{amp} = tilt^{f0}_{amp} - tilt^{I}_{amp} \tag{5.11}$$

If $skew_{dur}$ is zero, the pitch and energy contours have a nearly identical shape. Contours with positive $skew_{amp}$ values have a pitch contour that "tilts" more positively, that is, has a greater rise than fall, than the concurrent energy contour. Negative $skew_{amp}$ described a pitch contour that has greater fall than the corresponding energy contour.

The results of ten-fold cross-validation experiments using only tilt and skew coefficients are reported in Table 5.9. We find that these parameterizations of the pitch and intensity

Corpus	Baseline	J48	SVM
BDC-read	78.24 / 0.5	78.24 ± 0.771 / 0.500 ± 0.0156	78.24 ± 0.886 / 0.500 ± 0.0177
BDC-spon	84.57 / 0.5	84.50 ± 0.771 / 0.499 ± 0.0147	84.57 ± 0.738 / 0.500 ± 0.0148
BURNC	69.99 / 0.5	66.74 ± 0.820 / 0.490 ± 0.0123	70.00 ± 0.640 / 0.500 ± 0.0128

Table 5.9: *Accuracy (%) and CER of pitch accent type classification with collapsed downstepped variants using tilt coefficient features*

contours perform worse than the extrema-based features explored in Section 5.6.3 on all corpora. On the two BDC corpora they are not able to improve over the baseline classification performance. We do observe a 0.01 improvement in *CER* when using J48 decision trees to classify pitch accent using these tilt-derived features. This is not a dramatic improvement – only a 2% reduction in error, which is not statistically significant (p=0.379). While we found some interesting correlations between Tilt coefficients and pitch accent types, this parameterization is not able to successfully classify pitch accents in isolation. In the experiments described in this section we extract Tilt coefficients over the full duration of the accent bearing word. This parameterization is not designed to operate on this region, but rather, to capture only the pitch contour of the "accent event", approximately the accented syllable. In Section 5.6.3, we revisit the use of Tilt coefficients, applying this parameterization technique to narrower regions of analysis.

Quantized Contour Model

The shape modeling features described in Sections 5.6.3 and 5.6.3 fail to significantly capture the differentiating qualities of the inventory of pitch accent types. The features explored in these sections are parameterizations of the pitch and duration contour shape. The formulation of these parameterizations are all based on the assumption that the shape of a pitch accent is based on the relative location of extrema in pitch and intensity. While this is consistent with the annotation definitions of pitch accent types, it seems likely that these features are not capturing the same qualities that the human listeners identified when annotating the data. In this Section we present a modeling technique, Quantized Contour Modeling, which attempts to model the contour of a word without explicit parameterization of pitch and energy extrema.

In this technique, we quantize the f0 contour of a word in the time and pitch domains, generating a low-dimensional representation of the contour. The pitch of the contour is linearly normalized to the range between the minimum and maximum pitch in the contour, and quantized into N equally sized bins. The time domain is normalized to the range [0,1] and quantized into M equally sized bins. An example of such a quantization is presented in Figure 5.24 where $N = 3$ and $M = 4$.

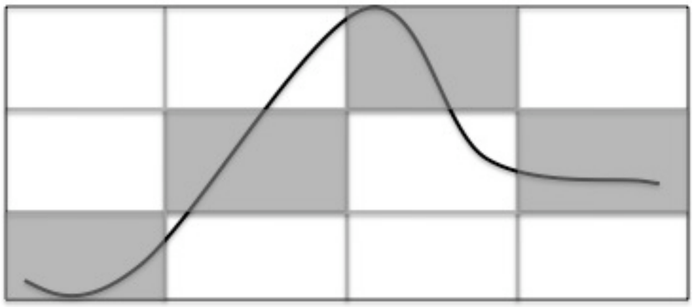


Figure 5.24: *An contour quantization with $N = 3$ time bins and $M = 4$ value bins.*

Using this quantized representation of a pitch contour, we train a multinomial mixture model for each pitch accent type. Let the quantized contour be an M dimensional vector C where $\forall C_i \in C, C_i \in \{0 \ldots N-1\}$. We indicate pitch (f0) contours by C^{f0} and intensity contours by C^I. We train a multinomial model $p(type|C_i, i)$ for each time bin $i \in \{0 \ldots N-1\}$ with Laplace (add-one) smoothing. We use these pitch accent type models to classify a contour using the following Bayesian classification function. The formula for this classifier can be found in Equation 5.12. Using this framework, we experiment with some extensions to this model. We model the slope over the course of the contour using Equation 5.13. Also, we can modify the model to include a sequential component by explicitly modeling the current and previous quantized values, as in Equation 5.14.

Moreover, we can extend these delta and sequential models to model the energy contour shape as well as the pitch contour. We constrain the time quantization of the two contours to be equal. The classification technique allows for the number of pitch and energy value quantization bins to be distinct. However, in these experiments, we keep these equal. The form of the classification functions using the energy contours are found in Figure 5.25.

We evaluate this quantized shape modeling using ten-fold cross validation, with a range of quantization sizes, $N \in \{2 \ldots 7\}$ and $M \in \{2 \ldots 7\}$. This produces thirty six quantization

Standard shape modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} p(C_i|type, i) \tag{5.12}$$

Delta f0 modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} \begin{cases} p(C_i - C_{i-1}|type, i) & \text{if } i > 0; \\ p(C_i|type, i) & \text{if } i = 0;. \end{cases} \tag{5.13}$$

Sequential f0 modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} \begin{cases} p(C_i|C_{i-1}, type, i) & \text{if } i > 0; \\ p(C_i|type, i) & \text{if } i = 0;. \end{cases} \tag{5.14}$$

Standard f0 + I modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} p(C_i^{f0}, C_i^{I}|type, i) \tag{5.15}$$

Delta f0 + I modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} \begin{cases} p(C_i^{f0} - C_{i-1}^{f0}, C_i^{I} - C_{i-1}^{I}|type, i) & \text{if } i > 0; \\ p(C_i^{f0}, C_i^{I}|type, i) & \text{if } i = 0;. \end{cases} \tag{5.16}$$

Sequential f0 + I modeling

$$\underset{type}{\operatorname{argmax}}\, p(type) \prod_{i}^{M} \begin{cases} p(C_i^{f0}, C_i^{I}|C_{i-1}^{f0}, C_{i-1}^{I}, type, i) & \text{if } i > 0; \\ p(C_i^{f0}, C_i^{I}|type, i) & \text{if } i = 0;. \end{cases} \tag{5.17}$$

Figure 5.25: *Quantized contour shape modeling classification formulae.*

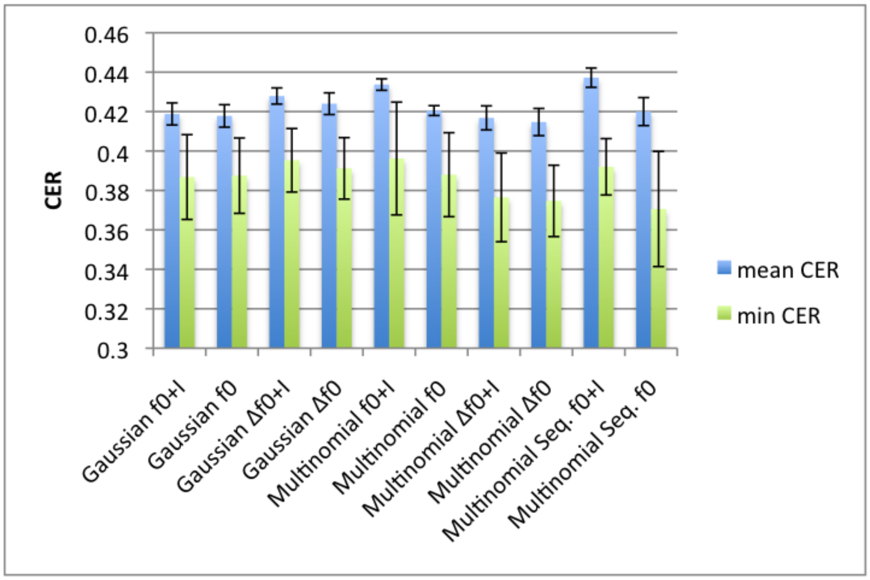


Figure 5.26: *Mean and Minimum CER of Quantized Shape Modeling evaluated on BDC-read. Standard, delta and sequential modeling are all evaluated on pitch only, and pitch and intensity.*

configurations, and six modeling techniques. To facilitate the visualization of these results on the three evaluation corpora, we plot the average and maximum Combined Error Rate from BDC-read in Figure 5.26.

We find that the sequential multinomial f0 modeling perform the best with a minimum *CER* of 0.371. The inclusion of an intensity dimension to the standard f0 modeling never significantly improves the mean or minimum *CER*. However, this performance is not significantly better than the best performing standard or delta multinomial modeling, each of which yield a minimum *CER* of 0.375. Simultaneously modeling the pitch and intensity contours performs worse than the modeling of pitch contours alone. This suggests that, at least in this modeling context, intensity is a source of noise, not discriminative information. In the case of sequential multinomial modeling the inclusion of the intensity dimension significantly *increases* the mean *CER* by 0.0172 (p=0.00968).

In these experiments we model the contour of the full word. The contour shape that differentiates on pitch accent type from another is localized to the area around the acoustic excursion that indicates that the word is accent-bearing. The failure of this modeling technique is likely due to the fact that we are attempt to model complex data with a simple tool. The presence of additional acoustic material and inconsistent accent placement makes even the 81-bit (7-by-7 quantization of pitch and energy) representation too impoverished to capture the differences in accent type, given the amount of training data available. In Section 5.6.3, we examine techniques to isolate the a region *w*ithin a word that corresponds to the pitch accent while ignoring potentially irrelevant acoustic material.

Varying the region of analysis

The acoustic excursions that characterize pitch accents, while typically localized on the lexically stressed syllable, are not strictly constrained to the segmental boundaries that define the syllable, or even the accent-bearing word. We see some evidence of this in the error analysis in comparing syllable- and word-based pitch accent detection experiments (cf. Section 3.4). In this work, we find that complex pitch accents can be detected better by classifiers which have access to acoustic information from unstressed syllables within an accented word.

In this section, we explore a different way of identifying the region of analysis for pitch accent type classification. In previous experiments we have been analyzing the full accented word. The results of these experiments have not, however, been able to consistently significantly outperform the majority class baseline. Instead of using the full word to calculate acoustic features, we use a crude syllabification technique to identify the accented region within the word. This technique is based on the assumption that the point within the word with the greatest intensity will lie within the accented syllable. We identify the point of maximum energy within the word. If it falls at the start of the word, we ignore this, and find the next greatest local maxima. This energy peak corresponds to an accented word preceding the current word, with this initial maximum being an artifact of the previous energy excursion. If the maximum within the word falls at the word-final boundary, however, we extend the region of analysis into the following word, or words, until the local maximum

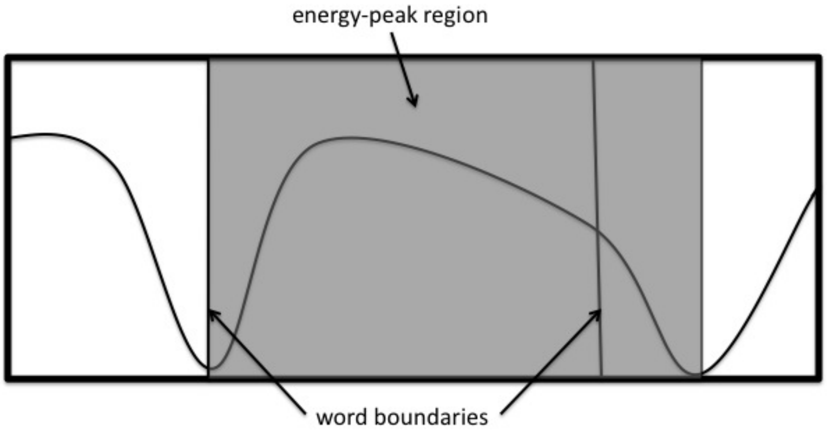


Figure 5.27: *An example of using the energy maximum to identify the region of analysis for pitch accent type modeling.*

is found. We identify the “region of analysis” as the span between the local minima on either side of this local maximum. These local minima, may or may not fall within the word that is currently being analyzed. Unconstrained by segmental boundaries, this region of analysis should have access to the full acoustic excursion corresponding to the pitch accent. This segmentation strategy attempts to omit unstressed syllables from analysis, thus hopefully, omitting irrelevant acoustic material from outside the pitch accent.

This segmentation technique, however, may introduce error if it incorrectly identifies the region of a word bearing the pitch accent. The energy maximum in a word tends to correspond to the stressable syllable aligned with the starred portion of the accent, but this is not necessarily so the case. Moreover, using the local minima to end the region may omit discriminative acoustic information from the analysis. The low preceding a L+H* may occur in a previous syllable, and the high following L*+H may fall well after the energy peak. An example of the region of analysis identification can be seen in Figure 5.27.

In addition to the energy-peak region of analysis, we also experiment with the pseudo-syllabification technique described in [219]. This technique generates hypothesized syllable boundaries based on the envelope, spectral information and zero-crossing rate of the speech signal independent of hypothesized word boundaries. As these pseudo-syllable boundaries do not align with the manual word boundaries, we select the pseudo-syllable which contains the maximum energy in the word as the representative syllable for modeling pitch accent type.

We evaluate the performance of these two regions of analysis using Quantized Contour Modeling (cf. Section 5.6.3) modeling the energy peak and pseudo-syllable regions, as opposed to the contour of each full word. The delta multinomial modeling configuration of Quantized Contour Modeling achieved the best performance in classifying the full word pitch contour. Therefore, we evaluate the performance of this configuration in classifying accent types using contours extracted from the three regions of analysis. The results of these evaluations are reported in Table 5.10. We find the energy-peak region pitch contour does not classify pitch accent types as well as the full word contour. The Quantized Contour Modeling approach classifies pitch accent with nearly equivalent accuracy whether modeling pitch contours extracted from the full word or from the pseudo-syllable region. On one hand, this shows that isolation of the pseudo-syllable with the greatest energy fails to eliminate any non-discriminative noise from the parts of a word that are outside the realization of a pitch accent. On the other hand, this result also indicates that the pseudo-syllable region contains all of the discriminative information contained in the word; the inclusion of material outside the pseudo-syllable does not improve performance. In Section 5.6.3, we evaluate the use of the pseudo-syllable region in the extraction of acoustic aggregations described in Section 5.6.1 and shape features described in Section 5.6.3 and 5.6.3.

Region	*CER*	Time bins	Value bins
Full Word	0.371 ± 0.0292	7	6
Energy Peak	0.440 ± 0.0196	5	7
pseudo-syllable	0.367 ± 0.0179	6	7

Table 5.10: *CER and quantization parameters of the best performing quantized contour model classifiers using sequential multinomial modeling and evaluated on BDC-read.*

Combination of Shape Modeling Features

In Section 5.6.3, we demonstrated that, using a pseudo-syllabification technique [219], we can isolate a potentially useful region of analysis for shape modeling. This region of analysis appear to capture the accented region within a word, without omitting an observable amount of discriminative information. In this section, we extract extrema-based shape features (cf. Section 5.6.3), Tilt coefficients (cf. Section 5.6.3) and aggregations (cf. Section 5.6.1) from this region instead of the whole word.

We will first evaluate the classification accuracy and CER of these feature sets in isolation. Then we will combine these features with the quantized shape modeling results from Section 5.6.3. For this combination, we extend the feature vector containing the other shape features with posteriors from multiple quantized shape modeling classifiers. We use the standard, delta and sequential multinomial f0 and two-dimensional f0 and intensity classifier posteriors with value bins, $N \in \{2 \ldots 7\}$, and time bins, $M \in \{2 \ldots 7\}$. Following this evaluation, we extend the feature vector still further, with the context normalized acoustic aggregations defined in Section 5.6.2. We evaluate each of these classification scenarios using J48 and SVM with linear kernels using ten-fold cross validation. The results on BDC-read can be found in Table 5.11.

Feature Set	J48	SVM
Extrema (**E**)	72.72 ± 1.492 / 0.420 ± 0.0330	78.24 ± 1.509/ 0.400 ± 0.0303
Tilt (**T**)	78.24 ± 1.378 / 0.500 ± 0.0277	78.24 ± 1.017 / 0.500 ± 0.0203
Model Output (**M**)	71.65 ± 2.150 / 0.469 ± 0.0288	73.10 ± 0.910 / 0.488 ± 0.0151
Aggregations	74.55 ± 0.689 / ***0.362*** ±0.0325	80.85 ± 1.033 / 0.403 ± 0.0253
E+T+M	71.40 ± 1.033 / 0.425 ± 0.0246	72.36 ± 1.310 / 0.443 ± 0.0271
E+**T**+Aggregations	75.42 ± 1.000 / 0.370 ± 0.0303	78.95 ± 0.951 / 0.475 ± 0.0300
E+**T**+**M**+Aggregations	72.30 ± 1.547 / 0.461 ± 0.0236	75.13 ± 0.751 / 0.465 ± 0.0152

Table 5.11: *Accuracy and Combined Error Rate of combination of shape modeling and aggregation features extracted over pseudo-syllable regions. These results are based on ten-fold cross-validation on the BDC-read corpus.*

We find the best performing classifiers to be the J48 decision trees trained using only aggregation features with a CER of 0.371 with aggregations calculated over the whole word and 0.362 when extracting features from the pseudo-syllable region. While this difference is not significant (p=0.735), the trend suggests that the pseudo-syllable region may be a useful region of analysis for pitch accent type classification.

The inclusion of shape features or Quantized Contour Model posteriors does not significantly improve classification performance. However, we find that the best performing Quantized Contour Model can classify pitch accent types with $CER = 0.371$ (cf. Table 5.10). This is a significantly lower Combined Error Rate than the J48 and SVM classifiers demonstrate when trained using only model posteriors. This suggests that there is discriminative information in the Quantized Contour Model posteriors that is not being efficiently employed by these classification techniques.

Moreover, the information captured by the Quantized Contour Model is quite distinct from the acoustic aggregation features. Specifically, the shape modeling technique normalizes f0 and intensity values before quantization, leaving it incapable of distinguishing contours in different ranges. For example, L+H* contours tend to reach a higher peak than H* accents; Quantized Contour Modeling does not have access to this information. On the other hand, the aggregation features reduce all the shape information from the contour, representing it as mean, standard deviation and extrema values. This suggests that if we combine the two classifiers we may be able to reduce the overall error rate. We thus explore a second combination strategy. In this approach, we combine an aggregation-based classifier (J48 or SVM) with a Quantized Contour Model by calculating the product of their confidence scores and identifying the pitch accent type with the greatest combined confidence. The results of this combination technique evaluated using ten-fold cross-validation on BDC-read, BDC-spontaneous and BURNC material can be found in Table 5.12.

Corpus	J48	SVM
BDC-read	76.36 ± 1.351 / 0.387 ± 0.0202	74.29 ± 0.891 / 0.373 ± 0.0264
BDC-spon	81.86 ± 0.853 / 0.422 ± 0.0183	84.38 ± 0.833 / 0.403 ± 0.0207
BURNC	62.29 ± 0.382 / 0.461 ± 0.00800	67.86 ± 0.635 / 0.472 ± 0.00785

Table 5.12: *Accuracy and CER of combining posteriors from Sequential Modeling of pitch and intensity with J48 and SMO confidence scores. Calculated over the pseudo-syllable region.*

We find that there is no significant classification performance by combining Quantized Contour Model posteriors with J48 or SVM classifier confidence scores. The best performance is achieved by J48 classification using aggregation features. This performs with a Combined Error Rate of 0.362. While the Quantized Contour Modeling performs competitively with the aggregation features in isolation, neither combination approach demonstrates improved overall classification performance. We believe the Quantized Contour Modeling captures distinct discriminative information about contour shape from the aggregation features. Though this belief suggests that there would be an opportunity for the two feature representations to combine to improved performance, neither approach we explored to this was able to demonstrate such an improvement.

5.6.4 Sampling strategies

Skewed class distributions can make it very difficult for machine learning algorithms to function successfully. The disparity in training data can lead to overwhelmingly high priors, and minority classes with data sparsity issues. Also, the modeling of the majority class will be more robust and yield more confident predictions. This disparity in modeling strength can lead to lower confidence in predictions of minority classes, classes which are already suffering from a low prior.

In situations in which class distributions are skewed it is common to train a model on a sample of the training data to make the classes more balanced. Skewed class distributions can negatively impact classification performance. Balancing the class distribution of training data can help ameliorate this problem. There are two commonly used sampling techniques: *undersampling* and *oversampling* [229]. When undersampling, a subset of the training data from the majority class is used in training the classifier. This subset may be selected either at random, or using feature analysis to be representative of the distribution of the full set training data. The size of the subset is chosen to generate a balanced class distribution. Oversampling, on the other hand, uses every training point of the majority class, but duplicates minority class training data points, such that the ultimate class distribution is equal. Undersampling has the problem of omitting viable majority class training data, and thus limiting the modeling capabilities of the classifier. While this may lead to poor modeling, it does not introduce any data integrity issues beyond potential data sparsity. Oversampling on the other hand, does introduce a data integrity problem. Say, for example, the data set has a 3:1 class ratio. When oversampling, each minority class data point is replicated three times. This makes the assumption that every minority class training point is equally likely to be observed two more times. However, there is no evidence for this assumption. Moreover, there is evidence against it. Data points near the centroid of the observations are much more likely to be observed again than data points far from this centroid. The impact of the presence of a minority class outlier is significantly increased by oversampling. It would be possible, of course, to impute new data points by sampling from a model of the minority class. Doing so would generate new data points near the observed centroid more often than far from it. However, this technique would lead to the generation of feature vectors corresponding to observations that do not exist. While statistically sound, the use of this technique raises questions about what exactly is being modeled under such a manipulation of training data.

There is a third sampling technique that is less commonly used than over- or undersampling. —em Ensemble sampling [235] divides the majority class into N subsets such that the size of each subset, when included with the minority class data points, generates a relatively balanced class distribution. Then, N classifiers are trained using the minority class data and one subset of the majority class data. The classifications of these N classifiers are then combined as a *post hoc* operation. This variation on undersampling is able to both use all of the available training data in the classifier and to avoid the problems of having heavily skewed training data. While other combination techniques may be used, we combine the classifications generated by ensemble sampling using the following formula,

$$type = \underset{type}{\text{argmax}} \prod_{i}^{N} conf_i(type|X) \tag{5.18}$$

where X is the classification feature vector and $conf_i$ the confidence of the classifier trained on training set i.

We experiment with each of these sampling strategies, evaluating J48 and SVM classification using ten-fold cross validation. In the experiments reported in Section 5.6.1, SVM classification demonstrated the greatest difficulty in classifying minority classes. Sampling the training data may be able to improve this. The feature vector for these experiments contains the acoustic aggregation features extracted over the pseudo-syllable region, as well as context-based normalizations. None of the shape features are included. The results of these experiments are reported in Tables 5.13 and 5.14.

Corpus	No Sampling	Under
BDC-read	74.55 ± 0.689 / 0.362 ± 0.0325	55.92 ± 1.542 / 0.356 ± 0.0279
BDC-spon	81.22 ± 0.689 / 0.370 ± 0.0236	53.27 ± 1.296 / 0.349 ± 0.0295
BURNC	61.37 ± 0.590 / 0.461 ± 0.0158	47.73 ± 1.312 / 0.441 ± 0.0305
Corpus	Over	Ensemble
BDC-read	70.64 ± 1.328 / 0.379 ± 0.0390	72.99 ± 1.164/ 0.383 ± 0.0276
BDC-spon	77.04 ± 0.886 / 0.407 ± 0.0318	78.80 ± 1.279 / 0.388 ± 0.0325
BURNC	55.08 ± 0.394 / 0.461 ± 0.0112	57.57 ± 0.771 / 0.461 ± 0.0225

Table 5.13: *Accuracy (%) and CER of J48 pitch accent type classification using undersampling, oversampling, and ensemble sampling with collapsed downstepped variants.*

We find that J48 classifiers do not benefit from sampling the training data. The Combined Error Rate is never significantly improved by sampling compared with an unsampled classifier. On the other hand, SVM classifiers show significant improvements to *CER* by undersampling and ensemble sampling. While J48 classifiers achieved a lower *CER* when trained on unsampled data, we obtain the best performance using undersampled and ensemble sampled SVM classifiers. In addition to the theoretical concerns about oversampling raised earlier, we find that oversampling performs significantly worse than

Corpus	No Sampling	Under
BDC-read	80.85 ± 1.033 / 0.402 ± 0.0251	64.95 ± 1.000 / 0.287 ± 0.0248
BDC-spon	84.57 ± 0.558 / 0.500 ± 0.0111	60.98 ± 1.099 / 0.286 ± 0.0253
BURNC	70.00 ± 0.607 / 0.500 ± 0.0121	57.49 ± 0.672 / 0.395 ± 0.0169

Corpus	Over	Ensemble
BDC-read	53.28 ± 1.4596 / 0.311 ± 0.0312	68.97 ± 1.164 / 0.270 ± 0.0325
BDC-spon	54.73 ± 1.558 / 0.296 ± 0.0290	61.50 ± 1.132 / 0.284 ± 0.0236
BURNC	40.98 ± 0.6724 / 0.390 ± 0.0117	60.26 ± 0.410 / 0.394 ± 0.0154

Table 5.14: *Accuracy (%) and CER of SVM pitch accent type classification using undersampling, oversampling, and ensemble sampling with collapsed downstepped variants.*

under- or ensemble sampling by all measures. Moreover, we find ensemble sampling to provide a consistent if statistically insignificant improvement over undersampling. Ensemble sampling on BDC-read with aggregation features yields the lowest observed CER at 0.270. Obviously, the associated 68.97% accuracy under a 78.24% majority class baseline shows that this technique has limited utility. However, this result does provide support for the claim that ensemble sampling is a helpful sampling strategy when classifying rare classes.

Next, we examine if undersampled or ensemble sampled SVM classifiers are more suited to combination with Quantized Contour Models described in Section 5.6.3. Results from a *post hoc* combination (product) of confidence scores can be found in Table 5.15. Note, the Quantized Contour Models are trained with the full training set; only the SVM classifier is trained with under- or ensemble sampling.

Corpus	Under	Ensemble
BDC-read	73.93 ± 1.108 / 0.363 ± 0.0168	75.92 ± 1.318 / 0.343 ± 0.0199
BDC-spon	81.08 ± 0.767 / 0.414 ± 0.0205	78.89 ± 2.363 / 0.380 ± 0.0426
BURNC	67.40 ± 0.503 / 0.461 ± 0.00900	68.17 ± 0.830 / 0.456 ± 0.0112

Table 5.15: *Accuracy (%) and CER of pitch accent type classification using* post hoc *combination of quantized shape modeling and SVM classification using undersampling and ensemble sampling.*

Combination of ensemble sampled SVM classification trained on aggregation features and Quantized Contour Model posteriors generates the higher CER than the ensemble sampled SVM classifiers in isolation on all three corpora; on BDC-read the ensemble of

SVM classification *CER* is 0.246, on BDC-spon it is 0.284, and on BURNC it is 0.394. While Quantized contour modeling may capture distinct information from the features comprising the SVM feature vector, combining these disparate representations to improve classification performance remains elusive. Identifying the best way to use Quantized Contour Modeling and the information it captures remains an unresolved issue.

Next, we evaluate the speaker dependence of the improvements generated by undersampling and ensemble sampling the training data in SVM classifiers. To evaluate this, we use leave-one-speaker-out cross-validation. Results from this evaluation can be found in Table 5.16. Along with the Accuracy and Combined Error Rates using unsampled, undersampled and ensemble sampled SVM classifiers with linear kernels, we report the difference in performance measures between this speaker independent evaluation and the speaker dependent, ten-fold cross-validation reported in Table 5.14.

Corpus	No Sampling	Under	Ensemble
BDC-read	78.69 ± 5.527 / 0.430 ± 0.0687	58.94 ± 7.708 / 0.352 ± 0.152	62.24 ± 5.363 / 0.339 ± 0.115
	-2.16 / +0.028	-6.01 / +0.065	-6.73 / +0.069
BDC-spon	84.55 ± 1.263 / 0.500/0.0163	63.41 ± 10.1188 / 0.325 ± 0.195	64.08 ± 7.938 / 0.329 ± 0.169
	-0.02 / 0.000	+2.43 / +0.029	+2.58 / +0.045
BURNC	70.00 ± 4.986 / 0.500 ± 0.0640	55.94 ± 5.9696 / 0.423 ± 0.123	59.94 ± 5.970 / 0.422 ± 0.132
	0.0 / 0.000	-1.55 / +0.028	-0.32 / +0.028

Table 5.16: *Speaker independent accuracy (%) and CER of pitch accent type classification using ensemble sampled SVM classification with aggregated acoustic features. Evaluated using Leave-one-speaker-out cross-validation.*

Under speaker independent evaluation, we still see the improvement due to training data sampling. While not significantly different, we observe a trend of ensemble sampled SVM classifiers performing with *CER* very similar to that of undersampled classifiers with a slight improvement on BDC-read material. The accuracy of the ensemble sampled classifiers is slightly, though not significantly, increased. These two trends leads us to prefer the use of ensemble sampling as a technique to address the skewed training data distribution.

We find the Combined Error Rate to increase under all speaker independent evaluations. The only exceptions to this are the unsampled classification scenarios on BDC-spontaneous and BURNC, where the performance was already equivalent to the majority class baseline. The differences between speaker independent versus speaker dependent training are not statistically significant. However, the consistency across experimental settings suggest that there is a systemic negative, though not very large, effect of speaker independent training. One surprising observation is that accuracy *increases* under speaker independent evaluation of BDC-spontaneous material. Since this increase in accuracy coincides with an increased Combined Error Rate, it suggests that speaker independent evaluation is correctly identifying H* (i.e. majority class) tokens. It is unclear why this phenomenon is only realized on the BDC-spontaneous material and on neither of the read speech corpora.

5.6.5 The influence of Phrase Accents on Pitch Accent Classification

Phrase boundaries have unique intonational characteristics. They are realized by increased disjuncture, indicated acoustically by the presence of silence, pitch and intensity resets, and pre-boundary lengthening. In addition to these indicators of disjuncture, pitch contours have a unique shape immediately preceding a phrase boundary. Under the ToBI convention, these phrase-final contour shapes are referred to as phrase accents, which describe the pitch shape following the final accent up to the end of each intermediate phrase. Additionally, intonational phrases are marked by a 'boundary tone' – describing the pitch contour behavior at the edge of the phrase. It is hypothesized that phrase accents and boundary tones indicate how the preceding and following phrase should be interpreted. This discussion (cf. Chapter 6) is somewhat tangental to the task at hand. However, the presence of phrase accents has the effect of changing the pitch contour of phrase-final words. If the phrase-final word is accented, the presence of a phrase accent or boundary tone may change the pitch accent contour shape, possibly leading to a degradation of classification performance.

To evaluate the impact of phrase-final intonation on pitch accent type classification, we inspect the confusion matrix from the ensemble sampled SVM classifier described in Section 5.6.4. The classifier is trained using acoustic aggregation features extracted from the pseudo-syllable region. We found this to be the most consistent pitch accent type classifier. We construct separate results for phrase-final and non-phrase-final tokens by partitioning the results of the ten-fold-cross-validation performed on this classifier. The results based on intermediate phrase boundaries are reported in Table 5.17.

Corpus	Phrase-Internal	Phrase-Final
BDC-read	72.91 ± 1.082 / 0.244 ± 0.0325	61.36 ± 1.771 / 0.343 ± 0.0430
BDC-spon	69.18 ± 0.886 / 0.266 ± 0.0380	51.01 ± 0.836 / 0.329 ± 0.0282
BURNC	63.90 ± 0.705 / 0.392 ± 0.197	55.67 ± 1.214 / 0.397 ± 0.0197

Table 5.17: *Accuracy (%) and CER ensemble sampled SVM classification results separated by phrase location.*

Across corpora, we find that phrase-internal accented words are more successfully classified than phrase-final accents. In addition to the interference from phrase accents, we also find that phrase-final accented words have a very different distribution of accents than phrase-internal accented words. For example, on BDC-spon, only 3.3% of phrase internal accents are L*, but 13.8% of phrase-final accents are. Also, 8.5% of phrase-internal accents are L+H*, while this accent type comprises only 3.3% of phrase-final accents. We see similar differences on BDC-read; internal accents are 13.8% L+H* and 2.4% L*, while final accents are 6.4% L+H* and 13.6% L*. On the BURNC data, we see roughly equivalent rates of L+H* accents on phrase-internal and phrase-final accented words; however, the L* rate nearly doubles from 2.6% to 5.1%.

To account for differences in accent distribution and to model the phrase accent interference, we train separate classifiers for phrase-internal and phrase-final pitch accents. We use manual phrase boundary annotations to determine the phrase position of each word. This is clearly a best-case, oracular, classification scenario. To determine the impact of phrase location on pitch accent type classification we want to isolate the effects of the presence of phrase accent and distributional changes from artifacts arising from automatic phrase boundary detection. Results from ten-fold cross validation experiments can be seen in Table 5.18. These experiments again use an ensemble sampled aggregation based SVM classifier.

Corpus	Phrase-Internal	Phrase-Final
BDC-read	76.52 ± 1.361 / 0.227 ± 0.0335 +3.61% / -.017	57.54 ± 1.738 / 0.355 ± 0.0469 -3.82% / +.012
BDC-spon	72.73 ± 1.410 / 0.247 ± 0.0428 +3.55% / -.019	66.60 ± 1.574 / 0.320 ± 0.0481 +15.59% / -.009
BURNC	61.42 ± 0.525 / 0.387 ± 0.0137 -2.48% / -.005	58.18 ± 1.099 / 0.376 ± 0.0225 +2.51% / -.021

Table 5.18: *Accuracy (%) and CER using class-based modeling to classify phrase-internal and phrase-final accents separately. Difference from baseline (i. e. non-class-based) classification is also reported.*

We find that this class-based modeling does not significantly improve classification performance. The *CER* is not significantly decreased in any corpusor phrase positions. Only on BDC-spon does the accuracy improve by a significant degree. We know from examination of the distribution of accent types by phrase position that there are significantly more phrase-final L* accents and fewer L+H* accents. When we compare the f-measures of pitch accent types on phrase-final and phrase-internal accents in the BDC-sponteanteous data, we find approximately equal f-measures on minority classes – a minor improvement to phrase-internal L+H* classification, a minor reduction in phrase-final L* classification – but we see clear increases to the f-measure of the majority class, H*. This indicates that the distinct pitch accent type distributions in these two phrase contexts lead to a *more* skewed class distribution, which, along with fewer minority class data points, leads to a classifier that is more likely to predict the majority class. That we do not observe a corresponding reduction in minority class classification performance suggests that there is a minor benefit of using this approach this corpus. However, this benefit appears to be due to the difference in pitch accent type distribution based on phrase context more than successful modeling or isolation of phrase accent effects.

Class based modeling is only successful if the improvements gained by treating each class separately outweigh the reduction in training data for each model. The small changes across the evaluations of BDC-read and BURNC suggest that it does not. Rather, having more available training data for each class is more valuable than isolating phrase-final from non-phrase-final tokens. This suggests that the aggregation-based features and quantized contour shape modeling are able to determine similarities between tokens of the same pitch accent class despite the presence of a phrase accent. That is not to suggest that the phrase accent does not have an impact on the shape of a pitch accent contour; it is clear that it does. Merely that these effects can either be modeled or ignored as noise. The later is more likely, as there is little training data to model for each pitch accent type/phrase accent combination. The salient extracted features capture qualities of the contour shape that are not modified by the presence of a phrase accent.

If we view a pitch accent as being made up of an accent onset, approaching the excursion correlating to the perception of accent, and offset, following the pitch and intensity peaks of the accent, these results suggest that the most discriminative features to pitch accent classification are extracted from the accent onset. Phrase accents describe the pitch contour between the final accent in a phrase and the phrase boundary. Thus, if phrase accents have any impact on phrase-final accent contours, this impact would modify the accent offset. The fact that we see little difference in modeling phrase-final accents separately from phrase-initial and phrase-medial accents suggests that the discriminative features that are used by the classifier are not extracted from the region affected by the presence of a phrase accent, specifically, the accent offset.

We find phrase-position-based modeling not to be particularly helpful to pitch accent type classification on BDC-read or BURNC material. We do find, however, that it provides some improvement to BDC-spontaneous classification, but that this is mostly due to changes in the pitch accent type distribution in phrase-final and phrase-internal contexts. The fact that phrase accents do not appear to dramatically reduce the performance of pitch accent type classification suggests that the pitch accent types are distinguished mainly based on their onsets, and that their overall shape may be less important than the pitch and intensity contour shape leading *into* the accent excursion.

5.6.6 BURNC syllable-based classification

In Section 5.6.3, we examined the use of energy-peak and pseudo-syllable segmentation to define a region of analysis for pitch accent type classification. Both of these techniques attempt to identify the accented region within a given word – hypothesized to be the syllable. On the BDC material, time-aligned phone or syllable data is unavailable, leaving these approximations to be the best available syllabification approaches.

The BURNC, on the other hand, includes forced alignment phone boundaries and a lexicon containing syllabification information. This provides another route towards identifying the syllable region of analysis. In this section, we evaluate the best performing pitch accent classification techniques from the previous sections extracting features and modeling contours extracted from this syllable region. Unfortunately the lexicon and time-aligned forced alignment output do not use the same phone inventory. Therefore, we align the syllable boundaries to the forced alignment hypothesis using a minimum edit distance routine to align the two phone sequences. In this dynamic programming alignment, we assign a cost of 0 to the alignment of any two vowels. This forces syllable nuclei to be aligned even if there are significant differences in the phone sequences.

First, we evaluate the use of the acoustic aggregations defined in 5.6.1. These are the mean, maximum, minimum and standard deviation of pitch and energy contours. These are extracted from raw and speaker normalized contours, as well as their slopes. Based on the success of the sampling approaches explored in 5.6.4, we evaluate the impact of undersampling and ensemble sampling on SVM classifier training. Based on the modest CER improvement obtained by phrase position based modeling (cf. Section 5.6.5), we also evaluate this classification modification on the BURNC syllable-based material. Finally, we evaluate the use of Quantized Contour Modeling (cf. Section 5.6.3 over the BURNC syllable region of analysis. We identify the best modeling strategy and quantization parameters and report the accuracy and *CER* obtained by this approach. Then we evaluate the posterior combination with the ensemble sampled SVM classifier. The results of these experiments can be found in Table 5.19.

Approach	Results	
	J48	SVM
Aggregate Features	62.73 ± 0.886 / 0.445 ± 0.0212	62.73 ± 0.886 / 0.446 ± 0.0212
Ensemble Sampling	63.27 ± 0.836 / 0.455 ± 0.0192	61.95 ± 0.394 / ***0.372*** ±0.0121
Undersampling	48.41 ± 1.542 / 0.439 ± 0.0287	58.23 ± 0.754 / 0.374 ± 0.0190
Phrase-pos'n Modeling	59.33 ± 1.197 / 0.456 ± 0.0200	60.86 ± 1.066 / ***0.364*** ±0.0202
Quantized Contour Model: f0 (4,4)	47.11 ± 0.8528 / 0.439 ± 0.0177	
	J48	SVM
QCM/Agg. combination	63.89 ± 0.754 / 0.453 ± 0.0210	50.48 ± 0.476 / 0.434 ± 0.0161

Table 5.19: *Accuracy (%) and CER of experiments evaluated on BURNC material at the syllable levels.*

We find the lowest Combined Error Rate on BURNC pitch accent type classification when using ensemble sampled SVM classification with aggregate features extracted over accented syllables. The lowest *CER* using the full word or pseudo-syllable regions was 0.394 (cf. Table 5.14), a difference of 0.022. This difference approaches significance with p= 0.098 when evaluated using a two tailed t-test. The performance of the ensemble sampled SVM classifier is improved by an additional 0.008 when phrase-internal and phrase-final accents are classified separately. This difference is not statistically significant. Moreover, this technique, obviously, requires intermediate phrase boundary information. As noted in Chapter 4, identifying intermediate phras boundaries is not a trivial task. These results demonstrate that forced-alignment based syllables represent the best unit of analysis for pitch accent type classification.

We also perform speaker independent evaluation of undersampled and ensemble sampled SVM classification of pitch accents, with features extracted at the syllable level. We find that performance is reduced when training material does not include any instances from the evaluated speaker. Ensemble sampled SVM classification performs with accuracy of 60.86 ± 5.888 and a Combined Error Rate of 0.408 ± 0.136. Undersampled SVM classification yields 57.21 ± 6.166 classification accuracy and a *CER* of 0.409 ± 0.135. Both the accuracy and *CER* of the speaker independent evaluation indicate impaired performance under speaker independent evaluation. However, the improvement due to the lexicon based syllable region of analysis over the pseudo-syllable region can still be observed (cf. Table 5.16). The accuracy is improved by approximately 1.5% and the *CER* is reduced by 0.014.

5.6.7 Use of Part-of-speech Information

While and accent's contour shape is an acoustic phenomenon, accent types are used to distinct communicative effect. In contrastive contexts, we are more likely to observe L+H* accents, for instance. In this section, we examine the potential of part-of-speech tags to predict the type of accent that is most likely to fall on a given word. That is, we address questions such as, is a *noun* more likely to be accented with an L* accent than a *verb*?

To address this and similar questions, we use automatically generated part-of-speech tags for the BDC subcorpora as well as the BURNC. These tags were generated using the Stanford Tagger [213], an automatic maximum entropy tagger trained on Penn Treebank [127] data – manually annotated Wall Street Journal text. The Penn Treebank tag set contains 35 part-of-speech tags. This dimensionality can lead to data sparsity problems. Thus we also collapse these tags into broader classes in two ways. First, we collapse the tags into six *broad* syntactic classes: *NOUN*, *VERB*, *ADJECTIVE*, *ADVERB*, *CARDINAL*, *FUNCTION*. Second, we collapse the 35 initial tags into just two part-of-speech classes, *FUNCTION* and *CONTENT* words, where *CONTENT* words are defined as members of the broad classes, *NOUN*, *VERB* (excepting modal or auxiliary), *ADJECTIVE*, *ADVERB*, and *CARDINAL*.

The first analysis we perform is to determine if the part-of-speech tag has any information about the likelihood of one accent type or another. That is, is the likelihood of a word being accented with an accent type independent of its part-of-speech tag. Figures 5.28, 5.29, and

5.30 contain the distributions of accent types by *broad* POS tags.

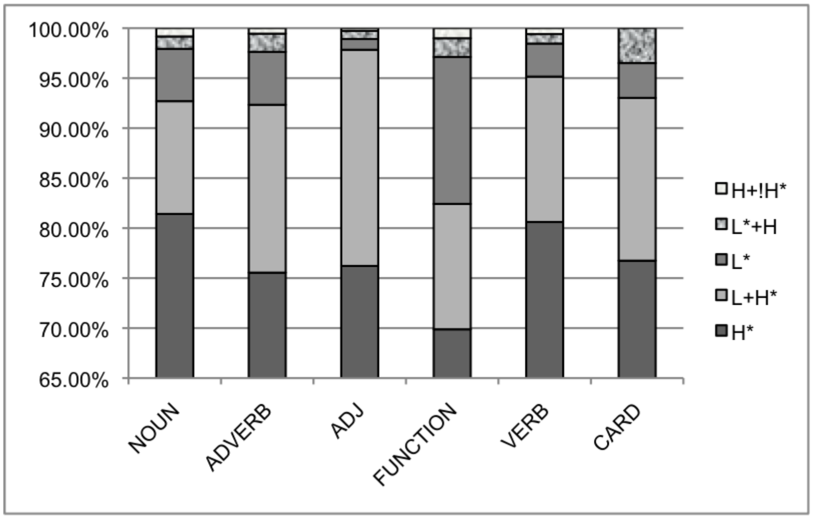


Figure 5.28: *Distribution of pitch accent types by broad POS classes on BDC-read.*

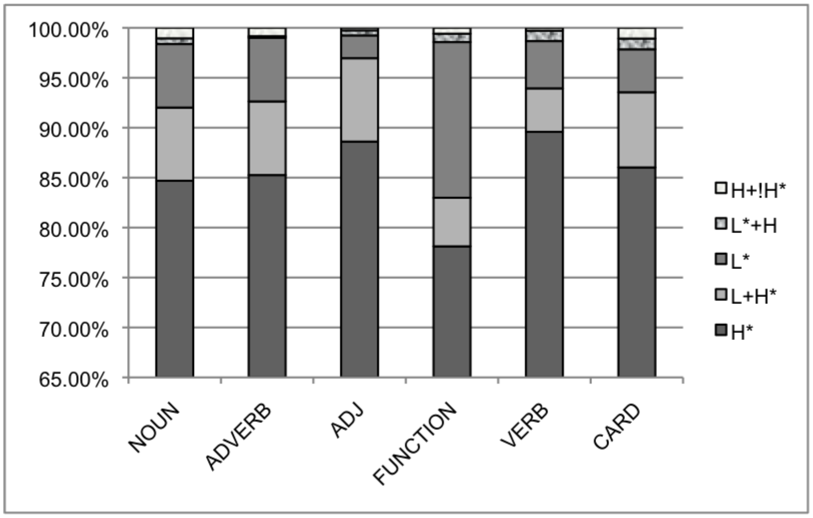


Figure 5.29: *Distribution of pitch accent types by broad POS classes on BDC-spontaneous.*

Pearson's χ^2 test indicates that distributions of *broad* POS tags and accent types are significantly dependent with $p < 1.0e - 10$ on all corpora. However, the contributing factors to this dependence are not consistent across corpora. In the BURNC material, we find a greater than expected rate of L+H* accents on *ADJECTIVE* and *ADVERB* tokens, and a greater than expected rate of H* tokens on *FUNCTION* words. On the BDC data, we see

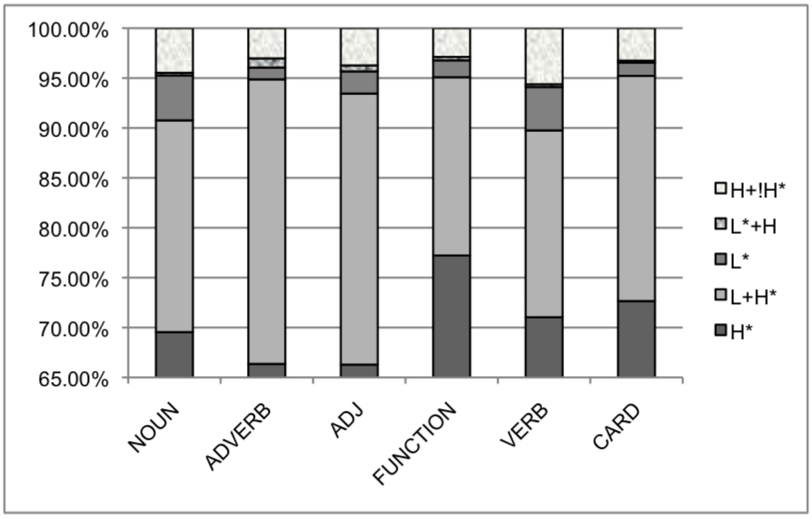


Figure 5.30: *Distribution of pitch accent types by broad POS classes on BURNC.*

significantly greater rates of L* accents on *FUNCTION* words, and lower rates of L* on adjectives. On the read subcorpora we observe higher than expected rates of L+H* on *ADJECTIVE* and *ADVERB* tokens; while this is also observed on the BDC-spontanteous data the distance from the expected value is reduced. Moreover, only on the BDC-read data do we find lower than expected rates of L* on *VERB* tokens.

There is a hypothesis that L* accents are more likely to fall on discourse given material. Modifiers such as adjectives and adverbs may be less likely to reiterate discourse given material. Modifiers are generally dropped after an entity enters the discourse. Take for example this simple dialog: “Could you pass me the *big* beaker?” “Thanks, now take the beaker back and hand me the notebook.” By occurring less frequently in discourse contexts where L* accenting is acceptable, these modifiers may be less commonly produced with this accent type. Conversely, we find that adjectives and adverbs *are* more likely to be accented with L+H* accents. This accent is often used to indicate contrast or emphasis. Modifiers may be used to distinguish one token from a previously unmodified (“Please, pass me the book.” “No, not that one, the *red* one.”) or contrastively modified token (“John has a red car, but I prefer *black* cars.”). This use of modifiers to draw contrast is a natural fit with the contrastive stress commonly indicated by L+H* accents.

We find that the *function/content* distinction is also not independent of pitch accent type. However, this is realized very differently on BDC data and BURNC data. On the BDC data, we observe that *function* words are accented with L* accents significantly more than expected. However, on BURNC material, *function* words are accented with H* accents more than expected, and are accented with other accent types significantly less often. In all three corpora, *FUNCTION* words are accented significantly less frequently than *CONTENT* words (cf. Chapter 3). It is unclear what communicative function the use of different accent types on *FUNCTION* words serves, though, this gives some evidence that it may be genre dependent. Another explanation of this is the labeling of accenting on *FUNCTION* words is annotator dependent, with some annotators being resistant to labeling *FUNCTION* words as accented (Hirschberg Personal Communication, April 2009).

Despite the fact that there is a statistically significant relationship between POS tags and accent types, POS tags alone cannot be used to successfully classify accent types. The most likely pitch accent type for each tag is H*, that is, $\text{argmax}_{type}\, p(type|tag) = H*$ for all tags. Sampling is not a viable solution to this, as it would have the effect of reducing, if not eliminating, the dependence between the two categories. However, we can include POS tag information in the feature vector of the aggregation-based classifiers introduced in Section 5.6.1 and used throughout Section 5.6. This allows the training algorithm access to POS tag information without tying it so tightly to the pitch accent type distribution.

We include three POS tag features, the *raw* tag, *broad class* tag, and *function/content* tag, in the feature vector along with the acoustic aggregations. We run ten-fold cross-validation experiments including these POS features with the aggregation-based feature set extracted from pseudo-syllables using ensemble sampling. This evaluation examines whether overall pitch accent type classification can be improved by access to POS tag information. Results from this evaluation are reported in Table 5.20. We include with these results the difference when compared to the classification experiments performed without the inclusion of POS attributes.

BDC-read	68.88 ± 0.869 / 0.278 ± 0.0266 +0.30% / +.008
BDC-spon	61.80 ± 1.214 / 0.281 ± 0.0385 +0.28% / -0.003
BURNC	61.92 ± 0.459 / 0.373 ± 0.00886 -0.03% / +0.001

Table 5.20: *Accuracy (%) and CER including nominal POS features in the aggregation-based feature vector with ensemble sampled SVM modeling. Differences from the same evaluation without POS features are also reported.*

We find no significant change in pitch accent type classification performance with the inclusion of part-of-speech based word class information to the SVM feature vector. While there are differences in pitch accent distribution depending on the word-class of the accented lexical item, the inclusion of this information does not significantly change the classification performance. It is possible that other lexical features may be used to predict pitch accent types. Parse tree features or a representation of information status may be able to differentiate likely pitch accent types based on lexical information. These features have not been examined and represent a direction for future work. The optimal approach to combining lexico-syntactic and acoustic information for pitch accent type classification also remains an open area for future research.

5.7 Conclusion and Future Work

In this chapter, we present a number of approaches to pitch accent type classification. A summary of the significant results contained in this chapter is presented in Table 5.21. We use the ToBI standard to define the inventory of pitch accent types, and, based on manual annotations using this standard, apply supervised learning techniques to the classification task. Human agreement with regard to pitch accent type under the ToBI standard is approximately 76.1% [156]. However, between 70 and 80% of pitch accents are H* accents, in the BURNC, BDC-read and BDC-spontaneous corpora. H* accents are characterized

	Classification Approach	BDC-read		BDC-spon		BURNC	
		J48	SVM	J48	SVM	J48	SVM
Word	Acoustic Aggregations	0.371	0.500	0.402	0.500	0.466	0.500
	Quantized Contour Modeling	0.375		0.401		0.435	
pseudo-syllable	Acoustic Aggregations	0.362	0.403	0.370	0.500	0.461	0.500
	Aggregations w/ Undersampling	0.356	0.287	0.349	0.286	0.441	0.395
	Aggregations w/ Ensemble Sampling	0.383	***0.270***	0.388	***0.284***	0.461	0.394
	Quantized Contour Modeling	0.367		0.407		0.439	
Syllable	Acoustic Aggregations	NA		NA		0.445	0.446
	Aggregations w/ Undersampling	NA		NA		0.439	0.374
	Aggregations w/ Ensemble Sampling	NA		NA		0.455	***0.372***
	Quantized Contour Modeling	NA		NA		0.439	

Table 5.21: A summary of select pitch accent type classification experiments. All evaluations use ten-fold cross-validation. Classification performances are reported using Combined Error Rate (CER). The best performance on each corpus is indicated in bold.

by a pitch excursion ending high in the speakers pitch range. This skewed distribution of pitch accent types complicates the classification task. First of all, there is relatively little training examples of minority classes. Second of all, the majority class baseline is high, making improvement difficult. Both of these, however, can be overcome. The rate of human agreement introduces another source of error. The annotation of the spoken material used in training and evaluation was not annotated by a single labeler. Thus there is some degree of inconsistency in the human annotation of the phenomenon. If this inconsistency is as high as reported by Pitrelli, et al. [156], it may not be possible to classify pitch accent types with accuracy greater than 76.1%. On BURNC and BDC material, a majority class baseline classifier – where every accent is classified as H* – will generate accuracy near the human agreement rate reported in the Pitrelli study. However, despite appearing to have high accuracy, such a classifier does not yield any pitch accent type information. To measure classification performance on minority classes, we define Combine Error Rate or CER (cf. Equation 5.6). We use this measure to evaluate the performance of pitch accent type classifiers throughout this chapter.

We explore a number of classification approaches using acoustic information to classify pitch accent types. We examine the use of a number of aggregation of pitch and duration information extracted from each accented word, as well as three pseudo-syllabification techniques: 1) one based on envelope valleys, 2) a more complicated technique which includes spectral and zero-crossing-rate information [219], and 3) alignment of lexicon based syllabification with forced-alignment predictions. We also examine a number of approaches to capturing the shape of the pitch contour within the accented word or region. When available, the lexicon based syllabification defines a region of analysis which leads to the best performing pitch accent type classification. However, forced-alignment and lexicon information is available only for the BURNC material. Moreover, we find that extraction of acoustic features from the full duration of an accented word does not perform significantly better than extraction of features from the pseudo-syllabification region derived from the technique presented in [219]. We find that the shape features explored in Section 5.6.3, including Tilt parameters, which were designed to differentiate types of prosodic events, are not able to classify pitch accents as well as the simple aggregations of pitch and intensity information. Furthermore, the combination of these shape features with the aggregation features does not improve classification performance.

In the discussion of features used to extract shape information from the pitch and intensity contours, we present a Bayesian modeling technique, Quantized Contour Modeling. Under this technique, acoustic contours are first quantized into a fixed number of time and value bins. Distinct models are trained for each time bin. (Gaussian and Multinomial models are evaluated in this chapter, but others could be used in their place.) During evaluation, the class that generates the greatest posterior is selected as the classifier prediction. Extensions of this approach to perform delta and sequential modeling are also proposed. The best performing parameterization of Quantized Contour Model is able to classify pitch accents with a CER, 0.375 that does not significantly differ from the best performance using acoustic aggregation features, 0.371. However, one of the approaches we investigate to combine the predictions of these two classification approaches is able to significantly reduce the error rate below that of the aggregation classifier.

One way to address skewed class distributions is to sample the classifier training data. We examine undersampling, oversampling and ensemble sampling. We find that by undersampling or ensemble sampling the training data for SVM classifiers, we can significantly improve pitch accent classification. Using ensemble sampling, the Combined Error Rate on BDC-read data is reduced from 0.371 to 0.270. This training modification is, by far, the most significant factor in improving pitch accent type classification above the majority class baseline.

We also notice that there is an influence of phrase accents on pitch accent classification. Phrase accents are phrase-final intonational phenomena which controls the shape of a pitch contour preceding an intermediate phrase boundary. When a pitch accent is realized on the final word of an intermediate phrase, classification performance is significantly worse. However, we find that overall classification accuracy can be improved if we model intermediate phrase-final accents separately from accents that fall earlier in an intermediate phrase. This, of course, requires high quality hypotheses of intermediate phrase boundaries, which is a difficult classification task.

There is an influence of part-of-speech based word-class information on pitch accent type distributions. That is, not all parts of speech are realized with the same pitch accent types. However, this distinction is not so significant as to be able to predict pitch accent type based only on word-class information with performance that rivals acoustic techniques. Moreover, the inclusion of this information with acoustic features does not improve classifier performance.

In addition to these experiments, we present descriptive analyses of pitch accent types, describing accented tokens from all three corpora used in the classification experiments: BDC-read, BDC-spontaneous and BURNC. In addition to this analysis, we focus on the distinction between H* and L+H* accents in the BURNC material. While many of the shape features examined in Section 5.6.3 are unable to significantly contribute to improve accent type classification, many show significant difference with respect to accent type. This descriptive analysis serves to inform future efforts in the classification of pitch accent type.

We would hope our experiments to show improvements to accuracy over the majority class baseline, but human agreement on this task is approximately 76%. This measurement was reported by Pitrelli et al. [156]; Syrdal and McGory reported agreement closer to 70%. Human agreement is lower than the majority class baseline on some corpora. However, we have demonstrated that significant error reductions can be observed using ensemble sampled SVM classifiers, and relatively basic aggregations of acoustic information. Even with perfect classifiers on the available data, we can expect to achieve accuracy in the range of 76% on unobserved data annotated by an unknown labeler. It is unfortunate that there are no reports of interlabeler reliability on different accent classes from the BURNC and

BDC material. Without this it is difficult to determine what the minority class pitch accent interannotator agreement is. This information could be used to calculate a more reasonable metric than accuracy to compare against automatic classifiers.

5.7.1 Key Observations

- **Accuracy is an imperfect measure of pitch accent type classification performance.** Due to the skewed distribution of pitch accent types, a majority class classifier, always prediction H*, is able to perform with accuracy near human agreement, despite having a degenerate decision making process. Accuracy is not sensitive to the classification performance on minority class tokens. To address this, we propose to use Combined Error Rate (CER), defined in Equation 5.6, to evaluate pitch accent type classification performance.

- **Ensemble sampling is a useful technique for training a classifier with a skewed class distribution.** Beyond the problem of evaluation, a skewed class distribution can make training a classifier quite difficult; the dearth of minority class training data can lead to overprediction of the majority class. In this chapter, we find ensemble sampling to be a sampling technique superior to undersampling and oversampling in addressing problem posed by the skewed class distribution.

- **The region of analysis significantly impacts classification performance.** Accents are typically realized within a subword region. While we find the full word to contain discriminative information to the presence or absence of pitch accent (cf. Chapter 3), limiting the region of analysis to either the pseudo-syllable or forced-alignment based syllable significantly improves pitch accent type classification performance.

- **There is an unrealized potential in using lexical features for pitch accent type classification.** The part-of-speech of an accented word has a significant influence on the pitch accent type. However, in our experiments, the inclusion of part-of-speech information does not lead to improved pitch accent type classification performance.

Chapter 6

Phrase-final Type Classification

6.1 Introduction

Just before a phrase boundary, a speaker's voice displays some unique characteristics. For example, typically, in Standard American English (SAE) when making a declarative statement, the pitch of the speech drops, as in the statement, "He drove a car.". Contrast this to the typical final rising pitch in the interrogative, "He drove a car?". Even from this simple example, it is clear that final intonation can significantly alter the intended interpretation of speech. This final rising pitch can also be used to indicate that the speaker is incredulous disbelief as in "I was wrong? And you were right?". There is also compelling evidence that in dialog, speakers signal their desire to hold or give up the turn, in part, via phrase-final intonation [65, 50].

This intonational behavior is a critical component of human communication, and therefore, important to spoken language processing tasks. The ability to recognize when a speaker is making a statement versus asking a question is very important for spoken dialog systems. Also, communication can severely suffer if one of these systems produces unexpected turn taking behavior – either by interrupting the user, or by not recognizing when he or she is waiting for the system to speak. Phrase ending intonation can also indicate that an utterance is incomplete in some way. High phrase ending tones are often used to indicate that there is more information in the next phrase that is relevant to the interpretation of the current phrase. In this way, phrase ending intonation informs a listener – human or machine – about the structure of the information conveyed via speech as well as its intended interpretation.

The ToBI standard segments phrase-final intonation into two components, a phrase accent and a boundary tone, either of which can be described as having a high (H) or low (L) tone. The phrase accent occurs at the end of each intermediate phrase, and describes the pitch contour from the last pitch accent to the end of the phrase. High phrase accents (H-) indicate that the pitch is rising or stable, while low phrase accents (L-) indicate a falling pitch. There is also a downstepped high phrase accent, which appears similar to a stable H- phrase accent, but falls to somewhere in the middle of the speaker's pitch range, as opposed to high in the pitch range. At full, intonational phrase boundaries, the phrase accent is followed by a boundary tone. High boundary tones (H%) indicate a final rise immediately prior to the boundary, while low tones (L%) indicate the presence of a stable or falling pitch.

Each intonational phrase boundary thus has associated with it a phrase accent and boundary tone. In this chapter, we treat these as a single unit. Thus, the inventory of types we consider here are L-L%, L-H%, !H-L%, H-L% and H-H%. For brevity, throughout this chapter we will refer to a pair of a phrase accent and a boundary tone as a "phrase-final". This is not an established term in the intonational literature, but will facilitate the discussion of the simultaneous classification of phrase accents and boundary tones. We treat this classification as a five-way problem instead of looking at independent classes of phrase accent and boundary tone due to their acoustic characteristics and their communicative uses.

The acoustic realization of phrase accents and boundary tones are heavily dependent on one another. Boundary tones appear quite different whether preceded by a low or high phrase accent. Conversely, phrase accents are realized with different slopes and at different locations within a speaker's pitch range depending on the following boundary tone. This strong dependency between the two events and the lack of an available technique to distinguish when a phrase accent ends and a subsequent boundary tone begins, leads us to treat the combination of phrase accent and boundary tone as a single unit. While the prototypical versions of these four contours can be neatly decomposed into phrase accents and boundary tones, in practice, the decoupling of these two effects is virtually impossible. More detailed examination of prototypical and ambiguous examples of the five phrase-final types is presented in Section 6.2. Stylized representations of the inventory of phrase-final types are presented in Figure 6.1.

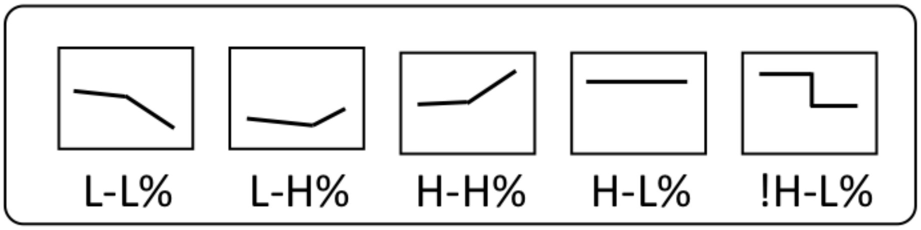


Figure 6.1: *Stylized representations of phrase-final types. All are assumed to be preceded by H* accents.*

Beyond the acoustic domain, the communicative uses of phrase-finals do not neatly decouple across the dimensions of phrase accent and boundary tone. Pierrehumbert and Hirschberg [153] who posited a compositional approach to the interpretation of categorical prosodic events had difficulty with this decoupling. They posit that H% can signal a hierarchical relationship between the intentions of the current and subsequent utterance, or satisfaction-precedence relationship [71], while phrases ending with L% can be interpreted without reference to a subsequent utterance. Moreover, they found that H- phrase accents and H% boundary tones both indicate that a phrase should be interpreted with particular respect to the following phrase, while boundary tones refer to relationships between intonational phrases, and phrase accents between intermediate phrases. They found evidence that H- can be seen to form part of a larger composite when used within an intonational phrase, with L- phrase accents highlighting the separation of the component intermediate phrases. However, when both a phrase accent and boundary tone are present, particularly in the case of intonational phrases containing a single intermediate phrase, they note “it is more difficult to separate the meaning of the phrase accent from the meaning of the boundary tone". In broad strokes, their findings suggest that high tones (H- or H%) indicate some degree of connection to the subsequent phrase, while low tones (L- or L%) indicate greater completeness of the phrase. Bartels similarly proposes that L- phrase accents are essential to "declarative intonation", where H- phrase accents are rarely used in statements, being much more common in questioning intonation [10]. She cites boundary tone effects as "potential discourse finality" and "continuation dependence". Phrases that end with a low boundary tone, L%, are potentially discourse final, while those containing high boundary tones, H%, contain some continuation dependence. This understanding of boundary tones is consistent with that of Pierrehumbert and Hirschberg.

In addition to these compositional approaches to contour interpretation, interpretations and communicative uses of phrase-finals, where the phrase accent and boundary tone are considered as a unit, have also been investigated. The declarative contours including the phrase-final intonation, L-L% tend to receive very little attention in these examinations. It is generally considered to be the default phrase-final, and indicative of a greater degree of finality, than the others. The the shallow rise, or "continuation rise", contour, ending with L-H%, communicates that there is "more to come" [26], and is frequently used in list intonation [92]. In addition to these structural implications, the L*+H L-H% contour can indicate uncertainty or incredulity [109, 224, 92]. H-H% convey the impression that the speaker is seeking confirmation. This is commonly used in contours to indicate yes-no questions, tag questions and declarative questions. However, contours ending in H-H% may also be interpreted the speaker showing deference to the listener [112], or seeking confirmation [73]. Gravano et al. [67] observed that utterances produced with downstepped contours, H* !H* (!H*) L-L%, are perceived as more certain than utterance produced with standard declarative contours H* (H*) L-L%, which are, in turn, perceived as more certain than those produced with a yes-no question contour, L* (L*) H-H%. Contours ending in a plateau, H-L%, convey a degree of incompleteness, and may be used to hold the turn [65, 50].

This turn-holding use of the plateau has been observed in many languages including German [179], Japanese [105], and Dutch [34]. The contours ending with a plateau also carry a bored, recitation effect with the implication, “you already know this” [82]. Moreover, it may also be indicative of negative emotion [18].

In this chapter, we address the task of classifying phrase-final intonation. The ToBI standard decomposes its description of phrase ending prosody at intonational phrase boundaries into two components: a phrase accent and boundary tone. In the work presented in this chapter we classify these two types as a single unit. That is, rather than classify phrase accents independently from boundary tones, we classify phrase-final intonation as a whole. This combination yields an inventory of five types of phrase-ending intonations: L-L%, L-H%, H-L%, !H-L% and H-H%. We make the decision to classify these two tones as a single event for two reasons. The value of a phrase accent hypothesis is severely limited without a corresponding boundary tone hypothesis, and vice versa. The intended interpretation of the two tonal components are inextricably tied together. Any spoken language processing task that can benefit by having access to information about phrase ending intonation requires information about both the phrase accent and the boundary tone. The communicative impact of phrase accents type distinctions as distinct from boundary tones type distinctions has not been isolated. However, such a requirement does not preclude the generation of independent phrase accent and boundary tone hypotheses. The two can be merged before delivery to a downstream system. This leads to the second reason to classify the two tones as a single event. The acoustic realization of the two tones are *also* inextricably linked. The realization of a boundary tone is significantly different based on the tone of its associated phrase accent, and vice versa. Isolating the boundary between the influence of a phrase accent and a boundary tone is at best difficult, at worst impossible; the two tones are tightly coordinated in their acoustic influence. By treating these two tones as a single prosodic event, we are able to assess their combined acoustic realization, rather than needing to separately identify the influence of pitch accent and boundary tone. For brevity we will refer to the combination of a phrase accent and a boundary tone as a “phrase-final” throughout this chapter.

This chapter is structured as follows. We will explore and evaluate acoustic (Section 6.5.1) and syntactic (Section 6.5.2) features to distinguish them. Much of the acoustic material that characterizes phrase-final intonation occur immediately prior to a phrase boundary. While this intuition suggests that the phrase boundary itself should represent the region of analysis should end, there is no indication of where the optimal region of analysis should begin. In Section 6.5.3, we address this question – where should we look for acoustic information discriminative to phrase-final type? The experiments presented in this chapter contain novel approaches to phrase-final classification, and yield state-of-the-art automatic classification results. However, there is still room for improvement, both in closing the gap from speaker-dependent to speaker-independent performance, and the overall classification accuracy. In Section 6.6, we will use the techniques examined in in Section 6.5 to classify intermediate phrase-final phrase accents. In Section 6.7, we summarize our findings and provide some directions for future research.

6.2 Examples of Phrase-Final Types

In this section, we present examples of the five intonational phrase-final types: L-L%, L-H%, H-L%, !H-L% and H-H%. Drawing on clear and confusable examples of each, we will discuss the defining characteristics of each of these types.

In read speech, the most common phrase-final is L-L%. This is most strongly associated with “the declarative contour”, H* (H*) L-L%, which is most commonly used at the end of a declarative utterance. This phrase-final is generally considered to be the most final, or complete. It is characterized by a steady decline in pitch from the final pitch accent leading into the phrase boundary typically with the final pitch low in the speaker’s pitch range. A clear example of an L-L% can be found in Figure 6.2. The characteristic decline in pitch leading into the phrase boundary can be clearly seen in this example. A more ambiguous

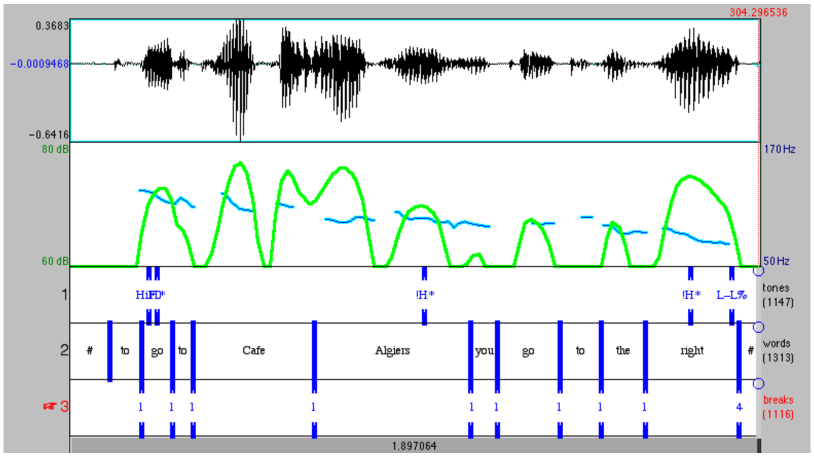


Figure 6.2: *A clear example of L-L%, drawn from file h2s7 of the BDC-spontaneous corpus.*

example of an L-L% phrase-final can be seen in Figure 6.3. This example contains pitch points that appear to rise following the decline in pitch. Without listening to the example, this increase in pitch could make the contour appear to be a L-H% phrase-final. When listening to the example, it is clear that the periodic acoustic material is a release of breath following the production of "Line" and not meaningful intonational rise. Pitch tracking artifacts such as this are a significant source of human disagreement and classification confusion in phrase-final classification. As a speaker ends an intonational phrase, the intensity of their speech signal commonly decreases significantly. This low intensity can lead a pitch tracking

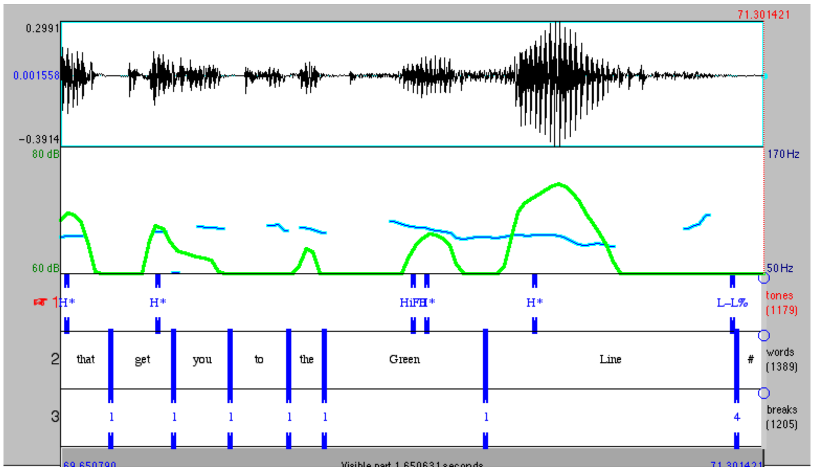


Figure 6.3: *A ambiguous example of L-L%, drawn from file h2s9 of the BDC-spontaneous corpus.*

algorithm to identify periodicity in vocal effects such as a phrase ending release of breath. This is not a pitch tracking error – there is periodicity in the acoustic signal – yet since this periodicity is not related to the intonational contour it can make automatic classification of phrase-finals more difficult.

The "plateau" contour, H* H-L%, is characterized by a sustained tone, towards the high to middle range of the speakers pitch range, from the last pitch accent to the intonational phrase boundary. Depending on the speaker, this sustained tone can rise or fall slightly, but in the canonical H-L%, the pitch is flat. A clear example of a H-L% phrase-final can be seen in Figure 6.4. Note the sustained, flat pitch from the final accent in the phrase until the speaker stops vocalizing. The small downward hook in the pitch is quite typical in plateaus and is due to the release of subglottal pressure as the speaker ends their vocalization. The example shown in Figure 6.5 is confusable with both L-L% and L-H% phrase-finals. The decline in pitch over the course of the word, particularly from the accent to the middle of "well" could indicate a L-L% contour. However, this contour falls rather too high in the speaker's pitch range to be considered a declarative contour. This is an example where having information about a speaker beyond the current phrase is critical for the analysis of their prosody. The same pitch contour produced by a speaker with a high pitch range may,

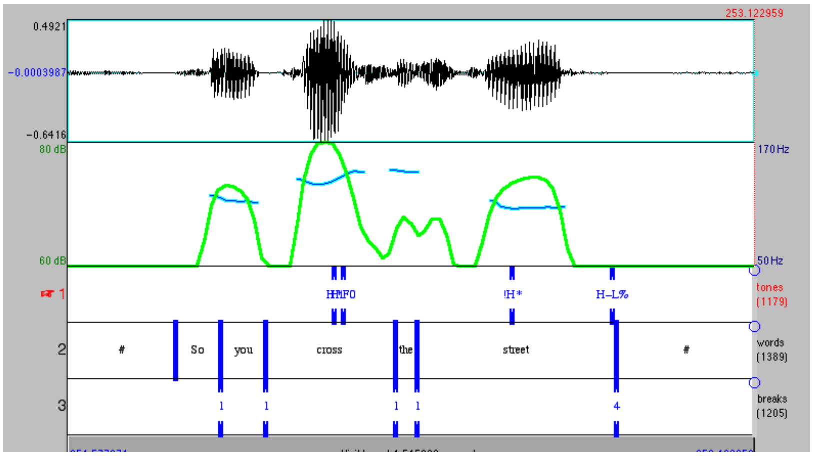


Figure 6.4: *A clear example of H-L%, drawn from file h2s9 of the BDC-spontaneous corpus.*

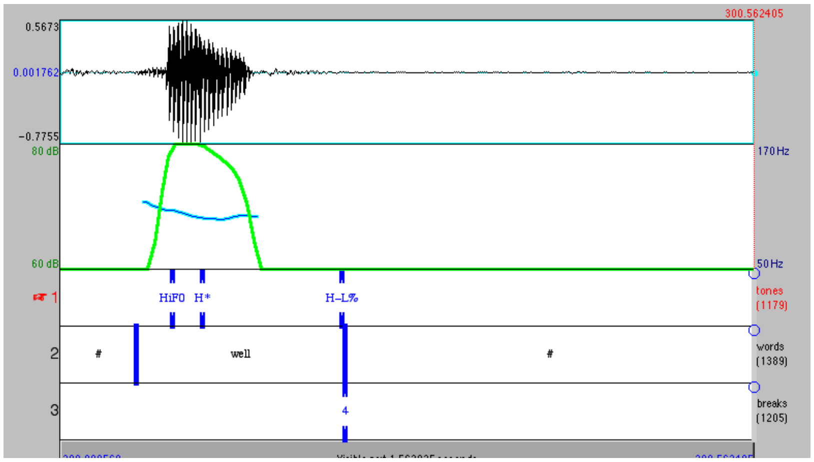


Figure 6.5: *An ambiguous example of H-L%, drawn from file h2s9 of the BDC-spontaneous corpus.*

in fact, be declarative. Normalizing pitch information to account for speaker differences can be used to reduce the impact of some of this ambiguity. Another source of ambiguity in this example, is that the contour reaches a local minimum approximately two thirds through its duration, followed by a very slight rise. While this is only a shallow rise, a local minimum followed by a rise is characteristic of a typical L-H% phrase final, as we shall see below. Differentiating a rise that should be ignored as part of a generally sustained tone from one that is indicative of a rising contour can be difficult for automatic classification techniques.

One of the least commonly used phrase accents, in the BURNC and BDC material, is the downstepped high (!H-). It is only used at intonational phrase boundaries with a low boundary tone (L%). This phrase accent/boundary tone combination has a sustained pitch, similar to the plateau (H-L%), but the sustain occurs at a pitch lower than height of the speaker's previous pitch range, but not higher than their low. An example of this can be seen in Figure 6.6. A typical characteristic of this phrase-final is an "elbow" in the pitch contour – a drop from a relatively high pitch to a sustained middle range pitch. In Standard American English, the !H-L% phrase-final tone combination is typically used in a "calling contour" – H* !H-L%. This is comprised of a "chanted tune on two sustained notes, stepping down from a fairly high level to a somewhat lower level" [110]. The calling

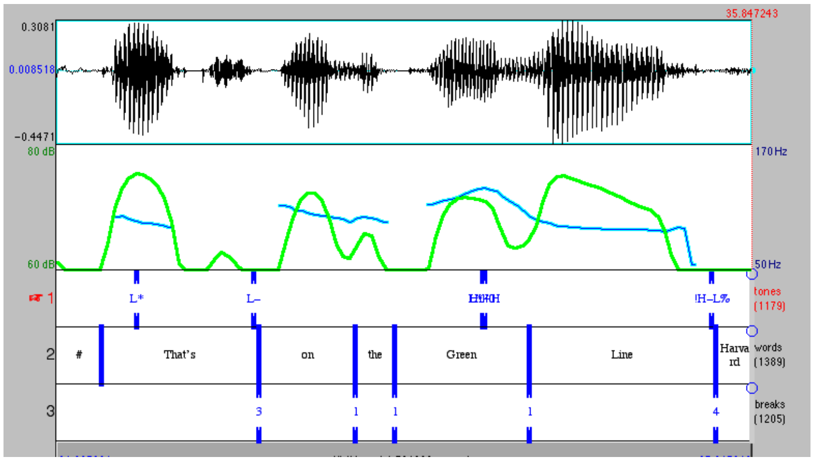


Figure 6.6: *A clear example of !H-L%, drawn from file h2s9 of the BDC-spontaneous corpus.*

contour has a stereotypical vocative use of a parent calling a child, but this phrase ending tone combination can be found in other contexts as well. For example, in the material in Figure 6.6, the !H-L% is the phrase-final tone combination of the intonational phrase "on the Green Line". The larger context of this phrase is "That's on the Green Line. Harvard Square's on the Red Line". The parallel structure of these two sentences suggests that they should be interpreted together, and the phrase accent and boundary tone has be impact of indicating the continuity between the two sentences. Careful investigation of the use and effect of this phrase-final is required to make a proper hypotheses about its communicative impact, but this anecdote suggests that there may be a compositional effect in addition to the stereotypical vocative use of !H-L%.

There are two issues that can make the classification of the !H-L% phrase-final difficult. One is the identification of the sustained tone preceding the phrase boundary. If the sustain is short, the phrase-final may be appear more like an L-L% that does not fully realize the typical low pitch. Second, once identified the sustain must be determined to be at a lower pitch than a typical H-L% would be, but not so low as to be considered an L-L%. In the example presented in Figure 6.7, despite having a pitch that drops to the speaker's low just before the phrase boundary, the labeler has identified a sustain immediately preceding this drop in the middle of the speaker's pitch range, and the characteristic elbow in the pitch track.

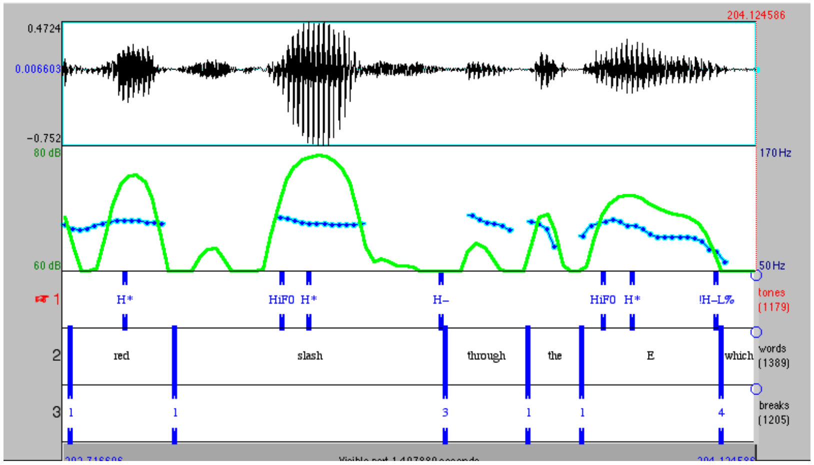


Figure 6.7: *An ambiguous example of !H-L%, drawn from file h2s9 of the BDC-spontaneous corpus.*

The L-H% phrase-final is typically used in "shallow rise" or "continuation rise" contours, H* L-H%. This contour is referred to a *shallow* rise refers to the fact that the pitch contour rises in a L-H% phrase-final – as indicated by the H% boundary tone – though this rise is not as great as in the *high* rise contour containing an H-H% phrase final, discussed below. The term "continuation rise" refers to the fact that this contour is often used, in Standard American English, to indicate that the speaker has more to say or that there is "more to come" [26]. More specifically, a continuation rise contour describes a subordination relationship between the current and following phrase [89]. The continuation rise may be used to indicate that a topic is incomplete – it is commonly used when speaking a list – or, in dialog, that the speaker intends to hold the turn [84]. Moreover, this phrase final, when preceded by one or more L*+H pitch accents, as in a L+H* (L+H*) L-H% contour, can be used to indicate uncertainty or incredulity [109, 224, 92].

A clear example of an L-H% phrase-final is presented in Figure 6.8. The characteristic features of this pitch contour are a fall following the nuclear accent in the intonational phrase, followed by a rise prior to the phrase boundary. This rise is typically ends in the middle to

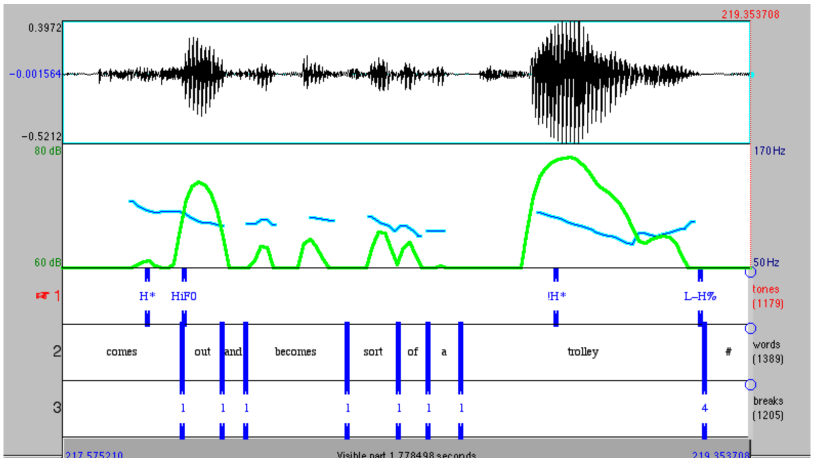


Figure 6.8: *A clear example of L-H%, drawn from file h2s9 of the BDC-spontaneous corpus.*

high range of the speaker's pitch range. This fall and subsequent rise can be easily observed in this example. Comparing this to the confusable plateau example (cf. Figure 6.5), we can see that it may not be sufficient to identify the presence of any fall and subsequent rise, but rather that the fall and rise each must be sufficiently steep in order to be considered a L-H% contour. On the other hand, if we examine a more confusable example of an L-H% contour, we can see that this criteria may not be sufficient. The example in Figure 6.9 shows an L-H% contour with a shallow rise. While the fall from the accented syllable in "market" is pronounced, the subsequent rise is quite slight. Differentiating a slight rise that is indicative

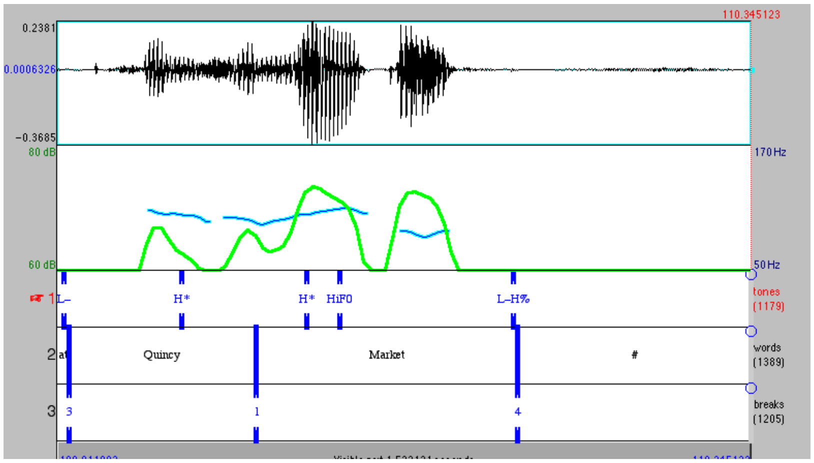


Figure 6.9: *An ambiguous example of L-H%, drawn from file h2s9 of the BDC-spontaneous corpus.*

of a L-H% contour from a minor perturbation in a plateau or a declarative contour can be a challenging task for humans as well as machines.

The tone of the nuclear pitch accent can modify the acoustic realization of a phrase-final. In the case of L-H% phrase-finals, the characteristic fall is only realized if the preceding pitch accent contained a high tone. While more than 90% of pitch accents in BURNC and BDC material do contain a high tone (cf. Sections 2 and 5.4), L* accents do not. In L* L-H% contours, there is no fall from the pitch accent to the phrase accent, rather a sustained low. This effect can also be seen in L* L-L% contours, where there is no fall associated with the L-L% phrase final, but rather a sustained low pitch. While these cases are not common in the material used by the experiments in this chapter, nuclear L* accents represent a potential source of difficulty for automatic phrase-final classification techniques.

The "high rise" phrase-final, H-H%, is commonly used interrogative contours, specifically, yes-no questions, L* H-H%. Ward and Tsukahara also found H-H% phrase finals to serve as a backchanneling cue in two-party dialogs, where it is used to request confirmation [225]. This use of the H-H% phrase final in contours to elicit confirmation was also observed by Hirschberg and Ward when examining high-rising, H* (H*) H-H%, responses to questions [93]. This contour, H* (H*) H-H%, is also associated with the much derided "uptalk" commonly associated with certain segments of adolescent populations in English speaking cultures [55, 226]. This phrase-final is characterized by a final pitch rise that starts in the middle or high in the pitch range and continues to rise until the phrase boundary. Figure 6.10 contains a clear example of a contour ending with an H-H% phrase-final. The in H-H% phrase finals can be similar to that observed in a typical L-H%, but the H-H% lacks the fall preceding this rise and contains a steeper rise. This is the case in the ambiguous example presented in Figure 6.11. In this example we see that the rising following the final pitch accent in the phrase to be very minor. The lack of a preceding fall, and the steepness of the rise, indicates that this is not a L-H% phrase-final. Moreover, this is somewhat ambiguous with an H-L%, plateau phrase-final. The rise is shallow enough that the contour may appear

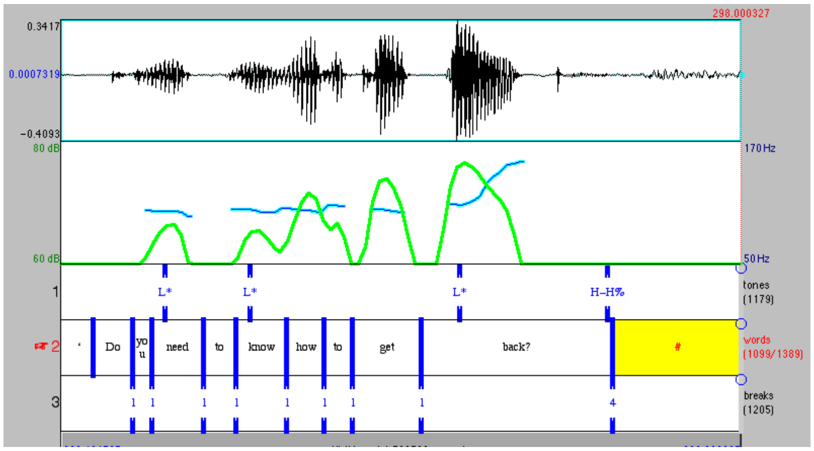


Figure 6.10: *A clear example of H-H%, drawn from file h2s9 of the BDC-spontaneous corpus.*

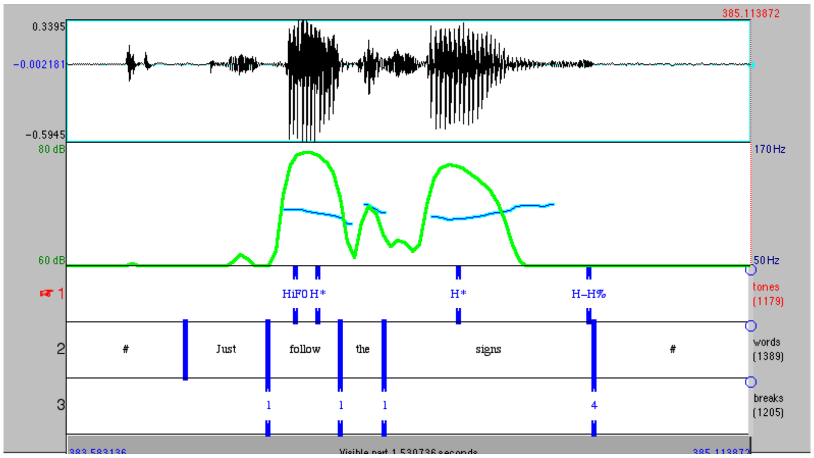


Figure 6.11: *An ambiguous example of H-H%, drawn from file h2s9 of the BDC-spontaneous corpus.*

to be a sustained tone, as opposed to a rising tone. The threshold for differentiating these must be empirically determined. This threshold most likely shows some degree of speaker dependence – some speakers produce greater rises than others – and some degree of labeler dependence – some labelers may require a greater rise to label a H-H%.

6.3 Related Work

There is less previous work on automatic phrase-final classification than other prosodic event detection and classification tasks. This is somewhat striking given the broad range of communicative effects in which phrase-final behavior participates.

A commonly cited evaluation of human agreement of ToBI annotation [196] evaluated pairwise interannotator agreement for phrase-finals (referred to as edge-tones) at approximately 85%. This established the highest automatic classification performance that can be expected without overfitting the behavior a particular human labeler. Moreover, manually annotated corpora are often labeled by multiple ToBI experts. In these situations, the internal consistency of the corpus itself is limited by the interannotator agreement, limiting the maximum performance achievable by supervised learning algorithms.

The earliest instance of phrase-final classification used only lexical information to determine the desired phrase-final type for a speech synthesis application [173]. In this work a decision tree was trained to assign one of three phrase-final types, L-L% L-H% and H-L%, using part-of-speech, information status, the location of the phrase boundary with respect to the surrounding sentence and paragraph, and the number of syllables in the phrase and since the previous pitch accent. When trained and evaluated on the material of a single speaker in the Boston University Radio News Corpus (BURNC), this approach correctly predicted 72.4% of phrase-finals over a majority class (L-L%) baseline of 61.1%. This is a heavily speaker dependent evaluation but represents an early success in the prediction of phrase finals.

Tepperman and Narayanan [208] used a tree grammar to model prosodic structure similar to that developed by Hirschberg and Rambow [91]. They were able to model the relationship between syllable stress and accents, intermediate and intonational phrases with this tree grammar better than with n-gram models. The authors evaluated this model's performance, trained as a Weighted Regular Tree Grammar, by predicting prosodic event categories with all but the target label given. Under this evaluation routine, the weighted tree grammar was able to correctly predict 75.08% of boundary tones and 88.22% of phrase accents, under a speaker-independent evaluation. This accuracy is quite high, however, the leave-one-label-out evaluation provides a substantial additional information to the classification technique that cannot be expected to be available in an automatic classification setting.

These approaches analyzed the use of lexical content to predict phrase-finals. Other work has addressed the acoustic correlates of phrase-finals in speech. Early work by Price, Ostendorf and Wightman [161] was able to distinguish L-L% and L-H% phrase finals with 76% accuracy. This classification was performed using an HMM model decoding quantized F0 values and was evaluated on a single radio news speaker.

Murray [136] used boundary tone classification as an application to evaluate the negative impact of pitch tracking errors. This work classified final-rise (H%) and final-fall (L%) boundary tones using a single, hand written rule. The rule was that if the mean pitch of the final vocalized region was higher than the mean pitch of the rest of the utterance, or if the slope of the pitch within the final vocalized region was positive, the phrase final was considered a final-rise (H%). Otherwise, it was considered final-fall (L%). While the goal of this study was to examine the impact of pitch tracking errors, not to generate high accuracy phrase-final predictions, this single rule was able to differentiate these two boundary tones with 68% accuracy using automatic pitch information. This accuracy increased to 77% if the pitch track is manually corrected. This work provides objective measurement of the degree to which pitch tracking errors negatively impact the performance of all automatic phrase-final classification techniques.

Ishi [97] classified phrase-finals in Japanese using an inventory of six types: long rise, short rise, long flat, short flat, weak flat and long fall as proposed by [211]. This work extracts aggregations of the pitch contour from the first and second half of the *phrase-final* region. The *phrase-final* region is defined as the vowel and coda of the phrase-final syllable. Using features that relate the aggregations drawn from the first and second half of the phrase-final region, Ishi trained a decision tree that was able to classify the phrase-finals of a single speaker with 79.2% accuracy.

In English, Ananthakrishnan and Narayanan [5] examined the use of F0 parameterizations for boundary tone classification. Evaluating the performance of Tilt (cf. Section 5.6.3) and RFC parameters, they were able to classify L-L% from L-H% phrase finals with 67.7% accuracy under speaker-independent evaluation. This performance was obtained using RFC F0 parameterization and a language model trained over categorical prosodic events. This work was evaluated on the BURNC.

In this chapter, we build on the syntactic findings of Ross using representations of part-of-spech information and including syntactic parse tree features. We are inspired by the acoustic findings of Ishi and Ananthakrishnan in modeling the pitch contour shape. We examine the use of structural syntactic features in Section 6.5.2. In Section 6.5.1, we examine the features that attempt to capture the shape of the phrase-final pitch contour, and in Section 6.5.3, we evaluate a variety of regions-of-analysis for phrase-final classification.

6.4 Materials

In this chapter, we describe a number of experiments in classifying phrase-final types. These experiments are performed on the read and spontaneous subcorpora of the Boston Directions Corpus (BDC) and the Boston University Radio News Corpus (BURNC). The phrase-final type distributions from these three corpora can be found in Table 6.1. The ToBI standard allows for the annotation of a phrase accent as X-? and a boundary tone as X%? when the annotator is uncertain of the tone of these events. Phrase finals containing either of these labels are omitted from the experimental data used in the following experiment sections.

Corpus	L-L%	L-H%	!H-L%	H-L%	H-H%
BDC-read	49.00% (685)	35.62% (498)	4.29% (60)	9.67% (135)	1.43% (20)
BDC-spon	29.57% (642)	32.33% (701)	4.42% (96)	31.09% (675)	2.63% (57)
BURNC	56.16% (3130)	38.38% (2139)	0.68% (38)	4.25% (199)	1.20% (67)

Table 6.1: *Distribution of phrase-final types*

We observe rather dramatic differences across corpora in the distribution of phrase-final types. The Broadcast News (BN) speech that comprises the BURNC corpus, speakers use the declarative (L-L%) and continuation rise (L-H%) contours nearly to the exclusion of all others. This is likely due to the broadcast news speaking style. The difference between the professionally read speech in the BURNC data and the non-professionally read speech in the BDC-read subcorpus is largely in the use of H-L%. We can observe a reduced use of L-L% and a corresponding increase in the use of H-L% in the BDC-read corpus when compared to the BURNC material. The use of H-L% is still more pronounced in the BDC-spontaneous material. While comprising merely 4.3% of phrase finals in the BURNC corpus, H-L% represents 31.09% of phrase-finals in the spontaneous subcorpus of the BDC data. As the speakers in the two subcorpora are identical, the distributional difference between BDC-read and BDC-spontaneous can only be explained by a difference in genre.

6.5 Experiment Results and Discussion

In this section, we describe a number of experiments classifying phrase-final types. These experiments are divided into four groups: analyses of the usefulness of acoustic features (Section 6.5.1), analysis of the descriptive power of syntactic features (Section 6.5.2), variation of the region of analysis (Section 6.5.3), quantized contour modeling (Section 6.5.4), and segmental class modeling (Section 6.5.5).

In Section 6.4, we note strong genre effects on the distribution of phrase-final types. These effects impact the classification performance of classifiers trained and evaluated on each of the three corpora, BDC-read, BDC-spontaneous, and BURNC. The majority class membership of each of these corpora significantly differ, from 32.33% for BDC-spontaneous, to 49.00% for BDC-read to 56.16% for BURNC. In many of the experiments presented in this section, we observe the lowest accuracy on BDC-spontaneous, and the highest on BURNC material. This wide range in majority-class baseline classification accuracy makes it difficult to determine if the effects of genre on this task are a matter of speaking style – spontaneous phrase finals are more ambiguous than those produced by a professional speaker – or a matter of distribution – phrase-final classification on BURNC data may be easier because it approximates a binary classification problem as opposed to the three-way classification required on the BDC material.

6.5.1 Acoustic features

In this subsection, we examine the acoustic features which differentiate the four classes of phrase final types. In the ToBI standard, the description of intonational phrase-final behavior — phrase accents and boundary tones — refers only to the pitch of the speech immediately preceding a phrase boundary. Therefore, our analyses are focussed on pitch (F0) features. That said, pitch and intensity are closely correlated; typically, the subglottal pressure necessary to produce an increase in pitch leads to a concurrent increase in intensity. Measures of the intensity of the speech signal are more reliable than F0 measures; they operate in non-sonorant regions, and do not suffer from halving or doubling artifacts. For these reasons, we also extract acoustic features from the energy contour along with the pitch contour.

The initial set of acoustic features are comprised of simple aggregations of the pitch contour throughout the phrase-final word. There is some evidence that acoustic information early in a phrase may be informative of phrase ending intonation in Swedish [33]. As the ToBI standard describes these tones as realized at the end of the phrase, despite this finding, we believe examination of the phrase-final word will contain the most discriminative information for phrase-final classification. We extract the minimum, maximum, mean, and standard deviation of the pitch, as well as, the z-score of the maximum and include these in a feature vector to train and test automatic classification techniques. These features are extracted from the raw and a z-score speaker normalized pitch contour. Also included is the number of pitch points in the word. The number of pitch points can serve as a confidence measure in the approximation of the mean. Moreover, since pitch points are extracted from each 10ms frame containing periodic material, the count of pitch points provides some durational information to the classifier.

Results of ten-fold cross-validation experiments using J48 decision trees and support vector machines (SVM) with linear kernels and sequential minimal optimization (SMO) on BDC-read, BDC-spon and BURNC corpora can be found in Table 6.2. We use the weka machine learning toolkit [232] implementations of these machine learning algorithms in these experiments Results from experiments using a feature vector comprising the features extracted from the energy of the phrase-final word are shown in Table 6.3.

Corpus	Baseline	J48	SVM
BDC-read	49.00	65.88 ± 2.690	64.66 ± 1.738
BDC-spon	32.33	50.44 ± 1.837	47.89 ± 2.132
BURNC	56.16	60.51 ± 1.476	60.33 ± 0.853

Table 6.2: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using aggregations of F0.*

Corpus	Baseline	J48	SVM
BDC-read	49.00	55.29 ± 1.706	62.59 ± 2.296
BDC-spon	32.33	39.80 ± 1.492	46.75 ± 1.328
BURNC	56.16	57.65 ± 1.050	59.99 ± 0.787

Table 6.3: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using aggregations of intensity.*

These simple aggregations of pitch and intensity are able to distinguish phrase-final types with accuracy greater than chance. When we examine the features used in the decision tree trained on BDC-read data, we find that the most relevant pitch features are the number of pitch points, the minimum and maximum raw value, the z-score of the maximum value, and the speaker normalized minimum. These roughly capture the length and pitch range of the phrase-final contour. For example, words with L-L% phrase finals have on average 10.1 pitch values, where L-H% has an average of 21.4, H-L% 26.9 and H-H% 26.8. The most likely explanation for this is that the low pitch associated with L-L% may be more difficult to extract from the speech signal, leading to fewer pitch points hypothesized by the pitch tracker.

While pitch features outperform intensity features, it is notable that aggregations of intensity are also able to distinguish phrase-finals with accuracy significantly greater than chance. On the BDC-read corpus, the intensity features that are most significant are the raw and speaker normalized mean intensity along with the z-score of the maximum intensity and the speaker normalized maximum intensity value. We find that L-L% phrase finals are produced with lower intensity (average mean normalized intensity of -0.59) than L-H% (-0.059) which are, in turn, produced with lower intensity than H-L% (0.29).

These acoustic features have the advantage of being quite simple to extract from the speech signal. However, phrase-final types are also distinguished by the shape of their pitch contour. While aggregations such as mean and maximum may capture some discriminative qualities, features which represent the shape of the contour should be better able to capture the differences between different phrase-final types. The following features are extracted capture the shape of the pitch contour within the phrase-final word.

- **Slope aggregations** The basic aggregations, minimum, maximum, mean, standard deviation and z-score of the maximum, are calculated over the slope of the pitch contour.

- **Linear regression features** A linear regression is fit to the pitch contour. The slope, intercept and error of the fit line are included in the feature vector.

- **Tilt parameters** Tilt parameters [201] are calculated over the full word. (Cf. Section 5.6.3 for a detailed description of Tilt parameters).

- **Extrema location** The relative position, in terms of percentage of word length, of the minimum and maximum pitch values within the word are extracted. Also, the slopes leading into and out of the maximum pitch value are included in the feature vector.

The above features are extracted from both the raw and speaker normalized pitch contour. Results from ten-fold cross validation experiments using only these shape features can be found in Table 6.4. Corresponding results using the shape features extracted from the energy contour are presented in Table 6.5.

Corpus	Baseline	J48	SVM
BDC-read	49.00	56.94 ± 3.067	65.31 ± 1.935
BDC-spon	32.33	47.63 ± 2.509	52.33 ± 1.230
BURNC	56.16	64.58 ± 0.771	67.22 ± 1.246

Table 6.4: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using F0 shape features.*

Corpus	Baseline	J48	SVM
BDC-read	49.00	46.21 ± 1.443	52.86 ± 2.919
BDC-spon	32.33	34.41 ± 1.656	41.18 ± 1.394
BURNC	56.16	58.41 ± 1.050	62.55 ± 0.672

Table 6.5: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using intensity shape features.*

The pitch shape features are able to capture more of the differences between phrase-finals than the simple aggregations examined earlier. On BDC-spontaneous and BURNC, we observe significant increases in the performance of the J48 and SVM classifiers trained using these features. The most discriminative pitch shape features are the relative location of the maximum slope value, and the slope of a linear regression fit line.

To understand how these features relate to phrase-final type, we examine their values on BDC-spontaneous material. In H-H%, L-H% and !H-L% phrase finals the average relative location of the maximum pitch slope is in the latter half of the word – 0.578 on H-H%, 0.566 on L-H% and 0.514 on !H-L%. For other phrase finals, the maximum slope occurs in the early half of the word, 0.393 for L-L%, 0.448 for H-L%. We find an unexpected phenomenon when examining the slope of the linear regression fit line. While it makes sense that this linear approximation should be able to capture some aspect of contour shape, we find that the average slope of L-L% phrase finals is **higher** (2.66Hz/sec) than that of

L-H% (-7.39Hz/sec). Plateaus (H-L%) have a decline (-13.40Hz/sec), while H-H% phrase finals show a clear rise (208.26Hz/sec). In general, declarative contours (L-L%) are thought of as falling contours, while L-H% are rising contours. Here, we find that when we examine the entire word, that declarative contours have a tendency to rise slightly. This is quite likely within the first half of the word where other prosodic and segmental effects may impact the pitch, rather than at the boundary tone. Despite being considered a "rising" tone, L-H% accents have a declining trend. This is quite likely due to the L- phrase accent dominating the pitch movement of the H% boundary tone. Also, as mentioned previously, the tone of the nuclear pitch accent can influence the contour shape in phrase-final words. Preceding pitch accents may also contribute to this unexpected slope behavior. This is an interesting result, and an potential direction for future study.

The intensity shape features are not especially discriminative to this task. This is not altogether surprising, as phrase-finals are defined in terms of pitch contours. However, these features do perform better than chance, and examination of the most discriminative features does reveal some interesting correlates between envelope shape and phrase-final type. The most effective features are tilt coefficients, the slope of a linear regression fit line and the slope following the point of maximum intensity. All of these features are able to differentiate between plateaus (H-L%) and other phrase-final types. The energy required to maintain a constant pitch is distinct from the "falling off" associated with declarative phrase-finals. This is realized in the fit line slope, where L-L% has an average slope of -20.3, the rising phrase-finals L-H% and H-H% are similar, with average slopes of -16.0 and -29.1, respectively. Conversely, plateaus have an average slope of -1.9. There is still a slight decline in intensity over the phrase-final word; however, it is much less pronounced in H-L% phrase finals than in any other. This phenomenon is also captured by the slope leading away from intensity maxima, where the mean value of H-L% contours is approximately 20dB/sec greater than any for other contour. Examining tilt coefficients, L-L% have the most "left leaning" intensity contours, with a mean $tilt_{dur}$ value of -0.344 and a mean $tilt_{amp}$ value of

-0.299. Rising contours tilt more to the right, but still have peaks in the early half of the word; H-H% contours have a mean $tilt_{dur}$ value of -0.274 and a mean $tilt_{amp}$ value of -0.253, while L-H% contours have a $tilt_{dur}$ value of -0.180 and a mean $tilt_{amp}$ value of -0.127. In clear contrast to these, plateau contours have intensity contours with almost no tilt to them, with a mean $tilt_{dur}$ value of 0.063 and a mean $tilt_{amp}$ value of 0.082. This indicates that the envelope of H-L% phrase-finals is centered nearly at the middle of the phrase-final word, and rises approximately as much as it falls.

Lastly, we examine two voice quality features. Jitter and shimmer quantify the perceived “shakiness” of the pitch contour and corresponding intensities, respectively. These two features are commonly used to diagnose speech pathologies, but have proved to be useful in identifying emotion [62, 121, 220] and disambiguating turn-taking behavior [65]. Gravano found increases in jitter and shimmer to be associated with smooth switches between conversation partners. These switches were also correlated with the use of H-H% and L-L% phrase ending types. In our labeling experience, we have observed that some speakers exhibit “creaky voice”, or “vocal fry”, immediately preceding intonational phrase boundaries, particularly when producing L-L% phrase- final behavior. Creaky voice is characterized by low irregular pitch and energy. Jitter and shimmer have been shown to be effective at recognizing this creaky voice [96]. Thus, we expect these qualities to be helpful in distinguishing L-L% from other phrase-final types. We measure jitter by Equation 6.1; shimmer is measured identically, though replacing pitch values with intensity values. This equation is drawn from [8] by way of [62].

$$jitter = \frac{\frac{\sum_{i=1}^{N-1} |f_i - f_{i+1}|}{N-1}}{\frac{\sum_{i=1}^{N} f_i}{N}} \tag{6.1}$$

The mean values of jitter and shimmer for each phrase-final type in BDC-spontaneous can be found in Table 6.6. Using ANOVAs, we find that the differences for the four phrase ending types in jitter are significant with p=$8.3 * 10^{-5}$ and the differences in shimmer are significant with $p < 2.2 * 10^{-16}$. The BDC-spontaneous material is used for this analysis due to the more equal class distribution represented in this corpora, but similarly significant effects are observed in the BDC-read and BURNC material. We find that H-L% phrase-finals have a significantly lower jitter and shimmer than L-L% phrase finals, with the two rising types falling in between these two. This observation supports our hypothesis that L-L% should contain greater degrees of jitter and shimmer. On the other hand, the sustained pitch associated with H-L% phrase finals leads to a reduction in these voice quality features.

Type	Jitter	Shimmer
H-L%	0.0199	0.0218
!H-L%	0.0171	0.0215
L-H%	0.0259	0.0246
H-H%	0.0278	0.0235
L-L%	0.0298	0.0269

Table 6.6: *Mean values of jitter and shimmer for each phrase-final type in the BDC-spontaneous corpus.*

The results of ten-fold cross validation experiments using only jitter and shimmer to classify phrase finals can be found in Table 6.7. While these features are hardly the most discriminative, they do capture acoustic qualities that are entirely distinct from the pitch and intensity shape and range features captured by the previous features. This distinctness and the success of these features when used in isolation suggests that they might be fruitfully combined to improve phrase-final classification performance.

Corpus	Baseline	J48	SVM
BDC-read	49.00	55.00 ± 2.394	49.00 ± 2.263
BDC-spon	32.33	46.06 ± 1.263	37.63 ± 1.312
BURNC	56.16	57.24 ± 0.869	56.16 ± 0.918

Table 6.7: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using jitter and shimmer.*

Next, we combine the acoustic features sets we have explored separately . First, we combine each pitch-based feature set with its intensity-based counterpart. Second, we collapse all of the pitch-based feature sets and all of the intensity-based feature sets. Third, we experiment with a feature vector containing all of the acoustic features described above.

The results of ten-fold cross-validation experiments using these extended feature vectors run on BDC-read, BDC-spontaneous and BURNC are presented in Tables 6.8, 6.9, and 6.10, respectively.

Feature Set	J48	SVM
pitch aggregate	65.88 ± 2.690	64.66 ± 1.738
intensity aggregate	55.29 ± 1.706	62.59 ± 2.296
all aggregate features	63.23 ± 2.837	65.59 ± 1.525
pitch shape	56.94 ± 3.067	65.31 ± 1.935
intensity shape	46.21 ± 1.443	52.86 ± 2.919
all shape features	57.94 ± 2.837	66.81 ± 3.034
agg+shape	65.45 ± 1.771	***70.10*** ± 2.345
jitter	55.01 ± 2.394	49.00 ± 2.263
agg+shape+jitter	62.02 ± 1.607	69.81 ± 2.034

Table 6.8: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using acoustic features on BDC-read. Baseline accuracy=49.0%*

Feature Set	J48	SVM
pitch aggregate	50.44 ± 1.837	47.89 ± 2.132
intensity aggregate	39.80 ± 1.492	46.75 ± 1.328
all aggregate features	47.90 ± 1.886	50.76 ± 1.755
pitch aggregate	47.63 ± 2.509	52.33 ± 1.230
intensity aggregate	34.41 ± 1.656	41.18 ± 1.394
all shape features	46.98 ± 2.083	53.29 ± 1.984
all aggregate features	50.16 ± 2.034	56.20 ± 1.935
jitter	46.06 ± 1.263	37.63 ± 1.312
agg+shape+jitter	51.17 ± 2.181	***57.12*** ± 1.410

Table 6.9: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using acoustic features on BDC-spontaneous. Baseline accuracy=32.3%*

We find, first of all, that the J48 algorithm is less suited to combining these feature sets than SVM classification with a linear kernel. On all corpora, we observe that the combination of pitch and intensity features leads to improved performance under SVM classification, though this improvement is not always statistically significant. The best performance on all corpora comes from the inclusion of both aggregation and shape features extracted over the pitch and intensity contour. On the BDC-read and BURNC corpora, the

Feature Set	J48	SVM
pitch aggregate	60.51 ± 1.476	60.33 ± 0.853
intensity aggregate	57.65 ± 1.050	59.99 ± 0.787
all aggregate features	61.24 ± 1.000	65.40 ± 0.968
pitch shape	64.58 ± 0.771	67.22 ± 1.246
intensity shape	58.41 ± 1.050	62.55 ± 0.672
all shape features	64.38 ± 0.804	68.29 ± 1.345
agg+shape	66.96 ± 0.935	72.94 ± 0.754
jitter	57.24 ± 0.869	56.16 ± 0.918
agg+shape+jitter	66.32 ± 0.968	***73.34*** ± 0.836

Table 6.10: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using acoustic features on BURNC. Baseline accuracy=56.2%*

inclusion of jitter and shimmer to the feature vector leads to improvements, but these are not statistically significant, while on BDC-spontaneous, the inclusion of these features *reduces* performance. This suggests that much, if not all, of the information captured by these voice quality features is already captured by the aggregation and shape features.

Using J48 classification, on the two BDC subcorpora, despite the addition of shape and jitter features, performance does not significantly improve over the initial performance demonstrated using only pitch aggregation features. Perhaps due to the availability of more training data, on the BURNC data, the additional features improve classification accuracy from this pitch-based baseline, but only by 4.47%. Moreover, the inclusion of intensity aggregation or shape features lowered the J48 classification accuracy. These results highlight the success of SVM classification to classify phrase-final tones using shape features – the source of greatest difference between the two classification algorithms.

To evaluate the speaker-dependence of these acoustic features, we evaluate the best performing feature sets using leave-one-speaker-out cross-validation and SVM classification. The results of these experiments can be found in Table 6.11. We find that on the BDC material, there is a strong speaker dependency. The absence of any training data for a speaker results in approximately a 10% reduction in classification accuracy. Naturally, this is a troubling finding. To achieve high accuracy performance on an unseen speaker,

Corpus	Feature Set	10-fold	Leave-one-speaker-out
BDC-read	agg+shape	70.10 ± 2.345	62.02 ± 4.805
BDC-spon	agg+shape+jitter	57.12 ± 1.410	47.81 ± 6.757
BURNC	agg+shape+jitter	73.34 ± 0.836	70.70 ± 2.362

Table 6.11: *Leave-one-speaker-out evaluation of SVM classification using acoustic features.*

novel representations or normalizations of contour shape will be required. The acoustic features evaluated in this section to not easily generalize to unseen speakers. The z-score normalization of pitch does not appear to be a successful technique to account for speaker differences. It is possible that a normalization that is more specific to this task is required. Such a procedure will account for the parameters that are typically varied in the production of phrase ending intonation, including pitch range, typical pitch contour shapes and timing – the relationship between the realization of phrase ending intonation and the phrase boundary. Development of more robust normalization procedure remains a source of future research.

Of note is that there is a comparatively degradation of performance on the BURNC material. There are at least two possible explanations for this. First, the broadcast news (BN) speaking style may be so well conventionalized [24], that classification of BN phrase-finals is more robust to speaker differences than non-professional speech or there may be fewer speaker differences in the use and production of phrase ending intonation. Second, the BURNC contains six speakers while the BDC material contains only four. It is possible that the availability of more training data from more speakers produces more robust models when trained on BURNC.

6.5.2 Syntactic features

In this Section, we examine the role of syntax on phrase-final behavior. Declarative (L-L%) and high-rise (H-H%) contours are considered be more indicative of finality than !H-L%, plateaus (H-L%) and continuation rise (L-H%) contours. Assuming that continuity and finality are conveyed through syntax as well as acoustics, access to syntactic information should be able to contribute to the classification of phrase accent types. In Section 6.5.2 we will examine the use of structural features extracted from automatically generated sparse trees, while in Section 6.5.2, we will investigate the use of part-of-speech (POS) tag information in the classification of phrase finals.

Parse tree features

In this Section, we explore the use of features extracted from hypothesized parse trees in the classification of phrase-finals. To automatically generate parse trees, we use an implementation of Charniak's "maximum entropy inspired" parser [36] trained on Switchboard data [64]. The Switchboard corpus is comprised of spontaneous telephone conversations. This is somewhat different in genre from the direction-giving monologues of the BDC corpus and the BN of the BURNC, but they are all spoken material. Another option might be to use a parser trained on Penn Treebank data [127]; the differences between Wall Street Journal articles and spoken data are, however, likely to be much greater than those between speech genres. From hypothesized parse trees we extract the following features. In these experiments, the lexical identity of the words in each parsed utterance are manually identified. Note that successful parsing of automatically recognized text can be quite difficult. Thus, the use of syntactic features in automatically classifying phrase-final types represents an idealized scenario where accurate lexical identity information is available. The analysis of the features, however, remains informative to the understanding of how speakers use phrase-final intonation in different syntactic contexts, and to the use of syntactic information in automatic classification systems. The results of these classification experiments provide an expected ceiling to the performance of the presented approaches when operating on automatically recognized material.

From the extracted parse trees we obtain the following features.

- **Constituent identity** We include nominal features representing the identity of the smallest constituent containing the current word. The inventory of constituent tags are defined in the Switchboard [64] and Penn Treebank [127] corpora.

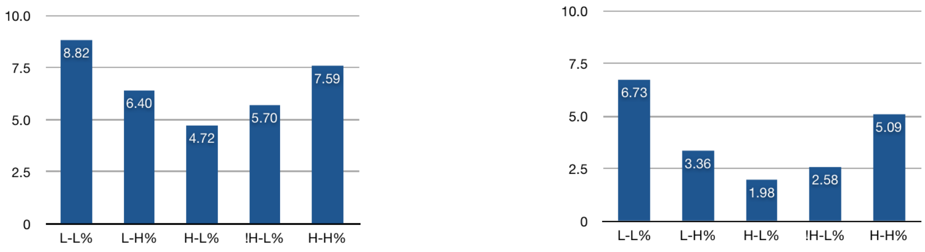
- **Positional features** We extract the absolute and relative position of the current word within its constituent, calculated from the start of the constituent.
- **Syntactic disjuncture** To measure the degree of syntactic disjuncture between the current, w_i, and following word, w_{i+1}, we calculate the distance in the parse tree between the parse nodes corresponding to w_i and w_{i+1} . In addition to this, to normalize this measure for the overall syntactic complexity of the current sentential unit, we also calculate the ratio of this measure to the depth of their closest ancestor. This captures the ratio of the cardinality of the disjuncture of ancestor nodes of the current and following words to the cardinality of the union. An graphical example of this normalization is presented in Figure 4.3, and discussed in Section 4.4.

Using descriptive statistics we are able to draw some conclusions about the relationship between phrase-final types and syntactic structure that can inform how phrase-finals are used. These analyses are performed on the BDC-spontaneous material. All of the numeric parse-tree features – the positional and syntactic disjuncture features – significantly differ with respect to phrase-final type. ANOVA on each feature rejects the null hypothesis with $p < 0.0001$. We find that H-L% phrase finals tend to occur at a shallower parse depth (mean 9.71) than L-H% (10.77) and !H-L% (10.24) which all occur at shallower depths than the L-L% (11.59) and H-H% (11.66). This feature is, most likely, more useful for the prediction of phrase-final types on some genres rather than other. Since each utterance in the BDC-spontaneoous corpus is used in a task-directed monologue, they share a similar degree of syntactic complexity. There is no expectation that the values of these numbers should remain informative when analyzing speech in other genres or domains. However, this implies that, at least within this corpus, H-L% phrase-finals are used either within less nested syntactic structures.

Examination of the absolute constituent position of the phrase final indicates that L-H%

(1.14), L-L% (0.95), !H-L% (1.07), and H-H% (0.98) phrase finals occur later in syntactic constituents than H-L% (0.69). This may indicate that most phrase boundaries more likely to fall at constituent boundaries, with H-L% more likely to occur within syntactic constituents. However, this may be an effect of these phrase finals falling within constituents of increased length. By examining the position from the end of the containing constituent, we can confirm that H-L% is less likely to fall at the end of a syntactic constituent than other phrase finals. H-L% phrase finals occur on average 0.75 words from the end of its containing syntactic constituent, while L-L% fall 0.39 words from the end, H-H% 0.36 words, !H-L% 0.52 and L-H% 0.31. ANOVA reveals a significant difference in distance from constituent end point with $p < 2.2 * 10^{-16}$. Similar relationships can be observed by examining the relative positions from the start and end of syntactic constituents.

The syntactic disjuncture measures are based on the intuition that the graph distance between two parse tree nodes can be used to quantify the degree of structural disjuncture between the two corresponding words. When two words are members of the same syntactic constituent, there is a minimal amount of disjuncture at the boundary between them. However, if there is a constituent boundary between the two words, the graph distance between them is increased, and continues to increase with the number of nested constituents that are either completed by the word prior to the boundary or started by the word following the boundary. Figure 6.12 contains the mean syntactic distance and syntactic distance ratio for each phrase-final type from the BDC-read corpus. When we examine the relationship



(a) Syntactic distance

(b) Syntactic distance ratio

Figure 6.12: *Syntactic distance measures calculated over BDC-read material.*

between syntactic disjuncture measures and phrase-final type, we again find evidence that L-L% is the phrase-final type used at boundaries with greater disjuncture with regard to syntactic structure. Moreover, H-H% tokens are used at greater boundaries than L-H% and H-L%. These measures clearly demonstrate that there is a relationship between syntactic structure and phrase-final type.

We also examine two nominal syntactic parse-tree based attributes and their relationship to the five phrase-final types. We identify the smallest constituent larger than POS category containing each word. This allows us to determine if certain phrase-finals are more likely to fall within, say, a verb phrase, or a prepositional phrase. Also, we identify the smallest syntactic constituent containing both words surrounding the phrase boundary associated with each phrase-final. This will determine if there is a difference in phrase-final usage within, for example, dependent or independent clauses.

Figure 6.13 contains a chart of the distributions of constituent identities broken down by phrase-final type. A χ^2 analysis reveals that these differences are significant with $p < 2.2 * 10^{-16}$. The point of clearest difference is the use of phrase-final types in Noun

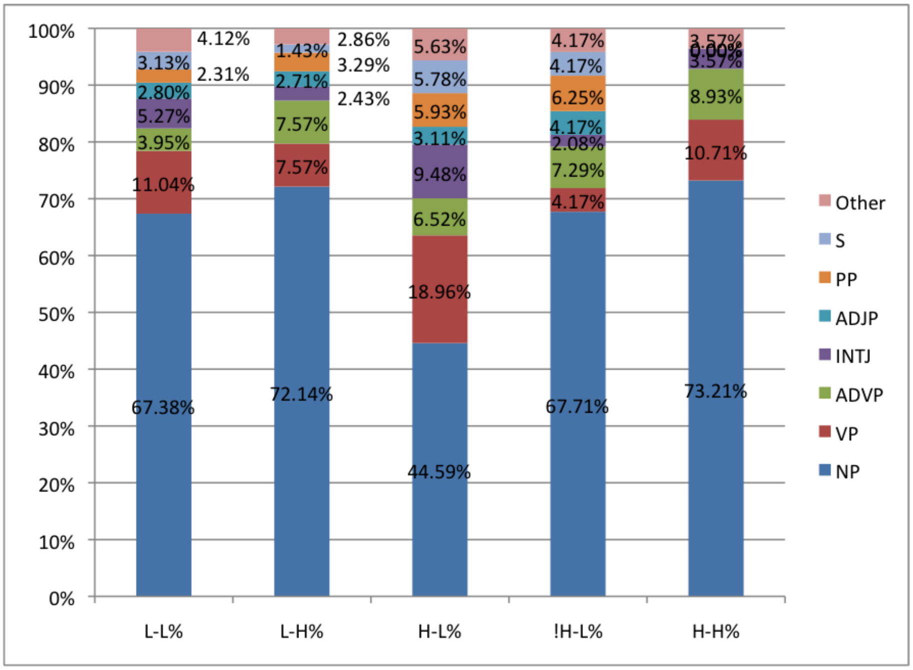


Figure 6.13: *Constituent identity distributions by phrase-final type from BDC-spontaneous.*

Phrases (NP). While approximately 70% of L-L%, L-H%, !H-L% and H-H% phrase-finals fall within NPs, only 44.59% of H-L% phrase-finals do. These plateaus (H-L%) are used more often in Verb Phrases (VP), Interjections (INTJ) and Prepositional Phrases (PP).

These differences are not great enough enough to draw clear conclusions about the relationship between syntax and phrase-final intonation. However, if we interpret them in the context of earlier analyses, we can draw some broad conclusions. The most distinctly used phrase-final is the plateau contour, H-L%. These contours are more likely to be used within a syntactic constituent as opposed to at the end of one. Regardless of their position within a constituent, when a phrase boundary occurs between two words that are syntactically “close”, as defined by parse-tree based graph distance, the more likely the phrase-final intonation is to be a plateau. Finally, they are less likely than others to be used in NPs. Particularly striking are the differences between the !H-L% and the non-downstepped H-L% plateau. Despite having many commonalities in their acoustic form, the downstepped plateau has more in common with the L-H% phrase-final tone combination than the H-L%. The three of these however, all share some commonalities in comparison to L-L% and H-H% They all tend to fall later in a syntactic constituent, and are used at phrase boundaries between words that are syntactically close. However, while H-L% phrase-finals often occur within a syntactic constituent, L-H% and !H-L% phrase finals are used most often at the end of a constituent. The two most similar phrase-final types, from a syntactic point of view, are L-L%, the declarative contour, and H-H%, the high-rising contour. They fall at the end of short constituents and the constituents they are used in have a similar distribution. The largest point of difference in the syntactic qualities we examined is that L-L% are used at points of somewhat greater syntactic disjuncture than H-H%.

Results of ten-fold cross-validation experiments using all of these parse tree features can be found in Table 6.12. Note that in the BURNC the same stories are repeated by multiple speakers. Therefore, when constructing the cross-validation folds, we must ensure that, if a given story occurs in a training fold, it does not occur in the training set, even if produced by a different speaker. In Table 6.13, we repeat the evaluation using leave-one-speaker-out validation to isolate any speaker differences. Due to the repetition of lexical material by multiple speakers, this evaluation is not performed on the the BURNC material.

Corpus	Baseline	J48	SVM
BDC-read	49.00	66.31 ± 2.132	67.53 ± 2.099
BDC-spon	32.33	46.84 ± 1.673	48.09 ± 1.312
BURNC	56.16	58.59 ± 1.214	59.57 ± 1.082

Table 6.12: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using parse-tree based features.*

Corpus	Baseline	J48	SVM
BDC-read	49.00	62.66 ± 7.446	66.31 ± 8.889
BDC-spon	32.33	43.07 ± 4.330	46.11 ± 8.652

Table 6.13: *Leave-one-speaker-out cross-validation accuracy (%) of phrase-final type classification using parse-tree based features.*

Overall, we find that these parse features demonstrate significant improvement over the majority class baseline. While they perform significantly worse than the acoustic features explored in Section 6.5.1, these results indicate that syntactic structure-based features can be informative for phrase-final type classification. Moreover, while the speaker-independent experiments yield lower results, on average, than the cross-validation experiments, these differences are not significant. Of course, this does not provide enough information to say with certainty whether any of these syntactic features are speaker-dependent or not. However, these results are encouraging, and suggest that more study should be done to determine if they can be replicated on automatically recognized speech as opposed to manual transcriptions.

We then combine the syntactic parse tree-based features with the full set of acoustic features we examined in Section 6.5.1 to see if the two feature sets can combine to improve phrase-final classification performance. The results of these ten-fold and leave-one-speaker-out cross validation experiments can be found in Tables 6.14 and 6.15. Despite modest classification performance in isolation, the structural syntactic features investigated in this

Corpus	Baseline	J48	SVM
BDC-read	49.0	65.16 ± 1.574	71.60 ± 1.443
BDC-spon	32.33	50.48 ± 1.935	58.31 ± 1.935
BURNC	56.16	65.51 ± 0.820	73.50 ± 1.15

Table 6.14: *Ten-fold cross-validation accuracy (%) of phrase-final type classification using parse tree and acoustic features.*

Corpus	Baseline	J48	SVM
BDC-read	49.0	56.15 ± 4.264	61.52 ± 5.199
BDC-spon	32.33	44.04 ± 3.723	49.42 ± 7.101

Table 6.15: *Leave-one-speaker-out cross-validation accuracy (%) of phrase-final type classification using parse tree and acoustic features.*

section do not significantly improve phrase-final classification performance in the presence of acoustic information. Recall that on BDC-read the SVM accuracy with acoustic features alone is 70.10%, on BDC-spontaneous 57.12% and on BURNC 73.34%. These correspond to improvements between 0.16% and 1.50%, none of them statistically significant.

The experiments and descriptive statistics described in this section show that there is an interaction between syntactic structure and phrase-final type. However, these features are not able to improve classification accuracy in combination with acoustic features. It is unlikely that the syntactic information captured by the features explored in this section is represented in the acoustic features investigated in Section 6.5.1. A more probably explanation is that acoustic cues are more reliable indicators of phrase-final type than these syntactic cues. Despite this, there **is** an application of these syntactic results in prosodic assignment. When performing text-to-speech synthesis, a prosodic assignment module analyzes the text and assigns prosodic targets for the synthesized speech. The syntactic features explored here have demonstrated some reliability in the assignment of phrase-final types based only on text analysis. More sophisticated analyses of semantic and pragmatic information, discourse structure and coherence should be able to further improve this text-based assignment of phrase-final types. If this analysis is sufficiently improved, we expect the combination with acoustic information to lead to improved classification performance.

Part-of-Speech tags

In this Section, we examine the interaction between part-of-speech tag information drawn from words near intonational phrase boundaries, and the coocurring phrase-final types. In Section 6.5.2, we observe that parse tree based syntactic features were able to differentiate phrase-final types with accuracy greater than chance. Here, instead of querying constituent and structural information, we will investigate the correlations between sequences of part-of-speech tags and phrase-final types.

To extract part-of-speech (POS) tag information we use an implementation of the Stanford POS tagger [213] that is run on the manual transcriptions included in the BDC and BURNC corpora. We train multinomial models to capture the relationship between POS tag information and phrase-final types. These are used to classify phrase-final type using POS tag information following Equation 6.2, where POS is a part-of-speech tag based nominal variable.

$$type^{*} = \underset{type}{\operatorname{argmax}}\, p(type|POS) \tag{6.2}$$

This classification decision method is evaluated using the POS tag of the word immediately preceding the phrase boundary (**unigram**), the bigram (**bigram**) and trigram (**trigram**) immediately preceding the boundary, and the tags of the words immediately surrounding the boundary (**surrounding**). In addition to examining the raw tag set, we investigate two clusterings of the full, Penn Treebank [127] derived, POS tag set. The first clustering, which we refer to as Broad Class, contains six elements: *NOUN*, *VERB*, *ADJECTIVE*, *ADVERB*, *CARDINAL* and *FUNCTION*. The second contains two POS tag classes, *FUNCTION* and *CONTENT*.

Ten-fold cross-validated evaluations of this POS-tag based classification approach performed on BURNC and BDC corpora can be found in Table 6.16. Due to the repetition of lexical material, the BURNC fold assignment guarantees that no story appears simultaneously in a training and test fold, for accurate evaluation.

POS Feature	BDC-read	BDC-spon	BURNC
Full POS set			
unigram	47.28 ± 1.870	38.37 ± 1.328	56.16 ± 1.148
bigram	46.28 ± 1.722	37.03 ± 1.361	57.13 ± 1.197
trigram	47.57 ± 2.575	33.44 ± 0.919	58.44 ± 0.968
surrounding	***54.36 ± 1.919***	38.42 ± 1.099	***60.60 ± 1.050***
Broad Class			
unigram	40.00 ± 2.526	39.11 ± 1.246	55.16 ± 0.869
bigram	47.71 ± 2.280	40.35 ± 1.837	55.70 ± 0.902
trigram	45.71 ± 2.130	37.26 ± 2.034	55.97 ± 1.132
surrounding	48.28 ± 2.362	***43.30 ± 1.919***	58.25 ± 0.787
Function Content			
unigram	49.00 ± 2.05	37.72 ± 2.099	56.16 ± 0.902
bigram	48.86 ± 1.591	38.42 ± 1.935	56.16 ± 0.722
trigram	49.00 ± 1.935	37.91 ± 1.624	56.18 ± 1.066
surrounding	48.71 ± 2.772	41.09 ± 1.427	55.98 ± 1.132
Baseline	49.00	32.33	56.16

Table 6.16: *Ten-fold cross-validation accuracy using multinomial classification and a single POS-tag feature.*

The improvement of the best performing part of speech features (in bold) is significant or approaches statistical significance on all corpora: on BDC-read p=0.06612, BDC-spon, p=0.00619, and BURNC p=0.01703. While this does represent improvement, in comparison to the performance of acoustic features or parse-tree based syntactic features, this classification accuracy is quite poor. The collapsing of part-of-speech tag classes does not, in general, improve the classification performance. In fact, on BDC-read and BURNC, the accuracy is significantly worse when POS tags are collapsed to six or two classes. This suggests that the interaction between phrase-final type and syntactic tag is an effect of some particular commonly used constructions, rather than major syntactic effects. The one exception to this is that, on the BDC-spontaneous material, performance can be improved by collapsing to **Broad Classes**. On the BDC subcorpora, where speaker independent evaluation is possible, we do not observe a significant effect of speaker dependency on the best performing features. On BDC-read, the surrounding POS feature yields accuracy of 48.78% ± 4.674, while

5.58% lower than the ten-fold result, a t-test determines this difference to only approach significance with p=0.103. However, on BDC-spontaneous, the surrounding **Broad Class** POS feature achieves accuracy of 42.38% ± 7.1668, 0.92% and not significantly lower than the ten-fold cross-validation result.

While the performance of the POS features is not particularly outstanding in comparison to the other feature sets explored in this chapter. However, POS features capture distinct information from the parse tree based features examined in Section 6.5.2. Moreover, they may contribute to improved classification performance in combination with acoustic features. We evaluate the impact of including those POS tag features which perform better than baseline in the acoustic SVM classifier feature vector presented in Section 6.5.1. Ten-fold cross-validation accuracy of these experiments can be found in Table 6.17; leave-one-speaker-out evaluation results can be found in Table 6.18;

Corpus	Acoustic Only	Acoustic + POS
BDC-read	70.10 ± 2.345	65.95 ± 1.689
BDC-spon	57.12 ± 1.410	51.22 ± 1.935
BURNC	73.34 ± 0.836	73.43 ± 1.689

Table 6.17: *Ten-fold cross-validation accuracy including discriminative POS tag features in the SVM acoustic feature vector.*

Corpus	Acoustic Only	Acoustic + POS
BDC-read	62.02 ± 4.805	59.16 ± 5.740
BDC-spon	47.81 ± 6.757	42.42 ± 5.691

Table 6.18: *Leave-one-speaker-out cross-validation accuracy including discriminative POS tag features in the SVM acoustic feature vector.*

The inclusion of part of speech tag information fails to significantly improve phrase-final classification accuracy in the presence of discriminative acoustic information. When we examine the effect of these features in isolation, we find a minor interaction between syntactic information and phrase-final type. This effect is observed in the investigation of part-of-speech tags and structural parse tree features. Regardless of the syntactic information extracted, the impact on phrase-final type classification is minor. This suggests that we reexamine the motivation behind the investigation of the interaction between syntactic information and phrase-final type.

Phrase-final types are used to communicate a desired semantic, pragmatic or discourse effect to a listener. It has been hypothesized that high phrase accents (H-) are used to indicate an semantic incompleteness of the preceding phrase. In the case of the plateau (H-L%), incompleteness is used by a speaker to hold the turn [65] and to provide additional relevant semantic content in a subsequent phrase. The high rise (H-H%) is commonly associated with yes-no questions, where the semantic completeness is expected to come from a conversation participant. The high rise contour can also be used to request verification from a listener, without the presence of an explicit yes-no question. These hypothesized and observed effects of phrase-final type on communicative expectations are not strictly syntactic in nature. They are more commonly discussed and understood as semantic and pragmatic effects. The syntactic qualities of the component lexical content are only informative insofar as the semantic and pragmatic effects are realized in distinct syntactic structures. In this Section, we find only scant evidence that this is the case – the differences in phrase-final type are not reflected in clear differences in part-of-speech content immediately surrounding a phrase boundary. In Section 6.5.2, however, we were able to find a realization of the hypothesized semantic effects in syntactic structure. In particular, we consistently found evidence that the plateau phrase-final contour (H-L%) is used more commonly at points of greater syntactic incompleteness than other phrase-finals. This is consistent with the hypotheses that the plateau is used to hold the turn, and indicate semantic incompleteness, or a "forward-looking resolution" of some communicative quality, whether syntactic, semantic, pragmatic or structural. It is likely that lexical analysis will be able to contribute to the generation of phrase-final type hypothesis, however, this analysis will require more sophisticated natural language understanding techniques than the syntactic analyses evaluated in this section. Based on hypothesis and observations about the use of phrase finals, it is likely that semantic, pragmatic, discourse structure, or coherence information should be effective to this end.

6.5.3 Regions of Analysis

The ToBI standard defines the phrase accent as describing the pitch contour between the final pitch accent in a phrase and the phrase boundary. If the pitch is flat or rising, the ToBI standard describes the phrase accent as having a high tone using the "H-" tag, while L- is used to indicate falling pitch. Boundary tones describe the pitch contour shape immediately preceding intonational phrase boundaries.

In the experiments described in Section 6.5.1, we extracted acoustic features over the full duration of the phrase-final word. However, the annotation standard indicates that phrase-final behavior, particularly the boundary tone, occurs immediately prior an intonational phrase boundary. Feature extraction from the full word may include acoustic material that is outside the phrase-final phenomena. To assess this possibility, and identify a more useful region of analysis we repeat the experiments from Section 6.5.1 while extracting features from a different set of regions prior to the phrase boundary. The candidate regions of analysis are as follows.

- **Full Word** The full phrase-final word. This corresponds to the analysis performed in Section 6.5.1.

- **Last half of a word** Features are extracted from only the second half of the phrase-final word.

- **Last 200ms** The average length of a word in BDC-read is 391ms, in BDC-spontaneous 440ms and BURNC 514ms. To approximate the region corresponding to the latter half of a phase-final word without requiring word identity/boundary information, we extract features from the 200ms immediately preceding a phrase boundary.

- **Energy Peak** Using the energy-peak identification technique defined in Section 5.6.3, we extract features from the final energy peak in the word.

- **pseudo-syllable** Using the envelope-based pseudo-syllabification technique defined in [219], we extract features from the final pseudo-syllable in the word. This technique is a more sophisticated approach than the energy-peak regions. It uses amplitude onset velocity and spectral stabilities to determine whether a syllable boundary is present.

- **Last accent** Annotated accent locations are time-aligned. Using these oracular accent locations, we extract acoustic features from the region between the last pitch accent in a phrase and the phrase boundary. This region of analysis begins at the point in time where the accent annotation was placed. This feature follows the definition given by the ToBI standard, though it requires the presence of high-accuracy pitch accent location information.

- **Syllable** Using the syllabification available for the BURNC material, we extract features from the phrase-final syllable. This syllabification technique is more lexically grounded than the energy-peak or pseudo-syllabification approaches. As described in Section 2.3, this syllabification is based on the output of forced alignment and subsequent coordination with a lexicon containing syllable boundaries. Both the forced alignment and syllable alignment may introduce errors. Without manual or force-aligned syllable boundaries, this region of analysis is unavailable for BDC material.

- **Last “accentable” syllable** Some researchers have considered the phrase accent to describe the pitch between the last *accentable* syllable, rather than the last *accented* syllable, where any syllable with primary lexical stress is considered “accentable”. Again, using the BURNC syllabification and lexicon-based lexical stress assignment, we are able to isolate this region. This region is unavailable on BDC material, due to the lack of annotated syllable or phone boundaries.

We report performance of aggregate features, shape features, and voice quality features over each of the available regions of analysis. Since SVM classification generated superior performance in Section 6.5.1, this is the only classification technique used in these analyses. The results of ten-fold cross-validation on BDC-read, BDC-spontaneous, and BURNC can be found in Tables 6.19, 6.20 and 6.21.

Feature Set	Word	Last 400ms	Last Half	Last 200ms	Energy Peak	Pseudo-syllable	Last Accent
Aggregations	65.59±1.522	65.38±1.361	66.17±2.263	64.31±1.197	58.30±1.870	63.88±2.345	66.09±2.148
Shape	66.81±3.034	57.30±2.329	62.23±1.853	62.59±1.853	53.36±1.624	55.44±3.100	57.87±1.870
Voice **Q**uality	49.00±1.722	49.00±2.952	49.00±2.394	49.00±2.083	49.00±1.837	49.00±1.738	49.00±2.329
A+S+VQ	69.81±2.034	71.10±1.853	72.32±2.493	71.46±1.886	64.16±1.837	68.31±2.066	70.96±1.574

Table 6.19: *Ten-fold cross-validation accuracy (%) of SVM phrase-final type classification of BDC-read material using acoustic features extracted from a range of analysis regionsn. Baseline=49.0%*

Feature Set	Word	Last 400ms	Last Half	Last 200ms	Energy Peak	Pseudo-syllable	Last Accent
Aggregations	50.76±1.755	49.06±1.738	51.27±1.706	50.85±1.607	39.71±2.099	46.98±2.394	50.25±2.210
Shape	53.29±1.984	46.71±0.984	47.12±1.230	46.20±1.394	40.58±1.378	43.39±1.443	43.94±2.558
Voice **Q**uality	37.63±1.312	37.68±1.378	37.67±1.788	38.23±2.083	37.63±2.148	38.42±1.099	38.65±1.952
A+S+VQ	57.12±1.410	57.44±1.738	55.50±1.689	54.95±2.444	50.71±1.624	55.18±1.345	56.75±1.935

Table 6.20: *Ten-fold cross-validation accuracy (%) of SVM phrase-final type classification of BDC-spontaneous material using acoustic features extracted from a range of analysis regions. Baseline=32.33%*

Across all corpora, we find that the two pseudo-syllabification techniques – the energy peak approach defined in Section 5.6.3 and that defined by Villing and colleagues in [219] – perform worse than the **Full Word** region of analysis. The only exception to this finding is that the Villing approach yields slightly improved performance on BURNC, but this improvement is not statistically significant. However, when we have access to syllabification information from forced alignment and a lexicon, extracting acoustic features from the

Feature Set	Word	Last 400ms	Last Half	Last 200ms	Energy Peak	Pseudo-syllable	Last Accent	Syllable	Last Accentable
Aggregations	65.40±0.968	66.66±0.476	66.30±0.672	65.62±0.525	56.16±1.640	65.57±0.984	65.62±1.214	68.37±0.886	68.29±0.754
Shape	68.29±1.345	64.02±1.312	68.94±1.427	70.72±1.000	60.85±1.033	64.78±0.984	63.74±0.738	71.67±1.099	71.92±1.000
Voice **Q**uality	56.16±0.918	56.16±1.050	56.16±1.181	56.16±0.738	56.16±1.050	56.16±1.542	56.16±0.869	56.16±1.099	56.16±0.984
A+S+VQ	73.34±0.836	73.03±0.853	76.22±0.804	76.37±1.312	67.86±1.164	73.89±1.017	73.35±0.787	77.10±0.853	76.62±0.886

Table 6.21: *Ten-fold cross-validation accuracy (%) of SVM phrase-final type classification of BURNC material using acoustic features extracted from a range of analysis regions. Baseline=56.2%*

phrase-final syllable yields the best phrase-final classification performance. This finding suggests that improvements to the acoustic pseudo-syllabification techniques may be able to improve identification of a helpful region of analysis, even when phone alignment information is unavailable.

Modeling the region from the last accent performs approximately as well as extracting features from the whole word. On the BDC material performance is better, but not significantly so; on BURNC performance is worse. While the contour shape between the last accent and the phrase boundary may be indicative of phrase-final type, this result suggests that the part of the contour lying outside a phrase-final word does not contain additional discriminative information. We find aggregations drawn over the latter half of the word to be consistently more discriminative than those extracted from the full word. The shape modeling features do not demonstrate the same consistency – on BDC-spontaneous, they perform worse.

One of the more striking results from these experiments is that extracting features from the region starting 200ms prior to the phrase boundary performs better than examining the acoustics of the full word on both BDC-read and BURNC. The performance of features taken from the 200ms preceding the phrase boundary is not significantly worse than those extracted from the full word on BDC-spontaneous. This is a very encouraging finding. It suggests that phrase-final classification can be performed without additional external annotation; word identity/boundary and pitch accent location information does not significantly improve performance. The most powerful indicator of phrase boundary location is the presence of silence. Silence detection is a *relatively* simple task, in comparison to the recognition required to generate word or syllable boundaries, or to hypothesize pitch accent location – the information required to extract acoustic features from the other best performing regions of analysis. These findings suggest that phrase-final types can be classified by examining acoustic material within the 200ms immediately prior to a silent region, and performance does not dramatically suffer because of it.

To measure the robustness of these regions to speaker differences, we perform leave-one-speaker-out cross-validation using the full feature sets for each corpus and region of analysis. The results of these evaluations can be found in Table 6.22. Across all corpora and regions, speaker independent classification performs approximately 6% worse than ten-fold cross-validation. Similar differences in region-of-analysis observed in the ten-fold cross-validation evaluation are observed in these speaker independent experiments. On BDC-read and BURNC – both corpora of read speech – we find that by extracting acoustic features from the **Last Half**, **Last 200ms** and **Syllable** regions can significantly improve phrase-final classification in a speaker independent scenario. No region of analysis performs significantly better than word-level analysis on the BDC-spontaneous corpus. On this material, the **Last Half** region generates the highest performance. Given its performance

Corpus	Word	Last 400ms	Last Half	Last 200ms	Energy Peak	Pseudo-syllable	Last Accent	Syllable	Last Accentable
BDC-read	62.02±4.805	61.66±4.740	64.59±4.297	64.23±4.986	59.16±6.347	61.02±5.838	61.73±4.740	NA	NA
BDC-spon	47.81±6.757	46.02±6.954	49.52±3.887	47.49±4.690	43.34±4.723	47.03±5.560	46.11±6.954	NA	NA
BURNC	70.70±2.361	70.36±2.739	74.29±2.788	74.00±3.165	64.75±3.378	70.12±2.526	70.36±2.739	75.08±2.772	74.54±3.542

Table 6.22: *Leave-one-speaker-out cross-validation accuracy (%) of SVM phrase-final type classification using acoustic features extracted from a range of analysis regions.*

across corpora and the minimal requirements of additional information which may introduce noise, the **Last 200ms** region is likely to be the best option for classification of phrase-finals. If available, however, forced-alignment syllable boundaries perform better. The robustness of this region to non-professional speech has not been evaluated here. Thus this result should be considered preliminary until further confirmation on other speech genres.

Examination of feature sets extracted from different regions of analysis, reveals that there are differences in their performance. The use of these different feature sets for classification, however, are not mutually exclusive; a classifier training algorithm can take advantage of features drawn from multiple regions of analysis simultaneously. To evaluate the impact of combining these feature sets, we include features drawn from multiple regions of analysis in the feature vector. Three regions of analysis do not require lexical information for identification; these are **Last 200ms**, **Energy Peak** and **pseudo-syllable**. If we have access to lexical identity information, we are able to extract features from the **Word**, **Last Half** and **Syllable** regions. Finally, with access to accent location information, we can use the **Last Accent** region of analysis. Ten-fold cross-validation evaluation of SVM classification using features extracted with access to these additional data can be found in Table 6.23. These feature vectors include the aggregation, shape and voice quality features described in Section 6.5.1, extracted from regions that can be identified using available additional information – word identity/boundary or accent location.

Available Information	BDC-read	BDC-spon	BURNC
None	72.53±1.296	56.47±1.427	76.87±0.918
+Word Identity	72.10±1.919	57.02±2.050	77.80±0.672
+Accent Location	72.10±2.001	57.90±1.296	77.50±1.000

Table 6.23: *Ten-fold cross-validation accuracy of SVM phrase-final type classification combining analysis region feature sets.*

From this experiment we see that on the BDC material, the inclusion of word identity and accent information does not provide improved phrase-final classification accuracy. Moreover, on all corpora, the combination of **Last 200ms**, **Energy Peak** and **pseudo-**

syllable feature sets increases the classification accuracy, but not to a statistically significant degree. On the BURNC material, the inclusion of the **Syllable** feature set, and others that require word identity information, increases performance by a degree that is not significant (p=0.213). Moreover, the performance of this combined data set is only 0.70% higher than the performance of the **Syllable** feature set, in isolation, 77.10%. These results suggest – if high quality syllable segmentation is available – extracting features from the phrase-final syllable is the most fruitful region of analysis. Barring the availability of time-aligned syllabification information, extracting features from the final 200ms preceding a phrase boundary is a strong alternative, superior even to acoustically driven pseudo-syllabification techniques.

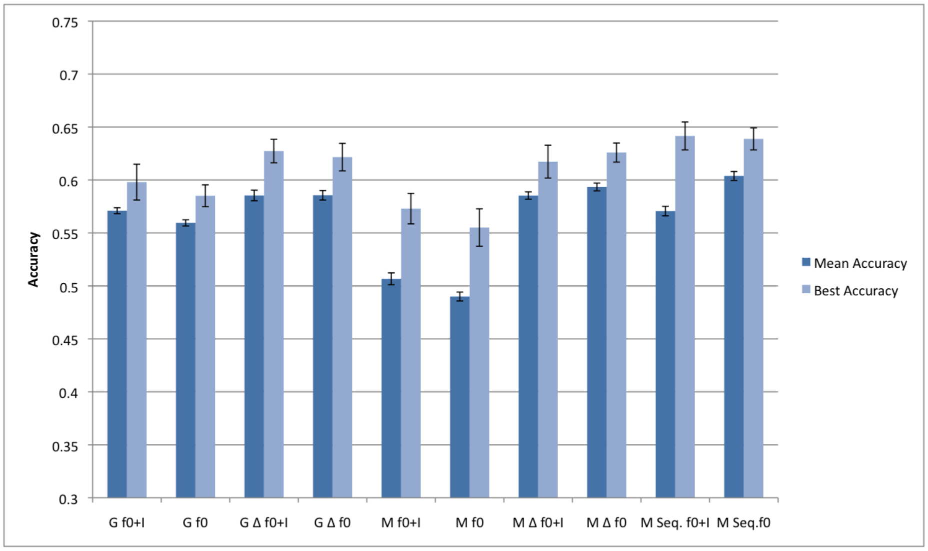
6.5.4 Quantized Contour Modeling

In Section 5.6.3, we define a Bayesian technique to model contour shape. The application of this model led to improved pitch accent type classification performance. In this section, we apply this modeling technique to the task of phrase-final classification.

The Quantized Contour Modeling technique is a relatively basic Bayesian modeling classifier. Each contour – pitch or intensity – is normalized and quantized into *N* time bins and *M* value bins. For each class and time bin, a multinomial model is trained, modeling the distribution of quantized values. At evaluation the probability that a test contour is a member of a class is given by product of the probabilities that each time model generated the input data point. A hypothesis is generated by taking the argmax over the set of classes. The formulaic definition of the classifier can be found in Section 5.6.3. Extensions to this modeling technique use the same core approach, but model the deltas of the contour, sequences of values, and simultaneously model both the pitch and intensity contours. Prior to quantization, during the normalization step, the contour range is normalized to the range [0,1].

In this section, we apply this modeling technique to the phrase-finals with a range of time and value bins between two and seven, using multinomial, multinomial delta and multinomial sequential models. Additionally, we replace the multinomial models with Gaussian models, modeling the quantized values, and their deltas, as well as unquantized, normalized contour values.

We first seek to identify the best modeling strategy for phrase-final classification. In these experiments we model the pitch and intensity contours over the **Full Word** region of each phrase-final word. Each modeling technique is evaluated using two to seven time and value bins. Graphs of the mean and best performing parameterization of each modeling technique with 95% confidence intervals can be found in Figures 6.14, 6.15 and 6.16.

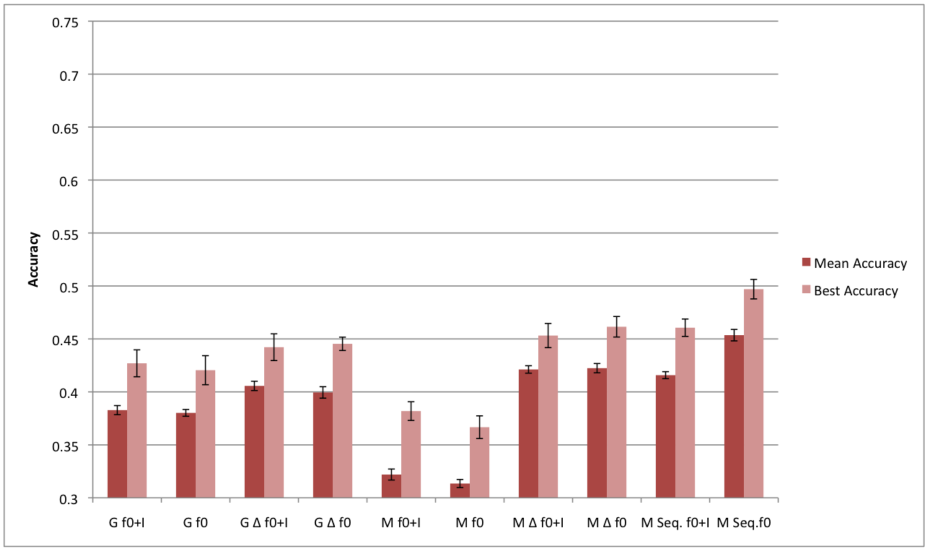


BDC-read

Figure 6.14: *Mean and Best Accuracy of Quantized Contour Modeling classification on BDC-read material. The number of Time and Value bins vary from 2 to 7.*

Averaging over all the examined configurations, we find that the multinomial sequential modeling of pitch and simultaneous modeling of pitch is the best performing modeling strategy.

Next, we attempt to identify the best performing region of analysis for this modeling approach. Based on the results observed in Section 6.5.3, we apply the Quantized Contour Modeling technique to the full word, last half of the word and the last 200ms and 400ms



BDC-spon

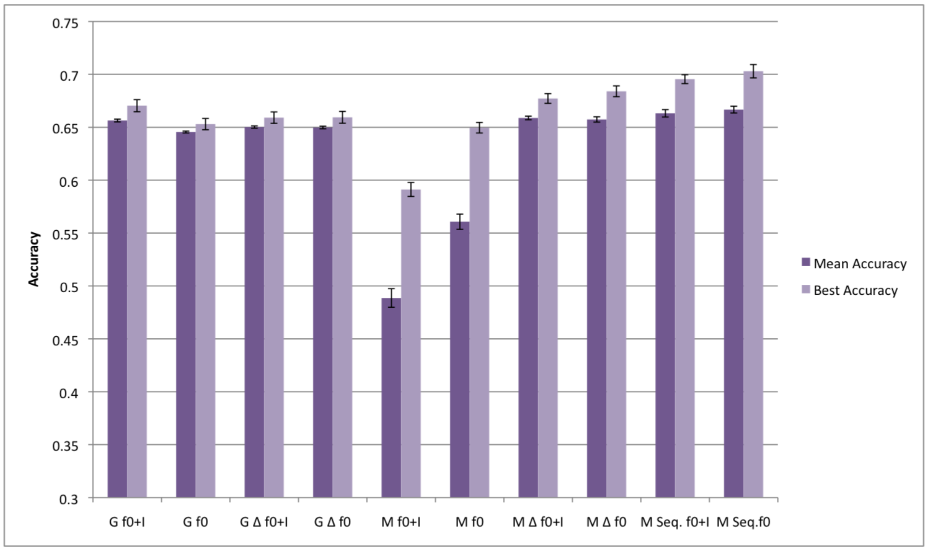
Figure 6.15: *Mean and Best Accuracy of Quantized Contour Modeling classification on BDC-spontaneous material. The number of Time and Value bins vary from 2 to 7.*

immediately preceding the phrase boundary. On the BURNC material, we also apply the model to the phrase-final syllable. Table 6.24 contains results of ten-fold cross-validation experiments on BURNC and BDC material using multinomial sequential modeling of the phrase-final pitch contour.

Region	BDC-read	BDC-spon	BURNC
Full Word	63.88 ± 1.04	***49.70* ± *0.920***	69.55 ± 0.421
Last Half	44.49 ± 1.35	43.67 ± 1.00	72.17 ± 0.603
Last 200ms	45.29 ± 1.18	43.90 ± 1.39	***73.25* ± *0.423***
Last 400ms	***64.02* ± *1.18***	49.01 ± 1.10	71.74 ± 0.620
Syllable	NA	NA	73.01 ± 0.630

Table 6.24: *Best performing Ten-fold cross-validation evaluation of Multinomial Sequential Quantized Shape modeling evaluated on different regions of analysis.*

This examination yields conflicting results. On the BDC data, the subword regions perform significantly worse than the modeling of the full word or 400ms. However, on BURNC, performance is significantly improved by modeling any of the other examined regions, with the **Last 200ms** region yielding the highest performance. Interestingly enough, this region requires the least amount of additional information for identification. It is unclear



BURNC

Figure 6.16: *Mean and Best Accuracy of Quantized Contour Modeling classification on BURNC material. The number of Time and Value bins vary from 2 to 7.*

why this region of analysis performs poorly on the BDC material, but successfully on the BURNC material, particularly in light of the acoustic classification results discussed in Section 6.5.1. Acoustic features extracted from the **Last 200ms** region were significantly more discriminative of phrase-final type when applying SVM classification. Due to its inconsistent performance, the **Last 200ms** region cannot be considered the best region of analysis for the application of Quantized Shape Modeling. The **Last 200ms** and **Last 400ms** regions are particularly attractive because they require no additional information beyond the phrase boundary location for identification. The **Last 400ms** region generates more consistent performance across corpora. On BDC-read and -spontaneous, its accuracy is not significantly less than the best performing region, **Full Word**. Yet, on BURNC the performance is significantly worse than applying the model to the **Last 200ms** region. The improved performance on the BDC subcorpora, indicates that this classification technique can be successfully applied without the addition of word identity/boundary information. The identification of the best region of analysis to apply the technique remains an open question, but regions that are approximately word sized appear to be more robust to differences in genre and speaking style. This result suggests that there may be large differences between the BDC and BURNC material. We know that the word segmentation information is more reliable on the BDC material; the BDC word segmentation is manually performed, while the BURNC word segmentation is the output of a forced alignment system. However, there may be speaker difference due to speaking style or physiological qualities or labeler differences. We are not aware of any investigation into the consistency of ToBI labeling across annotated corpora. Such analysis could be helpful in understanding these differences, and would provide a measure of the expected consistency of automatic prosodic analysis systems.

Overall, this classification technique performs significantly worse than the SVM classification employed in Section 6.5.1. However, the two modeling strategies may capture distinct aspects of the phrase-final contour. We attempt to integrate these disparate strategies in two ways. We combine their predictions and we include the shape model posteriors in the SVM feature vector. For each of these combination experiments, we use the Quantized Contour Model parameters which yielded the highest accuracy when evaluated in isolation. As a proof-of-concept evaluation of these combination strategies we evaluate the performance on BDC-read using the **Full Word**, **Last 200ms** and **Last 400ms** regions. Results of these experiments can be found in Table 6.25. The posterior combination approach performs significantly worse than the SVM classification in isolation. However, inclusion of the Quantized Contour Modeling posteriors in the SVM feature vector does demonstrate improved performance. However, the best combination performance, 71.75% accuracy, is not significantly better than the best SVM classification in isolation, using all acoustic

Region	Parameters (Time, Value)	Posterior Combination	SVM w/ Posterior Features
Full Word	(7,7)	64.23 ± 2.271	69.46 ± 2.597
Last 200ms	(7,7)	61.02 ± 5.752	71.75 ± 2.396
Last 400ms	(7,3)	64.74 ± 1.482	71.67 ± 1.68

Table 6.25: *Ten-fold cross-validation evaluation of combination techniques of Quantized Shape modeling and SVM classification of acoustic features. Evaluated on BDC-read material.*

features extracted from the last half of phrase-final words, 72.32% ± 2.4928. We extend this evaluation to the BDC-spontaneous and BURNC material and report these results in Table 6.26.

Region	BDC-spon	BURNC
Full Word	(5,6) 58.04 ± 1.696	(7,7) 75.63 ± 1.397
Last 200ms	(3,7) 58.36 ± 1.505	(5,7) 75.96 ± 0.528
Last 400ms	(6,5) 57.72 ± 1.894	(7,7) 76.28 ± 1.111
Syllable	NA	(7,7) 77.55 ± 1.027

Table 6.26: *Ten-fold cross-validation evaluation of including Quantized Shape modeling posteriors in an SVM classification feature vector. Evaluated on BDC-spontaneous and BURNC material. Parameters (Time, Value) are included.*

These accuracies also represent the highest phrase-final classification achieved in this chapter on these corpora. On BDC-spontantous, the accuracy of classification achieved over **Full Word** regions, 58.68%, is greater than the best performance obtained without the inclusion of Quantized Shape Model posteriors. However, this accuracy not significantly better than the SVM classification accuracy with features extracted from the **Last 400ms**, 57.44% ± 1.7384. There are no significant differences on these corpora between the **Full Word**, **Last 200ms** and **Last 400ms** regions of analyses. When available, syllable boundary information can be used to modestly improve accuracy, as we observe on the BURNC material this improves classification accuracy to 77.55%.

The speaker independent performance of this approach, evaluated using leave-one-speaker-out cross-validation can be found in Table 6.27. On the BDC corpora, we find speaker independent modeling to be between 5 and 10% less accurate than the cross-validation evaluation. On the BURNC data, the speaker independent modeling does not perform significantly worse. There are two main differences between the BURNC corpus and the BDC corpora. First, the BURNC material is read by professional newscasters, and second, there is more available training material. One or both of these two factors may lead to the success of speaker-independent modeling. The professional speaking style may confer inter-speaker commonalities that allow for easier generalization from one speaker to another.

Region	BDC-read	BDC-spon	BURNC
Full Word	(7,7) 63.09 ± 2.704	(5,6) 52.69 ± 3.594	(7,7) 73.30 ± 3.183
Last 200ms	(7,7) 62.80 ± 2.898	(3,7) 52.83 ± 4.806	(5,7) 74.63 ± 2.880
Last 400ms	(7,3) 62.66 ± 4.454	(6,5) 52.92 ± 3.067	(7,7) 74.21 ± 2.372
Syllable	NA	NA	(7,7) 75.09 ± 2.288

Table 6.27: *Leave-one-speaker cross-validation evaluation of including Quantized Shape modeling posteriors in an SVM classification feature vector. Parameters (Time, Value) are included.*

Additionally, the availability of more data may lead to the training of more robust models.

Overall, we find that Quantized Contour Modeling alone performs worse than SVM classification using acoustic features. However, the two classification approaches can be combined to modestly improve classification accuracy, suggesting that Quantized Contour Modeling is capturing salient information that may not be represented by other acoustic features. This modeling technique is quite sensitive to the settings of its parameters, and the modeling approach employed. While the best performing configurations can contribute to improving phrase-final classification performance, the best parameter settings are not consistent across corpora. This certainly limits the widespread application of this modeling technique, despite its ability to capture distinct shape information in a novel way.

6.5.5 Final segment class modeling

Phrase-final contours are realized immediately prior to a phrase boundary. As we observed in Section 6.5.3, and again in Section 6.5.4, much of the information discriminative of phrase-accent type can be found in the 200ms immediately before the phrase boundary – approximately the last half of the phrase-final word. Forced-alignment generated phone identity and boundary information is available for the BURNC data. Inspection of this information reveals that phrase-final phones have a mean length of 92ms – accounting for nearly half of the most discriminative region of analysis. We observed in Section 6.5.1 that the pitch contour contains the most discriminative information – yet, only 22.6% (1260 of 5573) of phrase-final phones are vowels. The extraction of pitch information from obstruent consonants is difficult at best, and entirely noise at worst. It is likely that the extraction of acoustic, or at least pitch features, from phrase-final regions should be handled differently when the region contains sonorant or obstruent segments (phones).

In this section, we attempt to use information about phrase-final segment identity to improve phrase-final classification performance. The experiments in this Section are all performed on the BURNC material, as this is the only corpus for which time-aligned segment identities are available. Since these experiments are based on the assumption that this information is available, we will use the phrase-final syllable as the region of analysis. This region generated the highest accuracy phrase-final classification in the experiments reported in Sections 6.5.3 and 6.5.4. The drawback of using this region of analysis is the additional data requirement; however, this burden is irrelevant in these experiments. We use the classification configuration that generated the highest accuracy results in all previous Sections. This is SVM classification using acoustic features extracted over the phrase-final syllable, as well as Quantized Contour Modeling posteriors.

To model phrase-final types distinctly for each phrase-final segment type, we train different models for each phrase-final segment. The BURNC phone inventory contains 54 elements. Thus we train 54 classifiers – one for each phrase-final phone id. Note, however, that this will greatly reduce the training data for each classifier. The phrase-final segment identity will need to be an extremely powerful indicator of phrase-final type for this classification scenario to improve performance with the fifty-fold reduction in available training data.

In addition to this segment identity based modeling, we cluster the phrase-final phones into a number of different sets based on their segment identity. These clusters are used to partition the data set before classifier training. In one experiment, we train distinct classifiers for phrase-final vowels (**V**), sonorant consonants (**S**), voiced obstruent consonants (**VO**), and unvoiced obstruent consonants (**UVO**). In the next experiment, we collapse both obstruent consonant classes, training three classifiers. Finally, we train two classifiers, one for sonorant consonants and vowels and a second for obstruent consonants. The results of all these segment-based experiments are reported in Table 6.28.

Segment Classes	Accuracy
None (1)	78.99 ± 0.951
Phone ID (54)	72.79 ± 0.853
V v. **S** v. **VO** v. **UVO** (4)	78.10 ± 0.836
V v. **S** v. **VO+UVO** (3)	78.02 ± 0.590
V+S v. **VO+UVO** (2)	78.36 ± 1.164

Table 6.28: *Ten-fold cross-validation evaluation of segment class based phrase-final type syllable classification.*

These experiments reveal that this class-based modeling technique is unable to improve performance by modeling phrase-final syllables separately based on the class of the final phone. While phrase-final types may behave differently depending on the final phone, the differences do not manifest themselves strongly enough to improve classification performance. Yet, the performance of these class-based classifiers only suffers significantly in the phone-identity-based classification – where the available training data is reduced more than fifty-fold. This suggests that while it is powerful enough to improve performance, there is sufficient discriminative information in the phrase-final segment to prevent significant reductions in performance due to the reduced training material. To measure the degree to which this improves classification, we compare the class-based modeling results contained in Table 6.28 to SVM classification performance with randomly downsampled training data. The downsampling approximates the amount of training data available to each classifier in the class-based modeling scenario, but the data is segmented randomly, as opposed to by the identity, or class membership, of the final segment. A graph of this comparison can be found in Figure 6.17.

What is most striking about this comparison is how robust the acoustic SVM classifier

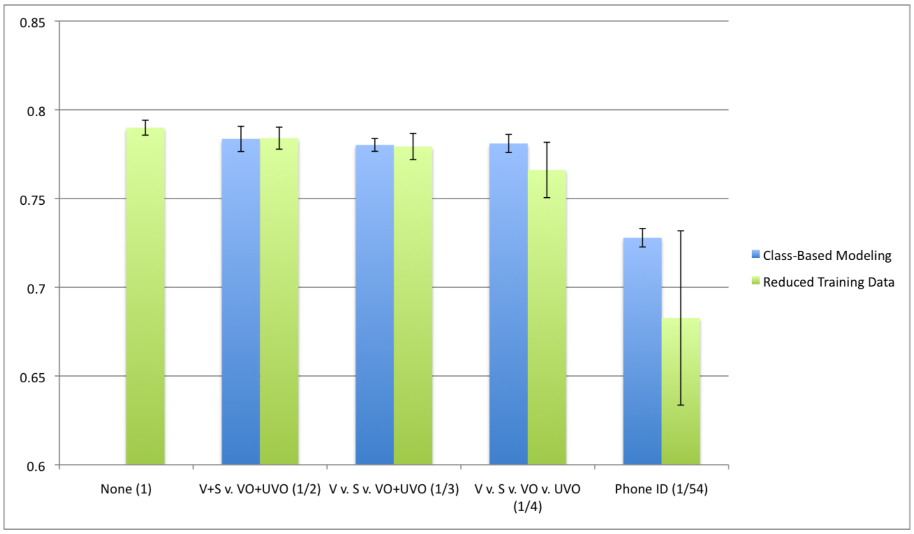


Figure 6.17: *Comparison of class-based classification and downsampled classification.*

is to reductions in available training data. The classification accuracy is not significantly lower when trained with one-fourth the amount of training data. The only situation in which class based modeling notably outperforms the downsampled SVM classifier is in the phone-identity-based modeling. In this scenario, 54 distinct classifiers are trained, one for each phrase-final phone. Conversely the SVM classifier is trained with 1/54 of the available training data. However, this difference is not statistically significant (p=0.3838). This casts doubt on the notion that there is even a minor effect of phrase-final segment identity on phrase-final type classification. If there is, in general this effect is clearly overwhelmed by the robustness of the acoustic based SVM classification to reductions in the amount of available training data.

6.6 Phrase Accent Classification

Intonational, or full, phrases are composed of one or more intermediate phrases. While all intermediate phrases have an associated phrase boundary, under the ToBI standard, only intonational phrases carry a boundary tone. Thus, the phrase-final behavior at these internal phrase boundaries is characterized by a either a high phrase accent (H-) or a low phrase accent (L-). High phrase accents correspond to rising or sustained high pitch contours, while a low phrase accent indicate falling pitch. Hirschberg and Pierrehumbert hypothesize that high phrase accents indicate that the current and following intermediate phrases should be interpreted together, as part of a unit, while a low phrase accent indicates for the two phrases to be interpreted separately [153].

In this section, we apply the lessons learned in the classification of phrase-finals – pairs of phrase accents and boundary tones – to the classification of intermediate phrase-final intonation, phrase accents that occur without boundary tones. As SVM classification yielded the best classification performance when classifying intonational phrase-finals, we evaluate only this classification algorithm in this section. We will focus on presentation of results and drawing comparisons between this classification task and the automatic classification of intonational phrase-finals.

When examining the use of phrase-internal phrase accents, we see a strong influence of genre. In the read portion of BDC, 62.23% of phrase accents are L-, while in the spontaneous material of the corpus, 50.55% are H-. BURNC contains an even greater proportion of H- phrase accents with 64.09%. All three genres have similar rates of internal intermediate phrases, 54.1% of intermediate phrases in BDC-read are intonational phrase-internal, while 45.9% of BDC-spon and 47.76% of BURNC intermediate phrase boundaries fall within intonational phrases. The lexical material in the BDC-read and BDC-spontaneous corpora are identical, with the exception of the removal of disfluencies. This suggests that the disparate use of H- and L- phrase accents in these two subcorpora are evidence of a difference in speaking style when reading aloud versus speaking spontaneously.

In Section 6.5.1, we examine three distinct acoustic feature sets: 1) acoustic aggregations, 2) shape modeling features and 3) voice quality features. In Section 6.5.3, we find that extracting these features from different regions of analysis significantly impacts their discriminative power with regard to intonational phrase-final classification. Specifically, we find that extracting acoustic features from the **Last Half** of the phrase-final word, the **Last**

200ms immediately preceding the phrase boundary and the phrase-final **Syllable** leads to features that are able to classify phrase-final types with better performance than the **Full Word**. In this section, we evaluate the use of these acoustic feature sets in classifying phrase accents. These features are extracted over the **Full Word**, **Last Half**, **Last 200ms** and **Syllable** regions of analysis. Results from ten-fold cross-validation experiments on BDC-read, BDC-spontaneous and BURNC are presented in Tables 6.29, 6.30 and 6.31, respectively.

Feature Set	Word	Last Half	Last 200ms
Aggregations	69.97 ± 3.067	70.37 ± 3.723	68.65 ± 2.017
Shape	67.46 ± 2.608	67.20 ± 1.509	67.33 ± 2.066
Voice **Q**uality	63.23 ± 2.001	63.23 ± 4.166	63.23 ± 3.329
A+S	72.22 ± 2.558	70.37 ± 3.313	70.90 ± 2.837
A+S+VQ	72.09 ± 2.493	71.43 ± 2.540	71.69 ± 2.476

Table 6.29: *Ten-fold cross-validation accuracy (%) of SVM phrase accent type classification of BDC-read material using acoustic features extracted from a range of analysis regions. Baseline=63.23% (L-)*

Feature Set	Word	Last Half	Last 200ms
Aggregations	65.97 ± 2.657	66.80 ± 2.411	65.50±1.89
Shape	62.19 ± 2.148	61.38 ± 2.017	61.79 ± 1.788
Voice **Q**uality	54.06 ± 2.657	55.47 ± 2.411	54.36 ± 2.804
A+S	69.41 ± 1.525	68.41 ± 1.525	68.61 ± 1.656
A+S+VQ	67.90 ± 3.001	69.11 ± 2.312	66.60 ± 2.280

Table 6.30: *Ten-fold cross-validation accuracy (%) of SVM phrase accent type classification of BDC-spontaneous material using acoustic features extracted from a range of analysis regions. Baseline=50.55% (H-)*

One clear difference between the classification of intonational phrase finals, composed of a phrase accent and boundary tone, and the classification of intonational phrase-internal phrase accents is the discriminative power of features extracted from different regions of analysis. When classifying phrase-finals, we find that regions at the end of a phrase-final word contain more discriminative acoustic information; there is no such effect in the results of the internal phrase accent experiments. No statistically significant difference in classification accuracy is observable due of the region of analysis. This indicates both that, the material in the initial half of a word is not particularly discriminative to phrase accent type, and, on the other hand, it does not introduce any significant noise to the problem. This leads to the conclusion that the region of analysis for this task is relatively immaterial, and can be dictated by the available annotations of the input speech material.

Examining the contributions of different feature sets, we find that while voice quality features showed some differences based on phrase-final type. However, these features are unable to classify phrase-accents significantly better than baseline on any corpora. On the BDC corpora, when combined with the Aggregations and Shape feature sets, the inclusion of voice quality features does not significantly improve classification accuracy. However, and somewhat surprisingly, the inclusion of voice quality features *does* significantly improve classification performance on the BURNC material.

Feature Set	Word	Last Half	Last 200ms	Syllable
Aggregations	65.06 ± 0.951	64.99 ± 1.689	64.84 ± 2.476	64.31 ± 2.13
Shape	64.27 ± 2.493	64.31 ± 1.263	64.27 ± 0.590	64.09 ± 1.345
Voice **Q**uality	64.09 ± 1.574	64.09 ± 1.476	64.09 ± 1.099	64.05 ± 1.558
A+S	68.26 ± 1.788	68.22 ± 1.607	68.41 ± 1.492	67.96 ± 1.558
A+S+VQ	72.28 ± 1.853	72.09 ± 1.656	72.39 ± 1.410	70.14 ± 1.082

Table 6.31: *Ten-fold cross-validation accuracy (%) of SVM phrase accent type classification of BURNC material using acoustic features extracted from a range of analysis regions. Baseline=64.09% (H-)*

Examination of the mean voice quality measures, jitter and shimmer, of H- and L- phrase accents reveals the source of this improvement. While we find no significant difference in the shimmer of these two phrase accents, the jitter significantly differs. H- phrase accents have a mean jitter of 0.0152, while L- phrase accents have a mean of 0.0218, a difference that is significant with $p < 2.2 * 10^{-16}$ evaluated using a t-test. While not a powerful enough difference to be discriminative in isolation, this feature improves phrase accent classification accuracy in combination with other acoustic features.

Pierrehumbert and Hirschberg hypothesize that when an intermediate phrase ends with a high phrase accent (H-) it forms a larger whole with the following phrase. On the other hand, phrases separated by an intermediate phrase boundary and a low phrase accent (L-) are interpreted separately [153]. This hypothesis suggests that the structural syntactic features that are examined in Section 6.5.2 may be applicable to this classification task as well. It would be consistent with this theory that phrase boundaries with a lower degree of syntactic disjuncture would have a H- phrase accent, while those with greater disjuncture would be produced with an L-. Ten-fold cross-validation of the structural features is presented in Table 6.32. We also include, in this Table, the classification accuracy when combining these syntactic features with the full set of acoustic features.

Corpus	Baseline	Parse	Parse+ Acoustics
BDC-read	62.23	64.68 ± 3.936	71.69 ± 3.362
BDC-spon	50.55	53.16 ± 2.886	68.71 ± 1.771
BURNC	64.05	64.24 ± 1.443	72.23 ± 1.952

Table 6.32: *Ten-fold cross-validation accuracy (%) of SVM phrase accent type classification using structural syntactic features.*

While the syntactic features were helpful in classifying intonational phrase finals, their use in classifying intermediate phrase accents is clearly limited. On the BDC-read and BURNC material, the performance is not significantly better than the majority class baseline.

On BDC-spontaneous, performance is improved by less than 3%, an improvement that is not statistically significant. In combination with the acoustic features, the inclusion of these syntactic features does not impact classification performance significantly.

To examine the, albeit minor, interaction between phrase accents and the structural syntactic features, we look at some descriptive statistics from the BDC-spontaneous material – the corpus which demonstrated the highest classification performance using these features. We find that the difference in syntactic distance and distance ratios across boundaries with H- and L- phrase accents approach significance with p=0.0422 and p=0.0541, respectively. The mean syntactic distance at high phrase accented (H-) boundaries is 4.16 compared to a mean of 4.55 at low phrase accented boundaries. This is not nearly as significant as the difference observed between phrase-final types. The relative phrase position also shows a difference in means that approaches significance, with p=0.0309; H- phrase accents are used at a relative phrase position of 0.29, while L- phrase accents have a mean relative phrase position of 0.35. While there are some minor differences in the syntactic context in which the two phrase accent types are used, these differences are not very large. While phrase accents may indicate the intended compositional interpretation of the content of the intermediate phrases surrounding the phrase boundary, we do not see this composition reflected in syntactic parse trees.

Next, we examine the use of Quantized Contour Models to classify phrase accents. Table 6.33 contains the results of the best performing Quantized Contour Model configuration on each corpus. Also included in this Table are the results of including the posteriors of the best performing Quantized Contour Model classifier in an SVM feature vector with acoustic features. When evaluating the use of acoustic features (cf., e. g. Table 6.29), we find no significant difference in the performances of acoustic features drawn from any region of analysis. Table 6.33, thus, only reports the classification accuracy of Quantized Contour Modeling of contours extracted from the **Full Word** region of analysis.

The Quantized Contour Model successfully classifies phrase accent, especially on the

Corpus	Baseline	Modeling Approach	Accuracy	SVM Accuracy w/ Acoustic Features
BDC-read	62.23	multinomial f0 (5,6)	71.16±2.083	73.02 ± 2.148
BDC-spon	50.55	multinomial delta f0 (2,7)	66.00±2.542	67.70 ± 2.345
BURNC	64.05	gaussian f0 (2,4)	67.58±0.984	72.24 ± 1.148

Table 6.33: *Ten-fold cross-validation accuracy (%) of Quantized Contour Modeling phrase accent type classification with best parameterization.*

non-professional speech of the BDC. On the BDC material, the difference between the Quantized Contour Model accuracy and the SVM classification accuracy is not significant. However, the classification performance of the best configurations *are* lower than those obtained by SVM classification with the previously evaluated acoustic features. Moreover, when the posteriors of this classifier are included with the acoustic features in training an SVM classifier, the performance is not significantly different from that obtained using the acoustic features alone. This result, along with the burden of optimizing its model parameters, limits the broad appeal of applying Quantized Contour Modeling to the task of phrase accent classification.

Though the acoustic features showed no significant differences due to the region of analysis over which they were extracted, the Quantized Contour Model is a fundamentally different representation of this acoustic information. Thus, we experiment with modeling of pitch and intensity contours drawn from the **Last Half**, **Last 200ms** and, on BURNC, **Syllable** regions of analysis. When evaluating the Quantized Contour Model classification performance on these regions of analysis, we, again, find no significant differences in the accuracy of the best performing configuration on BDC-read and BURNC material. On the BDC-spontaneous corpus modeling only the **Last 200ms** region achieved a phrase accent classification accuracy of 61.38%, significantly below the best performance when modeling the **Full Word** (66.00%) or **Last Half** of the word (66.00%).

To evaluate the robustness of the approaches presented in this section to speaker differences, we evaluate their performance using leave-one-speaker-out cross validation. The results of this evaluation are presented in Table 6.34.

Feature Set	BDC-read	BDC-spon	BURNC
Baseline	62.23	50.55	64.05
A+S+VQ	61.90 ± 13.956	52.76 ± 13.104	68.78 ± 3.231
Parse	61.24 ± 7.150	46.24 ± 6.757	NA
Parse + **A+S+VQ**	59.26 ± 14.465	55.47 ± 10.545	NA
Quantized **C**ontour **M**odeling	70.37 ± 3.821	57.57 ± 9.463	66.08 ± 2.378
QCM + **A+S+VQ**	65.34 ± 14.711	56.37 ± 11.660	68.78 ± 2.936

Table 6.34: *Leave-one-speaker-out cross-validation accuracy (%) of phrase accent type classification.*

We find that the Quantized Contour Modeling classification to be *less* sensitive to speaker differences than the acoustic and parse based features – most notably on the BDC-read material. However, speaker dependent evaluation yields performance that is significantly greater than speaker independent evaluation on all feature sets. The presence of this training material improves classification performance of the acoustic feature set by 10.19% on BDC-read, 15.14% on BDC-spontaneous and 3.5% on BURNC. BURNC may be more speaker independent due to the conventional broadcast news speaking style, or due to the increased amount of available training data.

We are not aware of any earlier studies classifying intonational phrase-internal phrase accents. Syrdal and McGory [196] found that the rate of pairwise human agreement on these prosodic events to be quite low, less than 40%. The best speaker independent automatic phrase accent classification performance reported in this section is 57.57% on BDC-spontaneous, 70.37% on BDC-read and 68.78 on BURNC. This performance is significantly *higher* than the interannotator agreement found by Syrdal and McGory.

There are few intonational theories about the communicative use of phrase accents at intonational phrase internal intermediate phrase boundaries. Pierrehumbert and Hirschberg [153] hypothesize that high phrase accents lead to the interpretation of two surrounding intermediate phrases as combining to form a larger whole, while low phrase accents indicate some disjuncture in the interpretation of the two. However, it remains unclear how automatic extraction of this acoustic information should best be used by downstream spoken language processing modules. Moreover, identifying the location of intonational-phrase-internal intermediate phrase boundaries is quite difficult (cf. Section 4.5). This classification task is predicated on the assumption that the phrase boundary has already been located, a non-trivial task. That said, the experiments and results presented in this section suggest, given improvements to intermediate phrase boundary detection, phrase accent information can be extracted automatically at rates that approach human agreement.

6.7 Conclusion and Future Work

Intonation at the end of a phrase is used for a number of important communicative purposes. It is used to indicate relationships between one phrase and the next. These relationships define the discourse structure and coherence relationships between the two phrases, as well as the intended semantic or pragmatic interpretation of their content; it can be used to indicate a declarative or tag question. Phrase-final behavior also serves some crucial discourse functions; it is commonly used to hold or cede the turn or to request confirmation. These functions are important components of interactive voice response (IVR) systems which could greatly benefit from appropriate interpretation and production of these discourse behaviors.

In this chapter, we explore a range of techniques for automatic phrase-final classification. Some key results are summarized in Table 6.35. There are three major contributions of the research presented in this chapter. First, we identify acoustic features which yield high classification performance. Under speaker-independent evaluation these accuracies are 63.81% accuracy on BDC-read, 51.27% accuracy on BDC-spontaneous and 77.64% accuracy on BURNC. Using ten-fold cross-validation, these improve to 73.39% on BDC-

	Classification Approach	BDC-read	BDC-spon	BURNC
	Phrase Accent/Boundary Tone Pairs			
SVM w/ linear kernels and SMO	Pitch Aggregations	64.66	47.89	60.33
	Intensity Aggregations	62.59	46.75	59.99
	All Aggregation Features	65.59	50.76	65.40
	Shape Features	65.31	52.33	67.22
	All Acoustic Features	69.81	57.12	73.34
	Parse Tree Features	67.53	48.09	59.57
	POS tag Modeling[a]	54.36	43.30	60.60
	All Features	71.60	58.31	73.50
	Acoustic Features Region of Analysis			
	200ms	71.46	54.95	76.37
	Half Word	72.32	55.50	76.22
	Syllable	NA	NA	77.10
	QCM[b]	64.02	49.70	73.25
	Acoustic Features + QCM	***71.75***	***58.36***	***77.75***
	Phrase Accents Only			
	Acoustic Features	72.22	***69.41***	***72.28***
	QCM	71.16	66.00	67.58
	Acoustic Features + QCM	***73.02***	67.70	72.24

[a] Best POS tag results. Surrounding bigram raw POS tags on BDC-read and BURNC. Surrounding bigram broad class POS tags on BDC-spontaneous.

[b] Best QCM results. See Section 6.5.4 for parameters and regions of analysis.

Table 6.35: A summary of select phrase ending tone classification experiments. All evaluations use ten-fold cross-validation and are reported in Accuracy (%). The best performance on each corpus is indicated in bold and italics.

read, 58.18% on BDC-spontaneous, and 78.99% on BURNC. There is not a lot of other research on the classification of phrase ending intonation. The best previous published result on BURNC material yielded 72.4% accuracy on a single speaker subset of the corpus [173]. Other research on this task has treated this classification task differently, either classifying L-H% accents from L-L% [161] or classifying phrase accents and boundary tones separately [208].

The application of linear regression, and extrema location features are novel. Also, Quantized Contour Modeling has not previously been applied to this task. In isolation, the Quantized Contour Modeling classification technique does not yield the best performance. However, its posteriors can be used by an SVM classifier to improve acoustic-only performance. Second, we identify the region 200ms prior to an intonational phrase boundary as the optimal region of acoustic analysis. This region generates the most discriminative acoustic features across corpora and requires no additional lexical information for its identification. Third, we examine the interaction of structural syntactic features and phrase-final types, identifying syntactic features that can be used to classify phrase-finals with relatively high accuracy – as high as the acoustic features under speaker-independent evaluation.

In Section 6.5.1, we examine acoustic features that can be used to differentiate phrase-final types. We find that a combination of information about the shape of the pitch contour, and aggregations over the duration of the phrase-final word can combine to accurately classify phrase-final behavior. Evaluation of voice quality features indicates that there is a significant difference in voice quality during H-L% phrase-finals. However, the inclusion of this information does not significantly improve automatic classification performance. Extracting these acoustic features over the duration of the phrase-final word, we observe classification accuracies over 70% on both read corpora, and 57.12% on spontaneous speech. This performance is reduced when evaluated in a speaker-independent setting, to 47.81% accuracy on spontaneous speech, 62.02% accuracy on non-professional read speech, and 70.70% accuracy on professionally read broadcast news.

The hypothesized uses of phrase-finals to indicate structural relationships across intonational phrases suggests that there may be an interaction between syntactic structure and phrase-final types. For example, the continuation rise contour is generally believed to indicate a sense of incompleteness, that there is “more to come”. This would suggest that it may be used between two independent clauses to indicate to a listener that a sentence is not yet complete. Using automatically generated parse trees and part-of-speech tags, we investigated the use of syntactic information in the automatic classification of phrase finals. Part-of-speech information, as applied in this chapter, does not prove to be helpful in the differentiation of phrase-final type. However, we find that structural features extracted from parse trees to be able to classify phrase finals with moderate accuracy. While these perform worse than our acoustic features, they yield accuracies that are significantly above the majority class baseline on all corpora – 67.53% on BDC-read, 48.09% on BDC-spontaneous and 59.57% on BURNC. These features. however, appear to be more robust to speaker dependencies; their performance is not significantly worse when evaluated in a speaker independent setting. Moreover, under speaker independent evaluation the performance of structural syntactic information does not significantly differ from acoustic information. Unfortunately, when we combine the acoustic features with these syntactic features, we find no significant improvement to phrase-final classification accuracy. While this limits their usefulness in the analysis of human speech, these syntactic findings may be useful for prosodic assignment, where the desired prosody must be inferred from text input alone.

In the remainder of the chapter, we explore other techniques to extract discriminative information from the acoustic content of the phrase-final. When extracting acoustic features, we initially calculate the features over the full duration of the phrase-final word. However, phrase accents and boundary tones occur at the end of intonational phrases. Acoustic information from, for example, the beginning of the phrase-final word may fall outside the boundary of this prosodic event. To isolate the best performing region, we evaluate the performance of acoustic features extracted from a number of regions. We find that a 200ms region – approximately half a word – immediately prior to the phrase boundary generates the best performing acoustic features, while requiring no additional information for its identification. Word boundary information was not able to identify a region of analysis that performed significantly better than the final 200ms region. Extracting acoustic features from the phrase-final *syllable* generated the best performing phrase-final classification. Syllabification information, however, is rather resource intensive to extract. Acoustic pseudo-syllabification techniques were not able to identify a region of analysis as successful as forced-alignment syllabification based on manual transcription.

Evaluation of the Quantized Shape Modeling classification technique reveals that this technique can be used to successfully classify phrase-finals. While, the performance of this technique is worse than that of SVM classification using acoustic features, when used the posteriors are included in the SVM feature vector, the best ten-fold and speaker-independent performance is achieved across corpora. The downside to this modeling technique is its sensitivity to its parameter settings. Performance changes dramatically depending on the number of time and value quantization bins that are used. Moreover, the optimal modeling parameters are not consistent across corpora. This limits the broad application of this technique.

Finally, we explore the possibility that the identity of the phrase-final segment has a significant impact on phrase-final modeling classification. We find that given the amount of available training data, there is no impact of the phrase-final segment identity. Though if we dramatically reduce the amount of training data available to the SVM classifier, training classifiers separately for each phrase-final segment can significantly improve classification performance. However, to observe this effect, the amount of available data must be reduced by a factor greater than fifty. This indicates that there is not a major impact of phrase-final segment identity on phrase-final intonation classification performance.

We confirm through the experiments in this chapter that it is possible to extract phrase-final intonation information from speech with high accuracy. Due to wide range of communicative uses of this information, we believe that downstream spoken language processing tasks will be able to take advantage of these hypotheses to improve performance. We expect to evaluate this belief in the future. In Chapter 7, we use phrase boundary hypotheses to segment spoken broadcast news material prior to extractive summarization and story segmentation. We hypothesize that phrase-final information will help these, as well as other, tasks. We observe significant differences in the use of phrase-final intonation based on domain and genre. While we expect broadcast news, like that found in the BURNC material, to have a distinct speaking style, we observed significant differences across the two BDC corpora. These corpora differ mainly in speech genre – the set of speakers is identical and the lexical content, save the presence of disfluencies in spontaneous speech, is equivalent. In the future, we expect to perform a closer examination of the use of phrase finals in this corpus. Such a study may be able to provide insight into the difference between read and spontaneous speech.

6.7.1 Key Observations

- **Pitch is more useful than Energy for classifying phrase ending intonation.** While both contribute to improve classification performance, pitch is a more powerful signal distinguishing types of phrase-final intonation.

- **Examination of 200ms prior to a phrase boundary is sufficient for high accuracy classification of phrase ending intonation.** This result allows phrase-final classification to be performed without reference to lexical information like word or syllable boundaries. If lexical information is available, it can be used to improve classification accuracy. However, performance is not dramatically impaired by examining the 200ms region.

- **Syntax carries information that can be used to predict phrase ending intonation.** Parse tree information can be used to predict phrase-final intonation. While the syntactic features we explore are less discriminative than the acoustic features, this is a useful finding especially for prosodic assignment applications where acoustic information is unavailable.

- **The same approaches can be used to classify intermediate and intonational phrase ending intonation.** While the classifier parameters are distinct, the overall techniques used in classifying intonational phrase ending intonation – phrase accent/boundary tone pairs – are successfully applied to the classification of intermediate phrase ending intonation – phrase accents.

Chapter 7

Applications

Prosodic variation is critical to human communication. We have made the argument throughout this thesis that extraction of prosodic events will allow spoken language processing tasks access to information that is at least helpful and at most critical for achieving human-like interaction with speech. In this chapter, we apply prosodic event detection to three spoken language processing tasks: extractive summarization of broadcast news, story segmentation, and assessment of non-native speech.

These applications represent proofs-of-concept that there is important information in prosodic events for spoken language processing (SLP) applications. Hypothesized intonational phrase boundaries are used in both the extractive summarization of broadcast news and story segmentation applications to identify candidate data points for analysis. In the summarization application, we evaluate the use of intonational phrases as units for extractive summarization. We use intonational phrase boundaries to identify candidate locations for story boundaries in the segmentation application. In both of these investigations, we compare the use of intonational phrase segmentation to other possible segmentations. We find intonational phrases to be the best unit for extractive summarization of broadcast news compared to sentences, words, and pause-based segmentation. In the story segmentation application, we observe improved story segmentation performance on English, Arabic and

Mandarin Chinese broadcast news using hypothesized intonational phrase boundaries compared to hypothesized sentence boundaries. However, we find that word boundaries as well as boundaries indicated by short pauses, or low threshold sentence boundaries yield better segmentation performance than intonational phrase boundaries. In assessing non-native speech, we find that, using a bigram tone sequence model using hypothesized prosodic events, we can classify speech as native or non-native with high accuracy, over 95% accuracy with only seven words of material and over 85% with three words.

This chapter is structured as follows. Extractive broadcast news summarization experiments are described in Section 7.1. In Section 7.2, we describe work on story segmentation. Non-native speech assessment is presented in Section 7.3.

7.1 Broadcast News Summarization

Extractive speech summarization algorithms [94, 238, 241, 41] operate by selecting segments from the source spoken documents and concatenating them to generate a summary. Generally, the speech segments extracted for summarization should be semantically meaningful and coherent stretches of speech. Segmentations currently used or proposed for extractive summarization include words, phrases, sentences, or speaker turns [94]. Choice of segmentation unit greatly influences the length and quality of the resulting summary. If speaker turns are extracted, the shortest summary will be a single turn, which may contain many sentences, not all of which may be important. We have the most control over the length of the summary if we extract individual words. However, extraction of single words sacrifices the structural, semantic and syntactic information conveyed by the source document. Sentences might be a better choice of segmentation for extraction, since they are shorter than turns, affording finer control over the length of a summary. Moreover, sentences are semantically and syntactically meaningful units. However, longer sentences may include modifiers, phrases and clauses which are not essential for the summary. Syntactic phrase extraction is a promising alternative to sentence extraction, but identifying phrases in speech transcripts using current Natural Language Processing (NLP) tools is errorful, due to noisy automatic transcription and the disfluencies present in spoken language. However, as shown in Chapter 4 intonational phrases can be reliably identified using acoustic information from the speech signal.

In this section, we explore the use of intonational phrase units in extractive summarization of broadcast news. We present results of summarization experiments based on the extraction of four different kinds of segments: intonational segments based on pauses of 250ms and 500ms, hypothesized sentence units and automatically predicted intonational phrases. In Section 7.1.1 we discuss related work. We describe our corpus in Section 7.1.2. In Section 7.1.3 we discuss our experiments. We present results in Section 7.1.4, and conclusions in Section 7.1.5. This work on broadcast news summarization was first reported in [131].

7.1.1 Related Work in Extractive Speech Summarization

In recent years there has been a growing interest in speech summarization. Zechner [238] proposed a system to produce a summary of spontaneous speech using a Maximal Marginal Relevance technique. Hori [94] describes a word-based extractive summarization approach, selecting a set of words to produce a given summarization ratio, where words are selected using ASR word confidence scores, linguistic scores, word significance scores, and word concatenation scores based on a Dependency Grammar. Kolluru, et al. [106] extracted phrases by using a multi-stage filtering process in which perceptrons were employed at different stages of summarization to remove words with low confidence and to find significant segments. Zhu [241] extracted sentences of spontanous speech using a number of different feature sets. Maskey and Hirschberg [129, 130] proposed a sentence extraction system that identifies significant segments using acoustic, lexical, discourse and structural features in a machine learning framework. Each of these systems extracts some type of segment —

words, phrases, or sentences — although the approaches vary in the length of the segment as well as in their extraction technique. However, none of this previous work has examined the use of intonational phrases in extractive summarization or the impact of the extraction unit chosen on summarization performance.

7.1.2 Speech Summarization Corpus

The corpus we use for our experiments is a subset of the TDT4 corpus [193]. TDT4 consists of newswire and BN in three languages: English, Arabic and Mandarin. The subset used in these experiments consists of 12 CNN "Headline News" broadcast news shows. These broadcasts were manually segmented into 419 semantically homogenous stories. One human labeler generated a manual summary with a length of less than 30% of the original story for each of these. The labeler was asked to use words and phrases directly from the story in the summary whenever possible. These annotator-generated summaries were based on manual transcripts provided with TDT4. This resulted in training material comprising 419 human summaries of manually-segmented broadcast news (BN) stories.

We have access to ASR transcripts for these stories produced by SRI as a part of the DARPA GALE program [190]. These ASR transcripts contain automatically hypothesized words, their boundaries and confidence scores. Additionally, our system has access to automatically generated story boundaries [169] and automatic speaker segmentations (diarization) [233] for the 12 BN shows. The automatic story segmentation module produced 96 CNN stories. All of our summarization experiments are run on the automatically annotated and segmented stories, using only automatically generated words, word boundaries, and confidence scores.

We automatically align the manual summaries with the ASR transcripts to obtain the summary labels, that is, a binary label indicating whether word or phrase should be included in the summary or not. We use an alignment procedure based on minimum edit distance to align these summary labels originally annotated on manual transcriptions to ASR hypotheses.

The aligner finds the optimal match between the words of the manual summary with the ASR transcript words. The forced alignment of summary and ASR transcripts provides us with summary labels for each word in the ASR transcripts. We use these word level summary labels to generate summary labels for each candidate multiword segment. For example, to create an annotation for an intonational phrase, we count the percentage of aligned words in a given hypothesized phrase that appear in the human summary. If more than 50% of a segment – intonational phrase or otherwise – is aligned to a manual summary, it is labeled for inclusion in a summary, otherwise it was labeled for exclusion.

7.1.3 Speech Segmentation

In order to evaluate the use of intonational phrases in extractive speech summarization, we first produce candidate segmentations at different levels of granularity. Here, we describe these segmentations and the techniques used to generate them. We identify segments using automatically hypothesized intonational phrases, hypothesized sentence boundaries, and two pause-length thresholds – 250ms and 500ms. The number and length of each of these segmentations can be found in Table 7.1.

Segmentation	number per story	mean length (words)
250ms Pause	43.2	11.47
500ms Pause	19.1	25.97
Sentence	26.9	18.46
Intonational Phrase	71.2	6.96

Table 7.1: *Segmentation Statistics for Extractive Summarization of Broadcast News*

Pause-Based Segmentation

To generate pause-based segments, we calculate the pause duration between each pair of ASR-hypothesized words. We insert a segmentation boundary at every pause that exceeds a manually determined threshold. For these experiments, we construct two input segmentations — one using a 250ms threshold, another with a threshold of 500ms. Obviously, the set of boundaries selected with a 250ms threshold is a superset of those selected with a 500ms threshold. We hesitate to use a threshold below 250ms due to the potential confusion of stop gaps with phrasal boundaries which can be as long as 200ms [125].

Automatic Sentence Segmentation

We use an automatic sentence boundary detector developed at ICSI [124] to produce hypothesized *Sentence Units* (SUs). The sentence segmentation procedure provides both hypothesized boundaries and confidence scores for each hypothesis. This system is trained on human transcriptions of BN and combines information from both a language model and an acoustic/prosodic model. On automatically recognized speech, it operates with an error rate of 57.23%.

Intonational Phrase Segmentation

We use J48 decision trees, the weka [232] implementation of Quinlan's C4.5 algorithm [164], to hypothesize the location of intonational phrase boundaries. We classify each ASR word in the summarization material as either preceding an intonational boundary or not. This decision tree is trained using feature vectors containing only acoustic information: pitch, duration and intensity features. These acoustic features are described in greater detail in Section 4.3. The experiments on using intonational phrase boundary detection for extractive story segmentation were performed prior to the development of the best performing intonational phrase boundary detector described in Chapter 4. The impact of the improved intonational phrase boundary detection on summarization performance remains an open, and intriguing, research question.

The training for this decision tree classifier is the TDT material material; this material comprises approximately 20 minutes of annotated speech and 3326 hypothesized words. A more detailed description of this data can be found in Section 2.4. When evaluated on the manually labeled TDT material using ten-fold cross-validation, intonational phrase boundaries are predicted with 89.1% accuracy, and an F-measure of 0.665 (precision: 0.683, recall: 0.647). The intonational phrase boundary detection module used in these experiments is an early version of the best performing approach described in Chapter 4. In the experiments reported in Section 4.3, we find AdaBoosted single split decision trees to perform better than J48 classifiers. Also, the regression and narrow window reset features described in Sections 4.3.1 and 4.3.1 are not used in this intonational phrase detection model. Ten fold cross-validation using the best performing acoustic model predicts intonational phrase boundaries on the manually labeled TDT material with 91.8% accuracy with an F_1 of 0.737 (precision: 0.789, recall: 0.694). The potential impact of this improved intonational phrase boundary prediction on the summarization approach described in this section remains an area for future investigation.

7.1.4 Extractive Summarization Experiments and Results

We next train summarizers that extract segments based on the four segmentation units described in Section 7.1.3, namely, pauses of at least 250ms and pauses of at least 500ms, sentences, and IPs. We extracted features for the corpus for each segment type and assigned summary labels to each unit, using the forced alignments with manual summaries described in Section 7.1.2. We construct each of our four summarizers as a binary Bayesian Network classifier [99] where the summarizer's task is to determine whether a given segment should be included in the summary or not.

Feature Extraction

We use acoustic and structural features that were independent of the word identities hypothesized in the ASR transcripts. Using only acoustic/prosodic and structural features allows us to avoid the potential impact of word errors in the segments we extract. However, this approach is still effected by the ASR word boundaries which may introduce some amount of error.

We extract aggregated acoustic features over each candidate segment. We extract the minimum, maximum, standard deviation, mean of f0, Δ f0, RMS intensity (I) and ΔI over each segment. We also extract the z-score of the maximum and minimum within the segment over these four acoustic information streams. Using the pitch (f0) contours only, we extract three pitch reset features. These reset features are calculated by the difference between the average of the last 10, 5 or 1 pitch points in the current segment, and the average of the first 10, 5 or 1 pitch points in the following segment. We speaker normalize the f0 and intensity tracks using z-score normalization, and calculate each of these features using both raw and speaker normalized acoustic contours. We also include in the average word length within the segment in the feature vector. The raw pitch tracks were extracted using Praat's [20] 'To Pitch (ac)...' function, intensity tracks using 'To Intensity...'.

Using hypothesized story boundaries [169] and speaker turn boundaries provided by ICSI [233], we also extract a set of structural features. For each segment, we extract its length (in words and seconds), absolute and relative start time, relative position in the current speaker turn and relative position in the story. Additionally, based on the speaker identifications produced by the diarization module, we calculate the position of the current segment relative to all of the material spoken by its speaker.

Extractive Summarization Results

The results of 10-fold stratified cross validation evaluation of this extractive summarization approach are shown in Table 7.2. We find that the best summarization (evaluated using F_1) is obtained from the summarizer which uses intonational phrases (IPs) as its input segmentation

Segmentation	Precision	Recall	F-Measure (F_1)
250ms Pause	0.333	0.622	0.432
500ms Pause	0.255	***0.756***	0.381
Sentence	0.362	0.540	0.434
Intonational Phrase	***0.428***	0.650	***0.516***

Table 7.2: *Evaluation of Extractive Summarization using ten-fold cross-validation.*

unit. An example of two summaries of the same story generated by the extraction of intonational phrases and hypothesized sentence units can be found in Figure 7.1. The example hypothesized summaries generated by extracting sentence units and intonational phrases (IPs) are approximately the same length. However, the IP-based summary contains significantly more relevant information, information present in the reference summary. The obvious drawback of IP-based summarization is the coherence of the automatic summary. By concatenating relatively short units, the coherence of the resulting summary can suffer relative to other approaches which concatenate longer units. While the information contained in the IP-based summary is more relevant – as determined by F_1 and ROUGE – it may be useful to apply some *post hoc* editing to improve the readability of the automatic summary. The human readability of generated summaries has not been evaluated; the effect

Reference Summary

> Al Gore addressed the nation last night just hours after contesting the Florida results in state court. The Gore and Bush camps are facing filing deadlines and court appearances. In question are the hand recounts in specific Florida counties as well as the controversial butterfly ballots.

Hypothesized Summary using Sentence Unit Extraction

> Gore said all votes must be counted before a winner is declared his running mate joe lieberman echoed that theme today. And the election is not over until you if the gore and bush campaigns are facing several filing deadlines and court appearances today sir kit judge sanders sauls has scheduled a five p m hearing on an emergency motion by democratic attorneys. The state's highest court is requesting legal briefs explaining why it should or should not hear the case disputed ballots in palm beach county are getting yet another inspection this time its members of the conservative group judicial watch who are eyeing the ballots. They say they're looking for evidence of fraud and want to make sure the votes were counted correctly traditional watch members aren't allowed to touch the ballots so election officials must hold them up for examination.

Hypothesized Summary using Intonational Phrase Extraction

> Al gore took his battle for the white house to the american people last night. Televised. Address. Just hours. After contesting election results in a florida state court. Gore said all votes. Be counted before a winner is declared. His running mate joe lieberman echoed that theme today. Spoke to so eloquently. If the gore and bush campaigns are facing several filing. Sir kit judge sanders sauls has scheduled a five p m hearing. The group of voters behind the lawsuit argues the ballot caused confused. Gore supporters to vote for someone else. Disputed ballots in palm beach county are getting yet another inspection this time its members of the conservative group judicial watch. Who are eyeing the ballots. And want to make sure the votes were counted correctly

Figure 7.1: *Example summaries produced by the extraction of intonational phrases and hypothesized sentence units. Extracted units begin with capitalized letters and end with a period.*

of the extractive unit is evaluated strictly by automatic comparison to the manual reference summary.

The summaries produced represent a significant improvement of 8.2% in F_1 over sentence-based summarization. Note that, on average, there are about 2.75 intonational phrases for every sentence. This enables the summarizer to extract smaller segments for inclusion in summaries. However, the superior performance of the IP-based summarizer improvement is not simply due to its ability to extract smaller segments. When we compare the IP-based summarizer to the summarizer trained on 250ms pause-based segments, we see a considerable difference in F_1. Note that, while this pause-based segmentation does operate on more units than the sentence-based summarizer (there are 1.6 250ms-pause-based segments for each sentence) the sentence-based results are superior – by almost 5%. Thus, the extraction of shorter segments does not necessarily improve summarization performance. Using more linguistically meaningful units, intonational phrases, despite being somewhat errorful, provides the best summarizer performance.

F-measure evaluation assesses exact matches of predicted summary sentences to a labeled summary sentences. This measure is generally considered too strict for summarization purposes; a segment classified incorrectly as a part of a summary may be very close in semantic content to another sentence which *was* included in the gold standard summary. A measure commonly used in summary evaluation is ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [122]. ROUGE measures overlap between units of two summaries. We measure overlap between our gold standard and automatic summaries. ROUGE-N is calculated by evaluating n-gram overlap where n is a number of sequential words (cf. Equation 7.1). In addition to two ROUGE-N calculations, ROUGE-1 and ROUGE-2, we compute ROUGE-L, a ROUGE variant that measures the longest common subsequence between the initial document and the summaries.

$$\text{ROUGE-N} = \frac{\sum\limits_{S \in Ref.Sum} \sum\limits_{gram_n \in S} Count_{match}(gram_n)}{\sum\limits_{S \in Ref.Sum} \sum\limits_{gram_n \in S} Count(gram_n)}$$

Note that due to potential error in automatic story segmentation there may be a different number of automatic and manual stories for each show. In fact, there are more than 4 times as many manually annotated stories as those derived from automatic segmentation. Therefore, in order to evaluate the performance of the summarizer against human summaries, we compare summaries of entire shows against one another, rather than a story-by-story comparison. Thus, the target summaries for each of the twelve training shows are constructed by concatenating human summaries of manually-transcribed and manually-segmented stories. The automatic summaries of each show are concatenated automatic summaries of each automatically-segmented story.

Evaluation of this extractive summarization technique using ROUGE scores can be found in Table 7.3. This evaluation was also performed using stratified ten-fold cross-validation. We observe that the advantage of using intonational phrases (IPs) for extractive summarization is even more pronounced when we use ROUGE for evaluation. IP-based summarization ROUGE scores are higher than those obtained by extracting 250ms pause-based segments – 13.5% using ROUGE-1, 8%, using ROUGE-2 and 14.4% using ROUGE-L. Moreover, using ROUGE alternatives, IP-based summaries are dramatically better than sentence-based summaries. The improvements can be seen in all summarization scores, F_1, ROUGE-1, ROUGE-2 and ROUGE-L, confirming that hypothesized IPs are an excellent

Segmentation	ROUGE-1	ROUGE-2	ROUGE-L
250ms Pause	0.437	0.103	0.415
500ms Pause	0.440	0.128	0.412
Sentence	0.394	0.096	0.377
Intonational Phrase	***0.572***	***0.183***	***0.559***

Table 7.3: *ROUGE-based Summarization Results*

unit for extractive summarization of BN. A paired t-test shows the improvement of IP-based summarization over the next best competitor to be statistically significant whether evaluated under ROUGE-1 (p = 0.00337), ROUGE-2 (p = 0.0332) or ROUGE-L (p=0.00129).

ROUGE variants, like F-measure, are also imperfect measures of summarization quality. F-measure is a very strict measure of summarization performance; only those units that are present in a reference summary are considered relevant, semantically equivalent units that are not selected in the reference are considered irrelevant. On the other hand, ROUGE may err in the other direction; N-gram overlap considers two sentences to be semantically similar based on the similarity of their surface form. For example, "X was the winner" is considered to be 75% similar to "Y was the best" under ROUGE-1, while only 25% similar to "X came out on top". While neither F-measure nor ROUGE are ideal measures of summarization quality, by evaluating performance using both measures a more complete evaluation is possible. By demonstrating statistically significant improvements under F-measure – an overly strict measure – and ROUGE variants – an overly lax measure – we believe we show that extraction of intonational phrases leads to superior performance over the extraction of other units.

7.1.5 Extractive Summarization Conclusion and Future Work

In this section we present results of an empirical study evaluating the use of intonational phrases in extractive summarization of broadcast news. Based upon a comparison of four types of input segmentation, sentences, two pause-based segmentations, and intonational phrase segmentation, using a summarizers trained on the same corpus and using the same features, we find that IPs are the best candidates for extractive summarization. IP segmentation improves upon the second highest-performing approach, sentence-based summarization, by a 0.082 difference in F_1. Moreover, extracting intonational phrases produces a relative improvement of 45.2% on ROUGE-1, 90.6% on ROUGE-2 and 48.3% on ROUGE-L over the extraction of sentence units. We attribute the superior performance of the IP-based summarizer to the fact that IPs are typically shorter than sentences but are more linguistically and semantically meaningful units than naively defined pause-based segments.

This section evaluates the use of prosodic phrase boundaries for the segmentation of spoken material. In the future, we plan to evaluate the use of pitch accent detection and classification in extractive summarization. Salient phrases are likely to contain more accented words, and moreover, it is possible that important phrases are indicated with a distinct distribution of accent types. Bolinger observed that, when delivering a news broadcast, newscasters to accent important words more frequently than speakers do in ordinary speech. If this relationship between important information and accent behavior holds, hypothesized accent location and type information should be able to improve extractive summarization of broadcast news. Also, in addition to phrase segmentation, the classification of phrase ending intonation may be able to contribute to summarization decisions. Based on the success observed in using intonational phrase boundaries to improve extractive summarization of BN, we intend to explore the application of a broader set of prosodic event hypotheses to this task.

7.2 Story Segmentation

Broadcast News (BN) shows generally include a series of typically unrelated stories, with occasional commentary and commercials. Often each of the stories included in a broadcast are related only by virtue of their relevance at a particular moment in time – namely, the airing time. There is no semantic relationship between one story and the previous or following one. The stories are each “about” a different topic. The goal of story segmentation is thus similar to topic segmentation – identify where one story ends another begins. These boundaries between stories are commonly defined by changes in speaking style, speaker participation and lexical choice. The SRI NIGHTINGALE Y1 system searches a diverse news corpus to return answers to user queries. For BN, story segmentation is a necessary pre-processing step since information retrieval, information extraction, anaphora resolution assume the presence of single story “documents”, whether from text or from audio.

The use of prosodic events in the story segmentation task is very similar to that applied to extractive speech summarization in Section 7.1. Here we explore the use of intonational phrase (IP) boundaries to identify candidate story boundaries, where previously we examined the use of IP segmentation to identify candidate regions for extractive summarization. Through this investigation we compare the relative performance of different input segmentations in the story segmentation process. In previous work, [169], we defined potential story boundary segments as a subset of hypothesized sentence boundaries provided to our system by ICSI. However, since these boundaries can be errorful, we address this decision, testing whether story boundary detection can be improved by the use of IP boundaries to define candidate story boundary location.

In Section 7.2.1 we present a brief survey of previous approaches to story and topic boundary detection. We describe the story segmentation material in Section 7.2.2. In Section 7.2.4, we compare the performance of story boundary detection using IP boundaries as well as other possible segmentations. In Section 7.2.5 we conclude and discuss directions for future research. This work was first described in [171].

7.2.1 Related Work in Story Segmentation

The majority of previous approaches to story segmentation have focussed on lexical features, such as word similarily [107], cue phrases [148], cosine similarity of lexical windows [75, 61], and adaptive language modeling [16] to identify story boundaries, generally in text. Among these lexical approaches, it is rare for any multiword segmentation to be used to reduce the set of candidate boundary locations; each word boundary is considered to be a potential story or topic boundary. One exception to this is [148], where the candidate boundaries were prosodic phrase boundaries. BN story segmentation has included acoustic features in detection. These approaches often apply an initial segmentation to their source material. The set of candidate boundaries used by Shriberg, et al. [181] were pauses with duration greater than 650ms. Others [214, 169] have used an automatic sentence unit detection technique to construct a set of potential story boundaries. Work on non-English BN has generally combined lexical and acoustic measures, as [227, 114] for Mandarin and [147] for Arabic. These approaches considered each word boundary to be a candidate story boundary. Palmer [147] even went so far as to allow story boundaries to be placed within a word, using “multi-media events” which may be lexical, acoustic or visual to define the set of potential boundary locations. There is some evidence that intonational phrase boundaries can be used to aid topic segmentation, a task closely related to story segmentation [87].

7.2.2 Story Segmentation Corpus

The material used for our study is the TDT4 corpus [193], which includes newswire text and broadcast news audio in English, Mandarin and Arabic. A description of this material can be found in Section 2.4. Story boundaries were manually annotated on this material. In addition to the raw audio signal for each BN document, our story segmentation system has access to a number of automatically produced annotations, including automatic speech recognition transcripts with word boundaries [190] and inter-word durations, hypothesized sentence boundaries with confidence scores [124], and speaker segmentation (DIARIZATION) hypotheses [233].

We evaluate the use of intonational phrase boundaries to identify candidate story boundaries, along with the segmentations defined in Section 7.1.3. Recall that the hypothesized sentence boundaries have associated confidence scores. These boundaries are a significant source of error. In order to take advantage of some of the decision making processes that generated these boundaries, while reducing the impact of undetected boundaries, we evaluate the use of sentence unit (SU) boundaries identified by thresholding the confidence score of the hypothesis at 0.3 and 0.1.

In Table 7.4 we present statistics relevant to evaluating the 'goodness' of the candidate input segmentations. First, we calculate the percentage of manually annotated story boundaries that align with input segmentation boundaries (**Exact Coverage**). We also calculate the average distance in words from the gold-standard story boundary to the closest input segmentation boundary, a crude assessment of the minimum error introduced by the input segmentation (**Mean Alignment Error**) Not every segmentation boundary is exactly aligned with each manually annotated story boundary. In these situations, the task becomes identifying the candidate boundary *closest* to a story boundary. The Mean Alignment Error is the mean distance from a story boundary to the nearest candidate segment boundary. This measure represents a lower bound on the error of a story segmentation system using each set of candidate boundaries. Finally, we examine the ratio of target story boundaries to input segments (**Target Boundary Distribution**). In general, machine learning algorithms perform better on data sets with an even distribution of classes than those with heavily skewed class distributions. These numbers vary across language and broadcast; the aggregate statistics shown here demonstrate the gross behavior of the input segmentations.

Input Segmentation	Exact Coverage	Mean Error Alignment (words)	Target Boundary Distribution
Word	100%	0	0.48%
Hyp. SUs	68.3%	3.6	8.3%
SU thresh=0.3	74.4%	1.8	6.4%
SU thresh=0.1	82.9%	0.61	4.3%
250ms pause	83.5%	0.66	5.1%
500ms pause	71.8%	12.7	12.2%
Hyp. IPs	62.0%	1.1	2.6%

Table 7.4: *Input Segmentationstatistics for Story Segmentation*

7.2.3 Story Segmentation Approach

To detect story boundaries, we construct feature vectors of lexical and acoustic features for each candidate input segmentation as the unit of analysis. We use these feature vectors to train decision tree classifiers specific to each show using J48, weka's [232] implementation of C4.5 [164]. That is, we build unique models for ABC's "World News Tonight" and CNN's "Headline News". This style of show-specific modeling has been shown to significantly improve story segmentation accuracy [169, 181]. For training purposes, we match each manually annotated story boundary to its closest preceding input segment boundary. These 'matched' input segment boundaries represent the set of 'boundary'-class data points for classification. Note that we have ToBI labels for a single show. It would be interesting to evaluate if intonational phrase boundary detection is improved by show-specific modeling, as story segmentation is.

Lexical Features

To capture lexical cues to story boundaries, we extract LCSeg [61] hypothesized segments and TextTiling [75] coefficients based on window sizes of three, five and ten segments preceding and following the current boundary. TextTiling and LCSeg have been shown to be useful in topic segmentation in text news documents and meeting transcripts. We also compute features based on lexical consistency immediately or following story boundaries from those lexical items, for each show, that are statistically likely to occur within a three, seven or ten word window preceding or following a story boundary.[1] For English BN these lexical items are stemmed using an implementation of the Porter Stemmer [158]. We include in our feature vector the number of words that occur in a three, seven, or ten word window preceding or following the current boundary that also occur on the corresponding keyword list. Note that we do not include the identity of these words in the feature vector, only the number of matches. For English BN, we also include the number of pronouns in the segment preceding each boundary, identified by a part-of-speech tagger based on the Brill tagger [30]; our use of this feature is based on the hypothesis that a speaker may begin or end a story by identifying themselves with a pronoun – e.g. “I’m X reporting live for CNN” – , or more generally that pronoun use may change over the course of a story, e.g. persons may be more likely to be referred to by a pronoun at the end of a story, where their identity may already be established.

Acoustic Features

Acoustic information has been shown to correlate with story boundaries [181, 169], topic shift [87] and changes in discourse structure [89], so we include such features in our detection of story boundaries. We extract the maximum, minimum, mean, median, standard deviation and mean slope of pitch, and intensity from the segment immediately preceding the current boundary. Based on speaker diarization output, we also extract these features based on speaker (z-score) normalized f0 values. We include in the feature vector the length of the segment. In addition to these, we calculate the difference of the above features extracted from the segment preceding and the segment following the current boundary. We also extract features based on speaking rate, hypothesizing that segments at the end of stories will be spoken at different rates and that vowel length may be prolonged preceding boundaries. These features include frame-based speaking rate (ratio of voiced to unvoiced frames), mean

[1] Statistical significance is determined using χ^2 with a threshold value of 20 for inclusion in the list of keywords.

vowels per second, mean vowel length, and lengths of segment final rhyme and segment final rhyme. Each feature is also speaker normalized and, when possible, is normalized by vowel identity. We also extract differences in these values across each candidate boundary.

These acoustic features may capture some of the prosodic variation represented by accent rate and type information. However, in these experiments we do not use hypothesized categorical prosodic events other than intonational phrase boundaries.

Structural Features

To capture structural consistencies in each news broadcast, such as the airing of commercials or regularities in story length, we include the relative position of a candidate boundary within the show in our feature vector. We also calculate a set of features based on each identified speaker's participation in the current show. In some shows, story boundaries often co-occur with speaker boundaries. In others, one story is closed and another begun by the same (anchor) speaker. To capture such patterns we extract three binary features: Is the current segment boundary also a hypothesized speaker boundary? Is the word immediately preceding the current boundary this speaker's first spoken segment in the broadcast? last? We also include in the feature vector the percentage of segments spoken by the speaker of the segment immediately preceding the current boundary.

7.2.4 Story Segmentation Results and Discussion

Results of story boundary detection based on our different input segmentations is shown in Table 7.5. All results are based on ten-fold cross-validation evaluation. We evaluate these using the WindowDiff measure [149], an extension of Beeferman's P_k [16]. The WindowDiff score is incremented for each false alarm and each miss in a hypothesized segmentation such that near-errors, where a hypothesized boundary is placed close to a target boundary, incur a lesser penalty than more egregious misses or false alarms. Therefore, lower Window Diff scores represent better segmentations. The appropriate window size for applying both WindowDiff and P_k is approximately one half the length of the average segment. In the TDT4 corpus stories have a mean length of 215.9 words. We thus evaluate using WindowDiff a window size of 100.

The story boundary detection model produces a story-boundary/non-story-boundary prediction for each input segment. As each input segmentation defines a different data set, we need to insure that the evaluations of these data sets are comparable. To do this, we align every set of input segment-based predictions to the word level. This allows us to apply the WindowDiff evaluation technique equivalently to the results of story boundary detection based on each input segmentation, and determine which demonstrates the best segmentation performance.

Input Segmentation	English	Arabic	Mandarin
Word	0.300	0.308	0.320
Hyp. SUs	0.357	0.361	0.278
SU threshold=0.3	0.324	0.318	0.258
SU threshold=0.1	0.308	***0.304***	0.253
250ms pause	***0.298***	0.312	***0.248***
500ms pause	0.344	0.419	0.295
Hyp. IPs	0.340	0.333	0.266

Table 7.5: *Story Segmentation Results evaluated using WindowDiff with k=100*

Hypothesized intonational phrases do not represent the best candidate boundary for story segmentation. Pause based segmentation and low (0.1) threshold sentence unit (SU) boundaries yield better story segmentation performance. However, the use of intonational phrase boundaries is able to improve segmentation performance over that obtained by hypothesized SU boundaries, using a 0.5 threshold. While IP boundaries do not represent the best candidate segmentation, they do carry some useful information for story segmentation. We clearly hesitate to make any claims about the success of the hypothesized intonational phrase boundary detection in identifying phrase boundaries on Arabic and Mandarin material; the acoustic indicators of phrasing may be significantly different in these languages from the English training material. However, we note that hypothesized IP boundaries predict story boundaries with greater success than hypothesized SUs across all languages, despite the use of English IP detection models for Arabic and Mandarin speech.

Across all languages we find that hypothesized SU boundaries using the default threshold confidence level fail to produce the best story segmentation. SU boundaries detected with lower confidence (0.1) perform best for Arabic, while boundaries detected from 250ms pauses perform best for English and Mandarin. However, note that a simple word-based segmentation produces surprisingly good results; while not the best performing for any language, they are second best in English and Arabic. In general, our results show that shorter input segmentations tend to produce better results. We expected the contextual information captured in the feature vectors extracted from larger segmentations to be highly discriminative of story boundaries. However, these large segmentations introduce a significant amount of error based on their misalignment with target story boundaries. The smaller input segmentations are less affected by errors in the input segmentation. Despite using features with a narrow view of the source data, these segmentations are able to produce the best story boundary predictions, probably as a result of this small amount of baseline error.

7.2.5 Story Segmentation Conclusions

In this section we evaluate the use of intonational phrase (IP) boundaries and other input segmentations to define candidate boundaries for story boundary detection in English, Arabic, and Mandarin. These input segmentations include hypothesized sentences defined using a number of confidence thresholds, pause-based segmentations, as well as hypothesized intonational phrases. These experiments indicate that, in general, shorter input segmentations produce better story segmentations, with the best results being produced by low (0.1) thresholding of sentence unit (SU) hypotheses and short (250ms) pause-based segmentations. Intonational phrases perform better than standard SU hypotheses (threshold = 0.5) on all languages, but fail to improve story segmentation performance over that obtained by short pause segmentation on any languages.

The identification of intonational phrase boundaries to define candidates for story segmentation represents one way in which prosodic event information can be applied to the task of story segmentation. Accent location and type hypotheses probably contribute to perceptions of story changes. At the start of a story or topic, many items are new to the discourse, and are therefore likely to be accented. Towards the end of the story much of the topic material will have been introduced, leading to a lower accent rate, and likely a different distribution of accent types. Thus, in theory, accent rate information should be helpful in the detection of story boundaries. This hypothesis remains to be evaluated. Moreover, it is unlikely that English BN story boundaries co-occur with the phrase final intonation that is commonly associated with a sense of incompleteness, namely, L-H% and H-L% tone combinations. Also it is likely that story boundaries *do* occur with L-L% phrase final intonation – a typical indicator of finality. Hypotheses of phrase ending intonation may be able to contribute to the accurate detection of story boundaries by eliminating boundaries with this intonation from the set of candidates.

7.3 Non-native Intonation Assessment

Humans acquire language proficiency early in life. The first language a speaker acquires is called their native tongue, native language or L1. Frequently, and perhaps more frequently at this point in history than ever, at some later point in their life, human speakers may learn a second language or L2. When acquired later in life, L2 proficiency is often poorer than L1 proficiency. Frequently, influences of L1 characteristics are manifested in the language competence of L2. These can be realized as difficulty in the perception and production of particular phonemes, use of syntactic constructions, or in the display and interpretation of prosodic variation. Moreover, members of a language group have the ability to distinguish L1 speakers from L2 speakers.

For some speakers, a foreign accent can be a mark of cultural identity. However, heavily foreign accented speech can be more difficult to understand leading to less reliable communication [215]. Moreover, low proficiency in pronunciation and suprasegmental delivery can lead to negative perceptions of a speaker [63, 123]. Also, automatic speech recognition systems perform significantly worse when speech bears the markers of a foreign accent [60]. L2 speech frequently differs from L1 speech in both phonetic and intonational characteristics. While the pronunciation differences that are associated with phonetic differences are commonly addressed in language instruction courses, production of native sounding intonation is rarely part of L2 instruction.

In this section, we explore the use of prosodic events in identifying non-native intonation. This identification can be used for two purposes. Identifying points of non-native intonation can allow an automated language instruction system to provide informative feedback to an L2 learner, or to assess language proficiency. Identifying non-native speech can also be used to adapt or select acoustic, pronunciation and/or language models for automatic speech recognition systems.

There have been significant efforts to assess spoken language competence of non-native speakers. Many of these have focused on segmental pronunciation ([56, 45] *inter alia*), though some have incorporated prosodic information of non-native speech in proficiency assessment. Rhythmic properties have shown a significant difference between L1 and L2 speakers. L2 speakers typically speak with a slower speaking rate and insert pauses in their speech more frequently [46, 140, 204]. Teixeira et al. [205, 204] examined the use of global pitch features for assessment of a speaker's fluency or nativeness. In predicting nativeness scores, they extracted features based on pitch slope, pitch maxima, the location of maximum pitch and a representation of the variation of pitch. While they found only weak correlations between these pitch features and nativeness scores (r < 0.434), they did find that overall assessment of nativeness was improved by including pitch features with lexical and speaking rate features. They report a relative improvement of 7.7% in the correlation between human and machine nativeness scores. In his doctoral thesis, Liscombe [123] examined the use of ToBI tones to automatically predict nativeness scores. In this work, the rate of phrase boundaries and the relative frequency of phrase accent and boundary tone combinations were examined, along with measurements of the distance between subsequent instances of the same phrase ending intonation. Correlations between these ToBI-based features and human scores of fluency were reported. He found the rate (r=-0.46) and distance between phrase boundaries (r=0.38) to show significant correlations with nativeness ($p < 0.001$). Moreover, the distance between subsequent H-H% (r = 0.53) and L-H% (0.47) tone combinations also showed a significant positive correlation. This work provides evidence that the frequency and type of prosodic events that a speaker uses has an impact on the perceived fluency of his or her speech.

Tepperman, et al. [206] presents work that is similar to the research described in this section. In this work, Tepperman and colleagues use a tone recognition model to assess the fluency of an utterance. The tone model is an HMM that recognizes sequences of intonational tones from an inventory containing a reduced set of ToBI tones – %H, H*, L*, H% and L%. The acoustic model is trained on material from a single BURNC speaker and generates hypotheses using pitch and intensity features extracted from 10ms frames. Three bigram tone models, which were to constrain the HMM transitions, were trained on manually annotated tones from SAE (BURNC) and British English (IViE) data. One bigram model was trained on only SAE material, another on only British English data, and the third was trained on all of the data. The authors found that the posterior of this tone recognition model correlated with human ratings of "pronunciation" when evaluated on non-native material from the ISLE corpus. The human raters were asked to take a speaker's fluency, rhythm and articulation into their scoring. The best reported correlation between model posteriors and human pronunciation ratings (0.331) was obtained using the bigram tone model trained using both American and British English material.

While related to [206], both use tone sequence modeling to assess the nativeness of speech, the research described in this section is different in two significant ways. First, we evaluate the use of two tone sequence models, one trained on L1 speech and a second on L2 speech, for this task. Rather than measure the goodness of fit with a single L1 model, this approach allows us to determine whether the observation was more likely generated by an L1 or L2 model. We also evaluate the use of a single L1 model as in [206], but find that this model captures effects of genre in addition to nativeness (cf. Section 7.3.3). Second, rather than measure the correlation to human pronunciation or fluency ratings, we opt to classify regions of spoken material as "native" or "non-native". This impacts the evaluation of the two approaches more than their functionality. In the assessment of L2 speech, scores are used to measure the nativeness, or fluency, of the speech as a measure of degree. By using a classification approach, the system we describe in this section can identify speech as "non-native". In a language instruction domain, this identification can be used to give feedback to a language learner. In the context of speech recognition system, for example, such a decision can be used to select or adapt acoustic, pronunciation or language models to improve recognition performance on L2 speech.

Previous work in intonation assessment has frequently been used to automatically score the fluency or nativeness of speech. The work presented in this section, on the other hand, discretely classifies speech as native or non-native. This approach is relevant to either adapt or select spoken language processing modules that are specifically suited to handle non-native input or to identify regions of 'non-native' intonation for instruction. The approach we take in this work is similar to that described by Tepperman, et al. [206]. We assess the use of a bigram tone model to model native and non-native intonation. This evaluation serves to decouple the performance of automatic prosodic event detection and classification performance from the tone sequence model. By applying this approach in a classification rather than assessment framework, this approach can identify specific regions of intonation that are 'non-native'. While previous work scores the fluency of an input utterance, this identification approach can be used to provide feedback about specific regions of non-native intonation.

This section is structured as follows. In Section 7.3.1, we describe a corpus of English material spoken by native Mandarin Chinese speakers. We describe the prosodic event characteristics of this corpus in Section 7.3.2. In Section 7.3.3, we describe and evaluate two proof-of-concept approaches towards distinguishing native and non-native speech using sequences of prosodic events. We conclude and present directions for future work in 7.3.4.

7.3.1 Material

We have collected a corpus of read speech from 12 (6 male, 6 female) native Mandarin Chinese (MC) speakers. At the time of collection, the speakers ranged from 20 to 35 years of age and had studied English for 6 to 23 years. One speaker had lived in the United States for 17 years, the remaining from 2 months to 5 years. Material was recorded in a sound-proof booth at Columbia University using a Tascam HD-P2 solid state recorder at 16 bit with a 44.1kHz sampling rate and a Crown CM-311 headset microphone.

We selected material from the Boston University Radio News Corpus (BURNC) (cf. Chapter 2) for this data collection. The BURNC material was selected in order to compare prosodic variation within and across speaking groups. Each story was read by multiple native Standard American English (SAE) speakers as part of the BURNC material. This collection of multiple native Mandarin Chinese speakers allows us to examine what dimensions of variation are common within native SAE speakers and which are unique to native Mandarin Chinese speakers. Subjects were asked to read an introductory transcribed broadcast news paragraph concerning NASA and a delayed spacecraft launch, to acclimate them to reading aloud and to the recording environment. We next had subjects read labnews stories **p** – computerized parole officers – and **r** – the Safe Roads Act. While story **j** – Massachusetts supreme court justice – has more available prosodic annotations in BURNC, it contains a high rate of proper names which caused difficulty for pretest subjects. We decided not to use story **t** as its subject was teen pregnancy, sex and contraception. We were concerned that this topic might make some subjects uncomfortable, modifying their intonation in unexpected ways. In total we collected 105.2 minutes of non-native read spoken material.

The experiments described in this section are run on a subset of this corpus. Material from the two BURNC stories from four speakers (2 male and 2 female) has been ToBI labeled by the author. The introductory paragraph has not been transcribed or annotated. This annotated material comprises 37.6 minutes of speech. At the time of recording, the four speakers of the annotated material were between 25 and 30 years old, with 6 to 19 years of experience with English; they had spent from 7 months to 3 years living in the United States.

7.3.2 Non-native Tone Distribution

We hypothesize that native SAE speakers and native MC speakers of English display distinct pitch accent and phrase boundary behavior. In this section, we compare the accent rate, distribution of pitch accent types, mean phrase length and distribution of phrase ending intonation of SAE and MC speakers. The set of annotated MC material we analyze is described in Section 7.3.1. We compare the prosodic event distributions observed on the MC data to those observed in the corresponding BURNC material and the BDC-read data.

We find that MC and SAE material have very different accent rates. In the collected material, non-native speakers accent 61.9% of words. This is a much higher rate than we observe in either corpora of SAE speech; 42.17% of BDC-read words are accented, while 52.0% of BURNC words bear accent. While L1 and L2 speech demonstrate quite different accent rates, the two L1 speech material differ greatly. Speech genre – read or broadcast news – influences accent rate to a degree as large as the speaker's native tongue does. It has also been observed elsewhere that L2 speech typically contains shorter phrases than L1 speech (cf. [123, 205], *inter alia*). Since intermediate phrases contain at least one accented word, shorter phrases may correlate with increased accent rates. We examine the average

	H*	!H*	L+H*	L+!H*	L*	L*+H	L*+!H*	H+!H*	X*?
MC	59.7	17.8	12.1	0.9	1.8	0.9	0.0	6.5	0.2
BURNC	38.1	10.7	19.9	3.8	3.7	0.4	0.0	3.5	19.8
BDC-read	47.3	29.4	12.2	1.3	5.8	1.3	0.0	0.7	2.0

Table 7.6: *Relative frequency (%) of accent type usage on MC, matching BURNC and BDC-read material.*

intermediate and intonational phrase length of each corpus. Results of this analysis of L2 material is consistent with other observations of non-native speech. The mean intermediate phrase length in the L2 material is 2.55 words compared to BURNC length of 3.87 and a BDC-read length of 5.03 words. The mean intonational phrase length shows an even greater difference between L1 and L2 speech. The L2 material has a mean intonational phrase length of 3.83 words, while the BURNC material has a mean of 6.16 and BDC-read 7.74. The difference between each of these means is significant with $p < 1 * 10^{-20}$ as determined by a t-test.

We also find that there are observable differences in the use of pitch accent, phrase accent and boundary tone types across L1 and L2 speech. The distributions of pitch accent types in the MC, matching BURNC and BDC-read material can be found in Table 7.6.

The distribution of pitch accents used in L2, BURNC and BDC-read material show some interesting differences. The clearest differences are the greater use of H* and H+!H* accent type in L2 speech and the decreased use of L* accents. It is possible that the relative infrequency of L* is due to the fact that there is no tonal equivalent in Mandarin Chinese, the speaker's native tongue. C-ToBI is a system for describing the prosody of Standard (Mandarin) Chinese [118]. Mandarin Chinese contains four lexical tones – and a neutral tone. These four tones are a sustained high, a rising tone, a fall rise and a falling tone. While C-ToBI has a mechanism for indicating that words are produced in a compressed pitch range, all four Chinese lexical tones contain a 'high' tone component. Also, as we observe in the BDC and BURNC material, accenting a word with L* accents is relatively uncommon in native speech. L* accents are often associated with the accenting of discourse

	L-	H-	!H-	X-?
MC	69.6	28.7	1.6	0.0
BURNC	68.1	15.9	14.6	1.4
BDC-read	77.1	19.0	3.9	0.0

	L-L%	L-H%	H-L%	H-H%	!H-L%	Other
MC	46.9	24.3	19.3	7.5	2.0	0.0
BURNC	57.6	34.5	4.6	0.4	0.0	2.9
BDC-read	49.0	35.6	9.6	1.4	4.3	0.0

Table 7.7: *Relative frequency (%) of phrase ending tone usage on MC, matching BURNC and BDC-read material.*

given tokens, words that refer to concepts that have already been introduced to the discourse. This intonational behavior may be difficult for L2 speakers to acquire due to the lack of low-toned prominence in their L1 and a lack of L1 stimuli from which to generalize. The increased relative frequency of H+!H* may also be due to the influence of the L2 speaker's native tongue. While this pitch accent type is relatively infrequent in native SAE speech, the fourth lexical tone in Mandarin Chinese is a falling tone – similar in shape to the H+!H* pitch accent type. The increased use of H+!H* pitch accents by native speakers of Mandarin Chinese may be due to the influence of their native tongue, and their familiarity with phonologically salient falling pitch.

We also observe differences between BURNC and BDC-read material. Broadcast news speech demonstrates some idiosyncratic intonational characteristics – including an high rate of accenting "given" items and function words and a lower than expected rate of accenting "new" items [24]. These idiosyncrasies make it somewhat unsurprising that non-professional read speech differs with respect to relative frequency of pitch accent, phrase accent and boundary tone usage. We discuss these differences in more detail in Chapters 5 and 6.

The distributions of phrase ending tones, phrase accents and boundary tones on the MC, matching BURNC and BDC-read material is reported in Table 7.7. C-ToBI does not have a correlate to the phrase accents in the standard ToBI system. Phrase ending intonation is described using only two boundary tones, high (H%) and low (L%). It is therefore possible that MC speakers may have difficulty with distinguishing the high rise (H-H%) from shallow rise (L-H%) when speaking English.

In fact, when we examine the distribution of phrase accent and boundary tone pairs used in the different corpora, we find that the MC speakers used a significantly higher rate of H-H%. In addition to maintaining the turn and indicating continuation, rising intonation is often used in SAE to indicate hesitation or uncertainty. If this were true in MC intonation as well, the increased rate of H-H% could be explained by lack of confidence held by the MC speakers in their English reading proficiency. Of course, more knowledge of how rising tones are used and how phrase final intonation varies in MC is needed to support this possible explanation. The other major difference between MC intonation at phrase boundaries is the increased use of the plateau, H-L%. The high sustained pitch that is characteristic of the plateau phrase ending intonation is similar to the first lexical tone in Mandarin Chinese. This increased H-L% use in the MC data may another example of an interaction between the tonal qualities of the speaker's native tongue and L2 intonation. Perception and production studies are required to determine how MC speakers perceive and use phrase ending intoantion. The use and perception of H-L% events, and perception and production of SAE rising phrase final tone combinations, L-H% and H-H%, are two phenomena that warrant further investigation.

We observe differences in the use of prosodic events in English read produced by native Mandarin Chinese speakers, professional native American English speakers, and non-professional native American English speakers. It remains to be seen if these differences are significant enough to be able to identify non-native speech by virtue of the prosodic events used by a speaker. We have found some differences in native SAE intonation and L2 intonation that may be explained by L1 interactions in MC speakers. It is possible that a MC speaker's intonation while reading English may be particular to tonal qualities found in Mandarin Chinese. The findings on non-native production of English reported in this section are unique to Mandarin Chinese production of English. The expected performance of the approaches and analyses presented in this section when applied to non-native speech from other L1 groups remains uncertain, and worth future investigation.

7.3.3 Sequential Modeling for Non-native Intonation Assessment

Having observed differences in the rate and type of prosodic events used by SAE and MC speakers in Section 7.3.2, we now turn our attention to the task of automatically identifying non-native intonation. There are two potential goals in the identification of non-native intonation. Spoken language processing (SLP) tasks, such as automatic speech recognition, have difficulty operating on foreign accented speech. If this effect of native tongue can be identified, SLP behavior can be modified to accommodate this effect, for example, models can be adapted or selected, or calls could be routed appropriately. On the other hand, this assessment can be used for language instruction. The ability to identify aspects of a speaker's production that may be non-native sounding can be used to either provide feedback in an instructive capacity or perform assessment of language proficiency.

In this section, we describe two experiments to identify speech as native (SAE) or non-native (MC). In both of these, we train bigram models of ToBI tone sequences to model the intonation of SAE and MC speakers. The likelihood of a sequence of k ToBI tones (pitch accents, phrase accents and boundary tones) T, is calculated by Equation 7.1.

$$P(T) = \prod_{i=1}^{k} P(T_i|T_{i-1}) \tag{7.1}$$

The calculation of "per tone" tone model perplexity is performed by the formula in Equation 7.2.

$$2^{-\frac{1}{k}\log_2 P(T)} = 2^{-\frac{1}{k}\sum_{i=1}^{k}\log_2 P(T_i|T_{i-1})} \tag{7.2}$$

Perplexity is a calculation of the average number of bits required to encode a single tone under a given model. Sequence likelihood is sensitive to k; specifically, longer sequences have lower likelihoods. The calculation of perplexity normalizes for sequence length, determining the expected number of bits required to encode a single tone, rather than the code length of an entire sequence.

In the first experiment, we apply a single model of SAE intonation to assess nativeness. We train a sequential model on manually annotated SAE tone sequences, and assess the model by evaluating the model perplexity when applied to an evaluation set of SAE data and a set of MC data. The MC evaluation data are the tone sequences of material described in Section 7.3.1. The SAE evaluation data is the BURNC material corresponding to the stories read by the MC speakers – BURNC labnews stories **p** and **r**. By using evaluation data from consistent lexical material, we can evaluate the difference between speaking style across speaker groups without any bias due to lexical content. We hypothesize that the SAE evaluation material will have lower model perplexity than the MC evaluation material. If this hypothesis is true, we can identify non-native speech using intonational information, represented as a ToBI tone sequence.

We find that if we evaluate the tone sequence represented by a full story – approximately 212 seconds of SAE speech and 282 seconds of MC speech – the per tone perplexity of SAE speech is 3.79 ± 0.08 bits, while the per tone perplexity of MC material is 4.28 ± 0.16 bits. In order to classify a tone sequence as 'native' or 'non-native', a perplexity threshold must be identified. We identify such a threshold using logistic regression. We generate perplexity values for the evaluation sets of SAE material and MC material. Then, using ten-fold cross-validation, we use logistic regression to identify a threshold, and evaluate the native/non-native classification accuracy. To avoid any sampling bias during this cross-validation evaluation, we use *undersampling* to enforce an equal distribution of SAE and MC perplexity values. When we compare the per tone perplexities calculated over full stories, we find a 68.75% classification accuracy; 11/16 tone sequences are correctly classified.

Recall that, the SAE intonation model in this evaluation is trained on BURNC material. The BURNC material is spoken by professional speakers reading broadcast news. This genre of speech has an idiosyncratic speaking style [24]. Therefore, it is possible that the tone sequence model is capturing the broadcast news style of intonation rather than artifacts of a speaker's native tongue. To assess this possibility, we evaluate the SAE and MC evaluation material on a tone sequence model trained on BDC-read material (cf. Chapter 2). The BDC-read corpus comprises read speech produced by non-professional speakers. We find the same cross validation accuracy in the classification of tone sequences extracted from full stories 68.75% – correct classification of 11/16 sequences. However, when we examine the mean perplexities, we find that the SAE material has an expected bit length of 4.82 ± 0.26 per tone, while the MC data requires 4.19 ± 0.32 bits per tone. That is, the non-native material fits the BDC model better than the native material. This indicates a strong genre bias on this technique. Genre, professional broadcast news speech versus non-professional speech, has a more powerful impact on tone sequence modeling than the native tongue of the speaker.

These experiments attempt to use a single model of SAE intonation in order to classify speech as native or non-native. However, we find that influences other than the nativeness of the speaker – specifically genre – can lead to poor modeling performance. In the second set of experiments, rather than using only a model of SAE intonation, we use two bigram tone models to classify material as native or non-native, one trained on SAE material and the second trained on MC material. Rather than identify a perplexity threshold on a single SAE model, in this scenario we evaluate the perplexity of a given tone sequence on both the SAE and MC tone models, classifying the sequence based on the model which fits the sequence better, i. e. yields lowest perplexity.

In the first set of experiments, we compare the MC material to the matching SAE material – the BURNC material with identical lexical content. The evaluation is structured as follows. We use a leave-one-speaker-out evaluation structure. No material from a test speaker is used in the training of the models. This removes the modeling of any individual speaker bias. Using the training data from remaining speakers, we train two models, one using SAE training data and another using MC training data. At evaluation, we compare the perplexity of the SAE model to the perplexity of the MC model. If an evaluation sequence has lower SAE perplexity than MC perplexity is it classified as 'native' and vice versa. We perform this evaluation using two sets of SAE material: 1) matching BURNC material, and 2) BDC-read material. The matching BURNC material has the advantage that the lexical content of the SAE and MC material is identical. This eliminates lexical items as a potential source of bias. However, from the previous experiment, we know that genre is a significant source of bias in this evaluation. It is possible that classification of native versus non-native speech using this evaluation set is, in fact, an assessment of 'broadcast news' speaking style versus 'non-professional' style. While the MC material is read news material, the BDC material is read direction giving monologues.

We find that this evaluation technique is very successful in distinguishing SAE from MC material when evaluated using manually annotated tone sequences. If we evaluate using a whole story from the BURNC or MC material, we can correctly classify 96.92% of stories as 'native' or 'non-native'. When we perform this experiment using the BDC-read material, this approach achieves 100% accuracy. This confirms the hypothesis that native and non-native speech can be differentiated by prosodic events.

However, this evaluation requires a significant amount of evaluation material. The current result is obtained by the tone sequence derived from roughly five minutes of speech. Therefore, we also approximate the performance of this technique with a smaller amount of material. Evaluation on a smaller test sample is important to determine the data requirements of this approach. Moreover, in online ASR applications, such as spoken dialog systems, a decision must be made as quickly as possible. In such applications, high accuracy classification of a small amount of input material is necessary. Moreover, identification of narrow regions of 'non-nativeness' is helpful for language instruction. It is helpful to direct a user to a small span of words where an intonational error occurred and provide feedback about the mistake. Thus, for each evaluation speaker, rather than evaluate each story as a whole, we generate a tone sequence from each sequence of M words, and classify these short tone sequences as 'native' or 'non-native'.

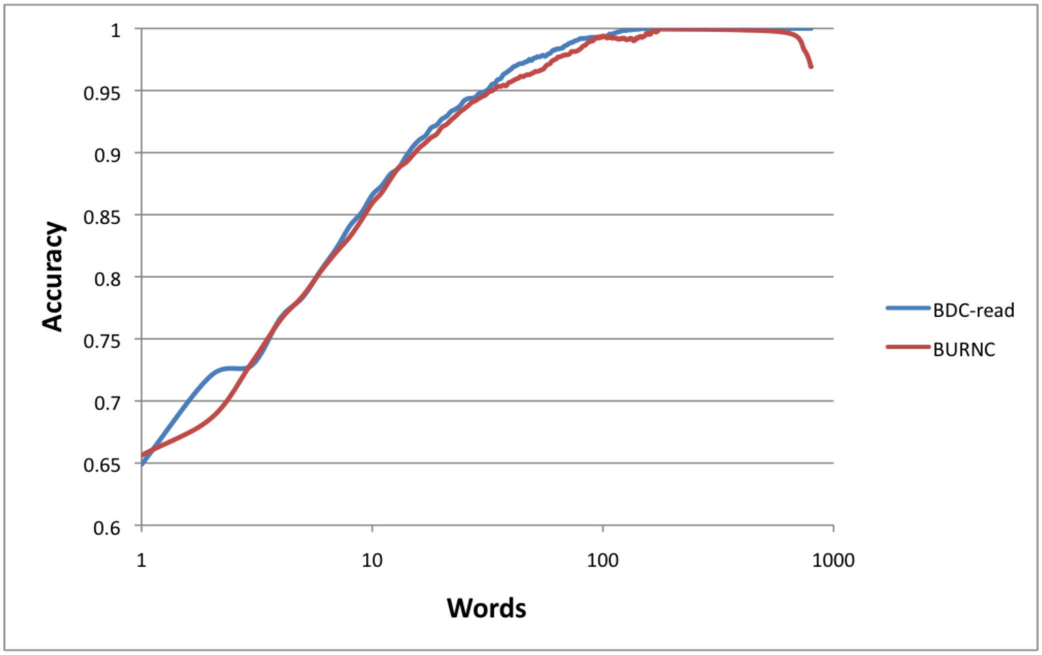


Figure 7.2: *Accuracy of native versus non-native classification using ToBI tone sequences derived from a variable number of words. Native evaluation material is either BDC-read or BURNC material.*

Classification accuracy is reduced when examining shorter spans of intonation. This phenomena occurs for two reasons. First, there is less information to be evaluated and therefore more noise. Second, not every non-native production will contain evidence of a foreign native tongue. The likelihood of a short utterance being indistinguishable from native speech by virtue of prosodic events alone is fairly high. Figure 7.2 contains the nativeness classification accuracy comparing MC material against both matching BURNC and BDC-read material, varying the amount of available evaluation material in one word increments. The majority class baseline classification accuracy using the BDC-read material is 68.9% while the BURNC baseline is 54.3%.

The first thing we notice from this evaluation is the similarity of performance using both SAE corpora. In the first experiment, we found a significant difference in the sequential modeling behavior whether the model was trained on BDC or BURNC material. Here we find no major observable difference in the SAE versus MC classification performance due to the training corpus. Second, we observe the expected degradation of performance as the observation window size decreases. Using a window of only 15 words, we observe accuracies over 90% on both corpora. Accuracies of over 80% are achieved with six word sequences using models trained on either BDC and BURNC material. The mean word length on the MC material is 359 milliseconds. Thus, 80% accuracy is achieved with analysis of approximately 2.16 seconds of speech, and 90% is demonstrated with 5.38 seconds.

Recall that these results are generated by tone sequence models that are trained and tested on manually annotated ToBI labels. To measure the degradation introduced by substituting automatic annotations for human labels on this task, we generate hypothesized tones for the MC material, and the BDC-read corpus. We generate hypotheses for all of this material using prosodic event detectors and classifiers trained on BDC-read material. The BDC-read hypotheses are generated using cross-validataion, while the MC hypotheses are generated from classifiers trained on the full BDC-read set. The MC material and BDC-read corpus both comprise non-professional read speech. The MC material is in the news domain, while BDC-read is made up of direction giving monologues. To limit the impact of domain, we use prosodic event detectors and classifiers that use only acoustic information – no lexical or syntactic features are used. The generation of hypothesized tones uses six classifiers, each of which have been described in detail in the thesis: 1) pitch accent detection, 2) pitch accent classification, 3) intonational phrase boundary detection, 4) intermediate phrase boundary detection, 5) classification of intonational phrase final intonation – pairs of phrase accents and boundary tones and 6) phrase accent classification at intermediate phrase boundaries. In general, we use the best performing acoustic model as reported throughout the thesis. However, due to the sensitivity of Quantized Contour Modeling (cf. Section 5.6.3) to its model parameter settings, we do not use this modeling technique in generating hypotheses for this application. The pitch accent detector uses Corrected Energy Based Classifiers combined using weighted majority voting (cf. Section 3.6). Both the intonational and intermediate phrase boundary detectors use AdaBoost with single split decision trees with acoustic features extracted from the word preceding a candidate boundary and the difference

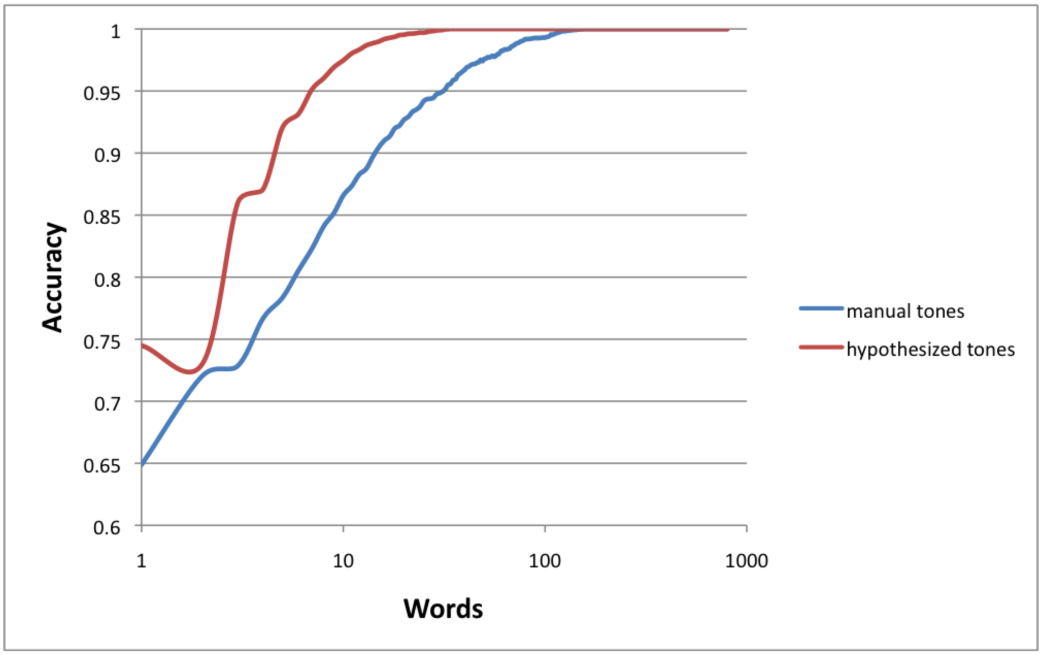


Figure 7.3: *Accuracy of native versus non-native classification using hypothesized ToBI tone sequences derived from a variable number of words. Native evaluation material is BDC-read.*

of features across each boundary (cf. Sections 4.3 and 4.5). To classify pitch accents we use ensemble sampled SVM classifiers using acoustic aggregations and feature to represent pitch contour shape (cf. Section 5.6.4). We classify intonational phrase final intonation (pairs of phrase accents and boundary tones) and intermediate phrase final intonation (phrase accents) using the same technique. We use SVM classification with acoustic features extracted from the **Last 200ms** preceding a phrase boundary (cf. Sections 6.5.3 and 6.6). Results of L2 classification using hypothesized tone sequences can be found in Figure 7.3. Recall that the majority class baseline classification accuracy using the BDC-read material is 68.9%.

We find a remarkable result when comparing the classification performance based on manual ToBI annotations to that based on hypothesized prosodic events. Contrary to expectation, we find that the classification accuracy using hypothesized prosodic events is consistently higher than that obtained from using manual annotations. Using manual annotations, 80% classification accuracy was achieved on sequences of 6 words; using hypotheses, this classification threshold is obtained with only 3 words. With manual

Corpus	No Accent	Accent				
		H*	L+H*	L*	L*+H	H+!H*
MC	29.2	***71.8***				
		89.5	4.6	5.2	0.0	0.6
BDC-read	57.2	42.8				
		60.2	29.8	9.7	0.0	0.3

Table 7.8: *Relative frequency (%) of hypothesized pitch accents and associated types on MC and BDC-read material.*

information, 90% accuracy is obtained with fifteen words; with hypothesized prosodic events, this 90% threshold is obtained with only 5 words. To classify speakers as MC or SAE with 95% accuracy manual annotations require tone sequences drawn from 31 words – approximately 11.13 seconds of speech; using hypothesized prosodic events, 95.15% accuracy is obtained using sequences of 7 words – approximately 2.51 seconds of speech.

We would expect that the errors introduced by using *hypothesized* prosodic events would *reduce* the classification performance. However, we find that the performance is improved. This indicates that there are regularities in the hypotheses produced on SAE and MC data which carry information about the native language of the speaker. To identify these regularities, we examine the distribution of hypothesized tones in each corpus. The distribution of pitch accents can be found in Table 7.8.

The distribution of hypothesized accents is informative in explaining why modeling hypothesized tones yields better non-native classification than manual annotations. Many more MC tokens are hypothesized to be accented than BDC-read tokens. It has been observed that L2 speech typically has a slower speaking rate, and, correspondingly, longer words [205]. The increased duration of L2 words is likely the source of the over-estimation of pitch accents.

Another contribution to the improved performance using hypothesized tones can be found by examining the hypothesized phrase ending tones. The distribution of phrase ending tones can be found in Table 7.9. The intonational phrase detection detects more phrase boundaries on the MC data than the BDC-read material. This is consistent with the literature

Corpus	No Phrase	ip boundary		IP boundary				
		L-	H-	L-L%	L-H%	H-L%	!H-L%	H-H%
MC	76.5	1.9		21.6				
		71.3	28.7	3.2	11.2	0.1	***85.3***	0.2
BDC-read	86.4	1.7		11.8				
		88.4	11.6	57.1	37.3	4.1	0.5	0.9

Table 7.9: *Relative frequency (%) of hypothesized phrase boundaries and their associated tones on MC, and BDC-read material. "ip boundary" is the rate of Intermediate Phrase Boundaries. "IP boundary" is the rate of Intonational Phrase Boundary rate.*

		Manual Phrase Ending Tone								
		H-H%	H-L%	!H-L%	L-H%	L-L%	H-	!H-	L-	NONE
Hyp. Phrase Ending Tone	H-H%	0	0	0	0	0	1	0	1	0
	H-L%	0	0	0	0	0	1	0	0	0
	!H-L%	62	129	13	221	391	21	0	21	8
	L-H%	2	19	3	11	42	11	0	13	13
	L-L%	0	2	0	1	10	1	0	7	11
	H-	1	2	0	0	0	5	0	2	15
	!H-	0	0	0	0	0	0	0	0	0
	L-	0	2	0	1	9	3	0	7	41
	NONE	27	82	9	64	123	157	5	357	2763

Table 7.10: *Confusion matrix of manually annotated and hypothesized phrase ending tones on Mandarin Chinese material. Hypothesized tones are generated from models trained on BDC-read material.*

on L2 speech; L2 speakers tend to have more, and therefore shorter, intonational phrases [205, 123]. A major anomaly can be found in the classification of intonational phrase ending intonation. The !H-L% contour is dramatically over-predicted on MC material. The over-prediciction of this tone occurs in a wide range of contexts. This can be observed by examination of the confusion matrix of phrase-ending tones which is presented in Table 7.10. This pervasive error does not instill particular confidence in the robustness of the phrase final classification module. A closer error analysis of this component is required to determine the source of these errors. The performance of the intonational phrase final classification module and the over-estimation of !H-L% tones likely contribute to the increased non-native classification performance using sequences of hypothesized tones.

While both the pitch accent detection and the phrase final classification components perform poorly on non-native speech, their limitations prove to be an advantage for classifying non-native speech. While both components show significant errors in the detection and classification of prosodic events on non-native speech, these errors are consistent. The consistency of the errors on L2 speech allows the tone model to use information about the errors to classify non-native speech with greater accuracy with hypothesized tone sequences than with manual annotations.

7.3.4 Conclusion and Future Work

The experiments in this section represent proof-of-concept results supporting the hypothesis that non-native speech can be identified using prosodic event information. We find that the distribution of tone types and phrase lengths are significantly different in native and non-native speech. Moreover, by comparing the perplexity of bigram tone sequence models, we can determine if speech is produced by a native Standard American English (SAE) speaker or a non-native, Mandarin Chinese (MC), speaker. Modeling sequences of hypothesized prosodic event location and type yields 100% SAE vs. MC classification accuracy on tone sequences extracted from 34 words, accuracy of 95.14% is obtained with sequences from 7 words, and 85.99% with tone sequences extracted from 3 words.

This supports the notion that prosodic event information can be used to identify regions of speech that are native and those that are indicative of non-native speech. This speaker origin information can be used for two tasks. Downstream spoken language processing performance can be improved by selecting or adapting lexical and acoustic models based on the native tongue of the speaker. Second, language proficiency instruction and assessment can be augmented by an automated technique to identify the "nativeness" of a non-native speaker's speech productions.

These results are preliminary, and the opportunity for future work is significant. One drawback of the classification approach described in Section 7.3.3 is that it requires Mandarin Chinese training data for the assessment of nativeness. It would be preferable to have a single model of native SAE speech, and to be able to identify tone sequences that are sufficiently divergent from this model, and flag these as anomalous and potentially 'non-native'. The development of this type of approach would be more suited to assessing non-native speech from different L1 speaker groups. The current approach requires training data from each non-native community – a significant drawback. Of course, a more general approach presents the problem of differentiating non-native intonation from other intonational anomalies produced by speech pathologies, disfluencies, or simply native speaker variation. Without some knowledge of a speaker's native tongue this may prove to be a difficult task.

When using a bigram tone model to characterize native and non-native speech, we observe significantly different behaviors between professional and non-professional read speech. This suggests that sequential tone modeling may be used to distinguish genre as well as native from non-native speech. It is also possible that sequential tone modeling could contribute to speaker identification and verification, and language or dialect identification.

We collected the Mandarin Chinese (MC) speech material from stories that have been read by multiple Standard American English (SAE) speakers to identify those specific areas of consistency among groups of language speakers. The goal is to identify those intonational characteristics that are consistent within SAE speakers, where MC speakers show significant divergence. We observed broad differences in the use of prosodic events. In the future we will continue this investigation to identify the lexical, syntactic and semantic contexts in which SAE and MC speakers display predictable divergence in their intonational behavior. The experiments in this section indicate that the intonation of non-native speech is different from native speech. Identifying more precisely how these differences are manifested will significantly contribute to the automatic assessment and processing of non-native speech.

7.4 Key Observations

- **Intonational Phrase Boundaries represent useful segmentation information** Intonational phrase boundaries are used as segmentation approaches in both extractive summarization and story segmentation. While not always the *most* informative segmentation strategy, on both tasks we observe the usefulness of this segmentation information in aiding downstream spoken language processing tasks.

- **Tone sequences can be used to model speaker differences.** In this chapter, we apply tone sequence modeling to successfully classify native Standard American English speakers from native Mandarin Chinese speakers. It remains an open question of whether or not this approach can be applied to other speaker classification tasks.

- **Hypothesized prosodic events can be used to detect speaker differences despite being errorful.** Though hypothesized prosodic events contain a significant degree of noise, we find that hypothesized tone sequences can be used to successfully classify the nativeness of a speaker. In fact, due to the consistency in the errors within each speaker group, the classification performance based on hypothesized tone sequence is *greater* than the performance based on sequences of manual tone annotations.

Chapter 8

Conclusion and Future Work

Prosodic events play a major role in human communication. Achieving human-like interaction with spoken material requires reliable analysis of the prosodic content of speech, and of prosodic events in particular. In this thesis, we present research on the extraction of prosodic information from spoken material, focusing our attention on the categorical prosodic events of accenting and phrasing. This research has led to state-of-the-art performance on the detection and classification of prosodic events, in some cases nearing human rates of agreement.

We identify five major contributions of the research presented in this thesis.

- **Novel Techniques** In each chapter, we apply novel machine learning and or feature extraction techniques to the tasks of prosodic event detection and classification. In some instances, these represent novel learning approaches, while in other cases, this is a matter of applying a known technique to the task for the first time.

  In Chapter 3, we use two classifier combination techniques to detect pitch accents. The Corrected Energy Based Classifier represents a novel classification technique (cf. Section 3.6). In this classifier, each member of an ensemble of classifiers is paired with a secondary classifier which is responsible for predicting whether the main classifier is correct or not. When the correcting classifier predicts that the main classifier is incorrect, the prediction is inverted before contributing to the ensemble. In this chapter, we also use part-of-speech (POS) class-based modeling (cf. Section 3.7.4). In this technique, we train separate acoustic models for tokens of different POS-based word classes.

In Chapter 4, we apply Boosting – specifically AdaBoost using single split decision trees – to the task of phrase boundary detection for the first time (cf. Section 4.3). We also explore the use of a top-down phrase detection approach, in which intonational phrase boundaries are detected first, and intermediate phrase boundaries are detected within intonational phrases (cf. Section 4.5).

In Chapter 5, we find accuracy to be an unreliable measure for evaluating automatic pitch accent classification (cf. Section 5.6.1). To evaluate our work on this task, we define a measure based on the Type I and Type II Classification Error Rates called "Combined Error Rate". We apply ensemble sampling to this task for the first time (cf. Section 5.6.4). We find that this sampling technique is able to significantly improve classification performance. Also, in this chapter, we present a novel Bayesian classification technique, Quantized Contour Modeling (cf. Section 5.6.3). In this technique, we quantize an acoustic contour into a fixed number of time and value bins, and learn value models for each time slice. At evaluation, we compare the posteriors of a set of pitch accent type models to identify the model which is most likely to generate an observed contour. We also apply Quantized Contour Modeling to the task of classifying phrase ending intonation in Chapter 6 (cf. Section 6.5.4).

The speech and natural language processing communities often treat machine learning techniques as black boxes – tools to associate feature vectors with annotations. We find throughout this thesis research that structuring classifiers – modifying their training and combining their results – can be used to improve overall performance. This opens up many opportunities for scientific creativity. Domain knowledge has always been used in directing the extraction of features. This work, through correcting classifiers, word-class based modeling, ensemble sampling and top-down phrase boundary detection, demonstrates that domain knowledge can be used to direct the construction of classification techniques as well as feature representations.

- **Novel Features** In each chapter of the thesis, we also apply novel feature extraction techniques, some of which evaluate the use of different regions of analysis while others examine new representations of previously studied phenomena.

  In Chapter 3, we investigate the use of acoustic and syntactic features for pitch accent detection. In this chapter, we measure the importance of acoustic context in representations of pitch and intensity features (cf. 3.3). Automatic pitch accent detection techniques operate by detecting accented *syllables* or *words*. We evaluate the impact of extracting acoustic features from the syllable or word level, finding that features extracted from the word are more discriminative of pitch accent (cf. Section 3.4). We extract energy features from 210 spectral regions for use in the corrected energy based classifier. While previous work has investigated the use of measures of spectral tilt for pitch accent detection [199, 207], this is the first time such a wide range of filtered energy features has been used in this task (cf. Section 3.6). Also in this work, we examine part-of-speech (POS) based word class features. In addition to evaluating previously studied word class definitions – function vs. content, and broad class (e.g. noun, verb, adjective) – we define three data-driven word-classes. We defined these word classes by grouping POS classes with similar accent rates together. These word classes are able to predict pitch accent locations with higher accuracy than the syntactically motivated groupings previously studied (cf. Section 3.7).

  We investigate techniques to automatically detect prosodic phrase boundaries in Chapter 4. Acoustic reset has been previously identified as an indicator of prosodic phrase boundaries. In Section 4.3.1, we explore a number of different features to capture acoustic reset. We find intensity reset to be a more reliable indicator of phrasing than pitch reset. Also, we find that intensity reset calculated from 40ms, only four intensity points, surrounding a candidate boundary to detect phrase boundaries better than other reset features calculated from larger regions. In addition, we explore a number of novel features based on linear regression fit lines to identify pitch and intensity reset (cf. Section 4.3.1). Finally, we examine the relationship between phrase boundaries and syntactic features (cf. Section 4.4).

In Chapter 5, we present research on the automatic classification of pitch accents. In examining acoustic correlates of pitch accent type, we extract acoustic features from a variety of regions of analysis. These regions include the **Full Word**, the **Syllable** and two acoustic pseudo-syllabification techniques – one defined by Villing et al. [219], and another defined in Section 5.6.3. We find the **Syllable** region to yield the best performance, with the Villing approach yielding classification performance slightly better than the features extracted from the **Full Word**.

We address the classification of phrase final intonation – phrase accent and boundary tone pairs – in Chapter 6. Similar to the investigation in Chapter 5, we attempt to identify the best region of analysis for classification of phrase final intonation. We examine the **Full Word**, **Final Syllable**, **Last half of a word**, **Last 200ms**, **Final Energy Peak**, **Final Villing pseudo-syllable**, the region from the **Last Accent** and the region from the **Last "accentable" syllable**. We find that classification using features extracted from the **Last 200ms** preceding a candidate boundary to yield the best performing phrase boundary detection results (cf. Section 6.5.3). We also apply syntactic parse tree based features to the task of classifying phrase ending intonation (cf. Section 6.5.2). While these features have been used in *detecting* phrase boundaries, we believe this a novel use of this type of information for *classifying* phrase final intonation.

- **Improved Understanding** In addition to achieving high performance on these automated tasks, we seek to improve the scientific understanding of prosodic phenomena. This analysis is performed as post-hoc analyses of classification results, examining the performance of different feature sets, and independent descriptive analyses of acoustic and syntactic correlates of prosodic events.

Automatic pitch accent detection approaches can be divided into those that detect accent bearing *words* and those that detect accent bearing *syllables*. However, there is no consensus regarding which region of analysis is optimal. To address this question, in Section 3.4, we compare equivalent approaches to pitch accent detection operating on syllables and words. We find that, all else being equal, word-based pitch accent detection approaches perform better than syllable-based approaches. We find that incorporation of acoustic context from surrounding words in pitch, intensity and duration features significantly improves automatic pitch accent detection (cf. Section 3.3). In Section 3.5, we examine the relationship between filtered energy information – spectral based features – and pitch accent detection. Previous work found that spectral tilt, spectral balance or high frequency emphasis correlate with the perception and production of pitch accents [185, 186, 53, 78]. This research considered the energy in four spectral regions – finding energy extracted from 500Hz-2kHz to show the greatest correlation with accenting behavior. We expand on this work by examining the correlation between energy filtered using 210 frequency bands and pitch accent. We find that the discriminative power of energy features extracted from different frequency regions varies widely and inconsistently. A single optimal frequency band to calculate spectral tilt cannot be identified. We find that the region between 2 and 20 bark is the most robust to speaker differences (cf. Section 3.5). However, we find that energy features extracted from this region perform equivalently to those extracted from the full spectrum when combined with pitch and duration features (cf. Section 3.6.2).

The relationship between syntax and phrase boundary location has received considerable attention ([7], [217], [86], [104], *inter alia*). In Section 4.4, we compare the use of a number of part-of-speech (POS) n-gram models. We find the best are those that incorporate POS information from tokens *surrounding* a candidate boundary, not only those tokens preceding a boundary. This finding is confirmed using both surrounding POS bigrams and surrounding 4-grams, compared to standard unigram, bigram and trigram models. Pitch reset has previously been identified as an indicator of prosodic phrasing. In addition to evaluating the use of pitch reset in automatic phrase boundary detection, we evaluate the use of energy reset measures. We find energy reset to be a more reliable indicator of intonational phrase boundaries than pitch (cf. Section 4.3.1). Further study is necessary to more completely understand the relationship between acoustic reset and phrasing; this finding indicates that energy reset should be considered an integral part of any such study, along with the more heavily studied pitch reset.

In Section 5.5, we perform descriptive analyses of acoustic qualities of pitch accent types. In Section 5.5.1, we narrow our analysis to the differences between H* and L+H* accents – the two most confusable, and most common accents in the examined corpora.[1] This analysis reveals some characteristics that define the differences of these two types in the ToBI standard. For example, we find that L+H* accents have a greater pitch slope, and reach a pitch peak later in the word than H* accents. These findings are most likely explained by the presence of the "sharp rise" that the ToBI standard identifies for L+H*. Moreover, we find that L+H* accents have greater energy and duration than H*. These qualities are not noted in the ToBI standard, but are probably related to the impression that L+H* accents are more *emphatic* than H* accents.

In our research on automatic classification of pitch accent and phrase final intonation, we explore the use of a number of regions of analysis from which to extract acoustic features (cf. Sections 5.6.3 and 6.5.3). Features extracted from regions defined

[1] H* and L+H* accents make up over 90% of accents in BDC and BURNC material.

by forced-alignment based syllable boundaries, and acoustic pseudo-syllabification [219] yield the best pitch accent classification performance. In the classification of phrase-final intonation – phrase accent and boundary tone pairs – we find that features extracted from phrase final syllables, and the 200ms preceding each candidate boundary achieve the best performance.

In Section 6.5.2, we examine the relationship between syntactic features and phrase final types. We find that H-L% and L-H% are used at word boundaries of less syntactic disjuncture than L-L% and H-H%. In discourse, L-H% and H-L% phrase endings often leave the impression that the speaker has more to say. The L-H% is commonly used in the "continuation rise" contour, while H-L% is used in dialog to hold the turn. Moreover, both of these at commonly used in "list intonation" when speaking a sequence of entities. Here we find that the continuity conveyed intonationally by these phrase final types is also observable in the syntactic structure.

In the two chapters concerning the classification of prosodic events – Chapters 5 and 6 – we discuss examples of pitch accent and phrase final types. Descriptions of the ToBI standard [183, 15] often contain descriptions and examples of canonical forms of the types of pitch accents, phrase accents and boundary tones. In Sections 5.3 and 6.2, we discuss both prototypical and potentially ambiguous exemplars of each type of prosodic event. By identifying some ways in which types of events can be ambiguous, we identify potential difficulty for human and machine annotators. We believe that discussion of the ambiguities between types of prosodic events defined by the ToBI standard serves to increase understanding not only of the annotation standard itself, but also the underlying prosodic phenomena.

- **High performance** The novel techniques and features described yield state-of-the-art performance on many of the prosodic event detection and classification tasks that are addressed in this thesis. A summary of the best results under ten-fold cross-validation on each corpus is available in Table 8.1. Results that we believe to be state-of-the art are indicated in bold and marked with an asterisk.

Identifying the current state of the art performance on the BURNC corpus when lexico-syntactic features are used is difficult. Many previous evaluations [237, 189, 187] which generated high performance *may* be fundamentally flawed. While not a one-to-one mapping, there is a significant relationship between lexical content and prosodic event type and location. The BURNC corpus contains multiple speakers reading the same news stories. Therefore, when using text based features such as syntactic information, POS tags, or word identity, it is critical to ensure that no news story appears simultaneously in training and evaluation material. Otherwise, the evaluation measures how consistently two or more readers of broadcast news produce the same material, rather than how reliably a set of features or a particular method can detect prosodic events. In some previous studies it is not clear whether or not the evaluation suffers from this problem [237, 4, 81]; in others, it is clear that the same story has been used for training and evaluation [189, 187]. The only previous work that has addressed this issue explicitly is [5].

The best performing acoustic pitch accent detection results on all corpora are obtained using the Corrected Energy Based Classification technique (cf. Section 3.6). This approach uses energy features extracted from 210 frequency regions to generate initial hypotheses, and pitch and duration features to correct these hypotheses. Context normalization is performed for all features. The inclusion of syntactic information slightly improves detection performance. This improvement is obtained by performing word class-based acoustic modeling. The Corrected Energy Based Classifiers are used for the acoustic model, and static threshold data-driven part-of-speech tag based word class representation yields the reported results (cf. Section 3.7.5).

We observe the best phrase boundary detection results by applying AdaBoost with single split decision trees to the task (cf. Section 4.3). The acoustic features we use include 1) aggregations of pitch and energy extracted from the word preceding a candidate phrase boundary, 2) reset features, where the difference of acoustic aggregations from regions on either side of a candidate boundary are calculated and, on the BURNC material, 3) segment length features to capture preboundary lengthening. We observe significant improvement by extending the acoustic feature vector with syntactic information (cf. Section 4.4). The syntactic features include part of speech model likelihoods, and parse-tree based features.

We use a measure called Combined Error Rate (CER) to evaluate pitch accent classification results 5.6.1. We obtain lowest CER by performing *post hoc* combination of Quantized Contour Model posteriors and ensemble sampled SVM classifier confidence scores. The acoustic features include aggregations of pitch and energy and representations of the contour shape including Tilt coefficients. These acoustic features are extracted from pseudo-syllable regions [219] on the BDC material and forced-alignment based syllables on the BURNC corpus. We do not find syntactic information to improve the pitch accent type classification. We are not aware of other approaches to pitch accent classification that have been evaluated on the BDC and BURNC material using the same type inventory used here. Without this comparison we hesitate to call these results superior to competitors, rather we believe they represent the first reported results on this task.

The highest accuracy phrase-final classification results are obtained by SVM classification using an acoustic feature vector including Quantized Contour Model posteriors. The acoustic features include aggregations of pitch and energy, representations of the pitch and energy contour shape and voice quality features. These features are extracted from the **Last 200ms** preceding a phrase boundary on the BDC-read material, the phrase final **Full Word** on the BDC-spontaneous material, and the phrase final **syllable** on the BURNC corpus. We do not observe any increase in phrase-final classification performance by incorporating syntactic features into this successful acoustic classification technique.

- **Applications** This thesis demonstrates that prosodic event information can be extracted from speech with high performance. We provide supporting evidence that these hypothesized prosodic events can be used to improve the performance of spoken language processing tasks. We provide examples of the application of prosodic event detection to extractive speech summarization (cf. Section 7.1), story segmentation (cf. Section 7.2) and the assessment of non-native speech (cf. Section 7.3). We believe this work provides strong evidence for the availability, theoretical importance and practical impact of the use of prosodic event information in a variety of spoken language processing tasks.

There are a number of directions in which this work can continue to develop. In this thesis, we have, for the most part, treated prosodic events as independent phenomena. Yet, there is evidence that this independence assumption does not hold. For example, the ToBI standard requires that each intermediate phrase contains at least one word bearing a pitch accent. This requirement has not been incorporated into either the detection of accent or phrase boundaries in this work. Incorporation of the relationship between prosodic events is a clear direction for future work on prosodic event detection.

A clear extension of the techniques presented in this thesis is the application of hypothesized prosodic event information to downstream spoken language processing tasks. The applications described in Chapter 7 can be expanded to take advantage of a wider range of hypothesized prosodic events. For example, topic, salience and information status appear to be correlated with the presence and type of pitch accent. Pitch accent presence and type information should prove useful to topic segmentation and extractive summarization.[2] Also, we can apply the tone modeling technique used in non-native assessment to assess other aspects of speech including genre, language and speaker identity. Moreover, the automatic

[2] Recall that, in Sections 7.1 and refapp:segmentation we only applied hypothesized intonational phrase boundaries to these tasks.

prosodic event detection and classification can be applied to many other spoken language processing tasks including native fluency assessment for medical assessment of speech pathologies, automatic speech recognition, emotion classification, turn-taking, speaker identification/verification and dialect identification.

The work in this thesis has employed supervised learning techniques for the detection and classification of prosodic events. These require the manual annotation of intonation, a very time-consuming task. Expert labelers can take up to 200 times real-time to perform full ToBI labeling of speech. The application of unsupervised and semi-supervised learning techniques is attractive to avoid the burden of these resource intensive data requirements. Both of these directions can take advantage of the feature representations and descriptive analysis performed in this thesis. The application of these results to learning with little or no manually annotated data is an attractive next step in this work.

Prosody in general and prosodic events in particular indicate how spoken information is structured. A substantial portion of spoken material is not syntactically well-formed, due to disfluencies, errors, and the use of sentence fragments. Prosodic information should be able to contribute to identifying the correct interpretation of these syntactically ill-formed utterances. One approach would be to use prosodic information in scoring candidate partial parses, building on work by Ostendorf and Veilleux [146]. Disambiguation of these ill-formed spoken utterances may lead to improvements in syntactic, semantic and pragmatic interpretation of ill-formed written material. Chat, blog content and comments, and email frequently contain information that is not syntactically well-formed. With the current widespread and growing use of this type of communication, the need for robust natural language processing techniques is becoming more and more pronounced.

We believe the contributions of this thesis advance the understanding of prosodic events and the use of prosody in spoken language processing towards the goal of human-like processing of speech by machines.

8.1 Key Observations

In the conclusion of each chapter, we outline key insights specific to the work described therein. Here we identify some broader insights from throughout the thesis.

- **Pitch vs. Energy** Prosodic variation has long been held to be dominated by changes in fundamental frequency, pitch. It this thesis, we find that energy is more useful in the *detection* of prosodic events, while pitch is more useful in their *classification*. This is clearly an oversimplification; both acoustic qualities contribute to high accuracy detection and classification of prosodic events. However, rather than being dominated by pitch, energy should be considered to have at least equal importance in the analysis of prosody.

- **Acoustic vs. Lexico-Syntactic Information** Lexico-syntactic information can be used to determine the sequence of prosodic events most likely to be produced when an sequence of lexical items is spoken. Acoustic prosodic analysis is concerned with the prosodic events that are *actually* realized, as opposed to the *expected* sequence of events given text. Therefore, using lexico-syntactic information in acoustic prosodic event detection and classification requires a balancing of these two information streams. Reliance on lexico-syntactic information will lead an automatic routine to miss anomalous prosodic events – which may be of particular significance. However, lexico-syntactic features can be used to improve prosodic analysis by resolving cases of acoustic ambiguity. In prosodic event *detection*, we have been able to use syntactic information to improve performance; identifying a way to use lexico-syntactic information to improve prosodic event *classification* remains an open research question.

- **Ensemble and Combination Methods** The main machine learning theme that emerges from the work in this thesis is the success of classifier combination and ensemble techniques. In all classification and detection tasks, classification combination and ensemble techniques are used to generate the highest performance. This is realized in the use of standard techniques, AdaBoostM1, for phrase detection, and ensemble sampling for pitch accent classification. The inclusion of Quantized Contour Modeling posteriors in an SVM feature vector improves phrase final classification. Finally, the corrected energy-based classifiers and word-class based modeling in pitch accent detection are more novel techniques where novel *classifier structure* rather than feature engineering is used to improve classification performance. The consistency of this theme points to the power of these approaches, and are encouraging to the application of creativity and engineering not only in feature extraction, but in classifier structure.

8.2 Limitations

Though we have advanced the state-of-the-art automatic prosodic event detection and classification performance in many of the tasks addressed in this thesis, the approaches used have some limitations. In this section we summarize these limitations.

- **Supervised techniques have high resource requirements.** By relying on supervised machine learning techniques, the approaches described in this thesis require manual ToBI labeling of large corpora. While there are some available annotated corpora, this annotation process is particularly resource intensive. Therefore there is a high resource requirement, to extend the described prosodic analysis techniques to new domains or languages.

- **Word and syllable boundary information is useful.** The highest performing approaches to prosodic event detection and classification rely on the presence of word or syllable boundary information. If manually defined, these boundaries require significant resources to generate, while if generated from forced alignment of manual transcriptions or automatic speech recognition, some amount of noise is introduced.

While this is a limitation, in each task we have presented approaches which do not rely on this information, detecting and classifying prosodic events on acoustically defined pseudo-syllable units.

- **Speaker identity information is important.** Many of the acoustic features used in the experiments in this thesis are normalized by speaker. Automatic speaker identification is a non-trivial task. While manual speaker identification is less resource-intensive than other annotations – often a speaker's identity is known when spoken material is collected – this remains a limitation of the approaches to prosodic analysis.

- **There is no guarantee that improvement to automatic prosodic event detection and classification leads to improvement of downstream spoken language tasks.** This is a limitation that is not unique to any approach taken to address these tasks, or to the tasks themselves. Any automatic annotation of intermediate data – syntactic parsing, semantic analysis, speaker identification, story segmentation, *inter alia* – may suffer from this limitation. It is unclear *a priori* to what degree improvement to intermediate steps are realized as improvement to the broader task. The assumption is that improvements to intermediate tasks will lead to downstream task improvements, but this assumption does not *always* hold. One counter-example can be found in the assessment of non-native intonation (cf. Section 7.3), where errorful automatic tone hypotheses lead to identification of non-native speech with higher performance than the manual tone annotations. In this case, if automatic prosodic event classification and detection were improved to 100% accuracy, the performance of non-native speech assessment would significantly degrade. While addressing this limitation is beyond the scope of this thesis, it remains a frequently ignored limitation of this and similar research.

	Task	Feature Set	Performance
BDC-read	Accent Detection	**A**	*84.4% ± 0.51
		A+S	*84.6% ± 0.41
	IP Detection	**A**	*95.71% ± 0.33 F_1: 0.826 ± 0.0121
		A+S	*95.71% ± 0.33 F_1: 0.826 ± 0.0121
	ip Detection	**A**	*92.83% ± 0.25 F_1: 0.442 ± 0.0218
		A+S	*93.94% ± 0.30 F_1: 0.543 ± 0.0197
	Accent Classification	**A**	*75.38% ± 0.48 CER: 0.246 ± 0.0253
		A+S	*75.38% ± 0.48 CER: 0.246 ± 0.0253
	Phrase Classification	**A**	*73.39% ± 1.43
		A+S	*73.39% ± 1.43
BDC-spon	Accent Detection	**A**	*83.2% ± 0.38
		A+S	*83.6% ± 0.46
	IP Detection	**A**	*93.00% ± 0.39 F_1: 0.809 ± 0.0108
		A+S	*93.75% ± 0.34 F_1: 0.834 ± 0.0090
	ip Detection	**A**	*91.29% ± 0.28 F_1: 0.484 ± 0.0277
		A+S	*91.79% ± 0.25 F_1: 0.541 ± 0.0213
	Accent Classification	**A**	*61.50% ± 1.13 CER: 0.284 ± 0.0236
		A+S	*61.50% ± 1.13 CER: 0.284 ± 0.0236
	Phrase Classification	**A**	*58.68% ± 1.59
		A+S	*58.68% ± 1.59
BURNC	Accent Detection	**A**	*85.5% ± 0.20
		A+S	85.9% ± 0.36
	IP Detection	**A**	*89.46% ± 0.38 F_1: 0.736 ± 0.0075
		A+S	91.48% ± 0.31 F_1: 0.761 ± 0.00115
	ip Detection	**A**	*88.68% ± 0.46 F_1: 0.257 ± 0.0129
		A+S	*89.50% ± 0.23 F_1: 0.394 ± 0.0213
	Accent Classification	**A**	*60.86% ± 1.07 CER: 0.364 ± 0.0202
		A+S	*60.86% ± 1.07 CER: 0.364 ± 0.0202
	Phrase Classification	**A**	*78.99% ± 0.95
		A+S	*78.99% ± 0.95

Table 8.1: *Summary of best results on all prosodic event detection and classification tasks. "IP" refers to Intonational Phrase Boundary. "ip" refers to Intermediate Phrase Boundary.* **A** *indicates Acoustic features only.* **A + S** *indicates the use of both Acoustic and Syntactic features. An asterisk indicates that we believe the result to be the best published on the task.*

Chapter 9

Bibliography

[1] J. F. Allen. A study on prosody and discourse structure in cooperative dialogues. *Phonetica*, 50:197–210, 1993.

[2] S. Ananthakrishnan, P. Ghosh, and S. Narayanan. Automatic classification of question turns in spontaneous speech using lexical and prosodic evidence. In *ICASSP*, 2008.

[3] S. Ananthakrishnan and S. Narayanan. An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model. In *ICASSP*, 2005.

[4] S. Ananthakrishnan and S. Narayanan. Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling. In *ICSLP*, 2006.

[5] S. Ananthakrishnan and S. Narayanan. Fine-grained pitch accent and boundary tone labeling with parametric f0 features. In *ICASSP*, 2008.

[6] A. Arvaniti and M. Baltazani. Greek tobi: A system for the annotation of greek speech corpora. In *Proceedings of Second International Conference on Language Resources and Evaluation*, pages 555–562, 2000.

[7] J. Bachenko and E. Fitzpatrick. A computational grammar of discourse-neutral prosodic phrasing in english. *Computational Linguistics*, 16:155–170, 1990.

[8] P. J. Baken and R. F. Orlikoff. *Clinical Measurement of Speech and Voice*. CA:Singular Publishing Group, San Diego, 2 edition, 2000.

[9] S. Bangalore and A. K. Joshi. Supertagging: An approach to almost parsing. *Computational Linguistics*, 25(2):237–265, 1999.

[10] C. Bartels. *The Intonation of English Statements and Question: A Compositional Interpretation*. Garland Publications, 1999.

[11] A. Batliner, R. Kompe, A. Kiessling, E. Nöth, H. Niemann, and U. Kilian. The prosodic marking of phrase boundaries: Expectations and results. In Rubio-Ayuso and Lopez-Soler, editors, *Speech Recognition and Coding – New Advances and Trends*, pages 89–92. Springer, 1995.

[12] A. Batliner, E. Nöth, J. Buckow, R. Huber, V. Warnke, and H. Niemann. Duration features in prosodic classification: why normalization comes second, and what they really encode. In *Prosody 2001*, 2001.

[13] C. M. Beach. The interpretation of prosodic patterns at points of syntactic structure ambiguity: evidence for cue trading relations. *Journal of Memory and Language*, 1991.

[14] M. Beckman. *Stress and non-Stress*. Foris Publications, Dordrect, Holland, 1986.

[15] M. E. Beckman and G. A. Elam. Guidelines for ToBI labelling. Ohio State University, 1994.

[16] D. Beeferman, A. Berger, and J. Lafferty. Statistical models for text segmentation. *Machine Learning*, 31(1-3):177–210, 1999.

[17] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago Computer Science, 2004.

[18] S. Benus, A. Gravano, and J. Hirschberg. Prosody, emotions, and... 'whatever'. In *Interspeech*, 2007.

[19] A. Bies. Bracketing guidelines for treebank ii style penn treebank project, 1995.

[20] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9-10):341–345, 2001.

[21] D. Bolinger. A theory of pitch acent in english. *Word*, 14:109–149, 1958.

[22] D. Bolinger. Contrastive accent and contrastive stress. *Language*, 37:83–96, 1961.

[23] D. Bolinger. Accent is predictable (if you're a mind-reader). *Language*, 48, 1972.

[24] D. Bolinger. The network tone of voice. *Journal of Broadcasting*, 26:725–728, 1982.

[25] D. Bolinger. *Intonation and Its Parts: Melody in Spoken English*. Stanford University Press, 1985.

[26] D. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford University Press, 1989.

[27] N. Braunschweiler. *Automatic Detection of Prosodic Cues*. PhD thesis, University of Konstanz, Germany, 2005.

[28] N. Braunschweiler. The prosodizer - automatic prosodic annotations of speech synthesis databases. In *Speech Prosody*, 2006.

[29] J. Brenier, D. Cer, and D. Jurafsky. The detection of emphatic words using acoustic and lexical features. In *Eurospeech*, 2005.

[30] E. Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992.

[31] G. Brown. Prosodic structure and the given/new distinction. In A. Cutler and D. Ladd, editors, *Prosody: Models and Measurements*, pages 67–77. Springer Verlag, Berlin, 1983.

[32] N. Campbell. Loudness, spectral tilt and perceived prominence in dialogues. In *ICPhS*, 1995.

[33] R. Carlson, J. Hirschberg, and M. Swerts. Cues to upcoming swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. *Speech Communication*, 46:326–333, 2005.

[34] J. Caspers. Local speech melody as a limiting factor in the turn-taking system in dutch. *Journal of Phonetics*, 31:251–276, 2003.

[35] L. Chaolei, L. Jia, and X. Shanhong. English sentence stress detection system based on hmm framework. *Applied Mathematics and Computation*, 185:759–768, 2007.

[36] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of the first conference on North American chapter of the Association for Computational Linguistics*, pages 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[37] F. Chen and M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *ICASSP*, 1992.

[38] K. Chen, M. Hasegawa-Johnson, and A. Cohen. An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic-prosodic model. In *ICASSP*, 2004.

[39] N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper and Row, 1968.

[40] H. Christensen, Y. Gotoh, and S. Renals. Punctuation annotation using statistical prosody models. In *in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding*, pages 35–40, 2001.

[41] H. Christensen, B. Kolluru, Y. Gotoh, and S. Renals. From text summarization to style-specific summarization for broadcast news. In *ECIR*, 2004.

[42] J. Clark and C. Yallup. *Introduction to Phonology and Phonetics*. Blackwell, 1990.

[43] A. Cohen. A survey of machine learning methods for predicting prosody in radio speech. Master's thesis, University of Illinois at Urbana-Champaign, 2004.

[44] A. Conkie, G. Riccardi, and R. Rose. Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events. In *Eurospeech*, 1999.

[45] C. Cucchiarini, H. Strik, and L. Boves. Using speech recognition technology to assess foreign speakers pronunciation of dutch. In *Proceedings of New Sounds*, pages 61–68, 1997.

[46] C. Cucchiarini, H. Strik, and L. Boves. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111(6):2862–2873, 2002.

[47] D. Dahan, M. Tanenhaus, and C. Chambers. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, 47:292–314, 2002.

[48] G. Demenko, S. Grocholewski, A. Wagner, and M. Szymanski. Prosody annotation for corpus based speech synthesis. In *Australian International Conference on Speech Science and Technology*, 2006.

[49] F. C. Diaz, J. van Santen, and E. R. Banga. Integrating phrasing and intonation modelling using syntactic and morphosyntactic information. *Speech Communication*, 51(5):452–465, 2009.

[50] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292, 1972.

[51] A. J. F. et al. Speech processing by man and machine. In T. H. Bullock, editor, *Recognition of Complex Acoustic Signals*. Report of Dahlem Workshop, Berlin, 1977.

[52] M. Fach and W. Wokurek. Pitch accent classification of fundamental frequency contours by hidden markov models. In *Eurospeech*, 1995.

[53] G. Fant, A. Kruckenberg, and J. Liljencrants. Acoustic-phonetic analysis of prominence in swedish. In A. Botinis, editor, *Intonation, Analysis, Modelling and Technology*, pages 55–86. Kluwer, 2000.

[54] I. Fischer and J. Poland. New methods for spectral clustering. Technical Report ISDIA-12-04, ISDIA, 2004.

[55] J. Fletcher and D. Loakes. Intonational variation in adolescent conversational speech: rural versus urban patterns. In *Speech Prosody*, 2006.

[56] H. Franco, L. Neumeyer, Y. Kim, and O. Ronen. Automatic pronunciation scoring for language instruction. In *ICASSP*, 1997.

[57] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *ICML*, 1996.

[58] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and Systems Sciences*, 55(1):119–139, 1997.

[59] H. Fujisaki and K. Hirose. Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis and synthesis of intonation. In *Preprints of papers, Working group on intonation, 13th International Congress Linguists*, pages 57–70, Tokyo, 1982.

[60] P. Fung. Fast accent identification and accented speech recognition. In *ICASSP*, pages 221–224, 1999.

[61] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing. Discourse segmentation of multi-party conversation. In *41st Annual Meeting of ACL*, pages 562–569, July 2003.

[62] B. García, I. Ruiz, A. Méndez, J. Vincente, and M. Mendezona. Automated characterization of esophageal and severely injured voices by means of acoustic parameters. In *EURASIP*, 2007.

[63] S. Gass and L. Selinker. *Second Language Acquisition: An Introductory Course*. University of Turku: Publications of the Department of Phonetics, 1994.

[64] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, 1:517–520 vol.1, Mar 1992.

[65] A. Gravano. *Turn Taking and Affirmative Cue Words in Task-Oriented Dialog*. PhD thesis, Columbia University, 2009.

[66] A. Gravano, S. Benus, H. Chávez, J. Hirschberg, and L. Wilcox. On the role of context and prosody in the interpretation of *okay*. In *ACL*, pages 800–807, Prague, Czech Republic, June 2007.

[67] A. Gravano, S. Benus, J. Hirschberg, E. S. German, and G. Ward. The effect of contour type and epistemic modality on the assessment of speaker certainty. In *Speech Prosody*, pages 401–404, Campesinas, Brazil, May 2008.

[68] A. Gravano and J. Hirschberg. Effect of genre, speaker and word class on the realization of given and new information. In *Interspeech*, pages 557–560, September 2006.

[69] M. Gregory and Y. Altun. Using conditional random fields to predict pitch accents in conversational speech. In *ACL*, 2004.

[70] M. Grice and M. Savino. Can pitch accent type convey information status in yes-no questions. In *Concept to Speech Generation Systems*, 1997.

[71] B. J. Grosz and C. L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 1986.

[72] J. Gundel. On different kinds of focus. In *Focus: Linguistic, Cognitive and Computational Perspectives*. Cambridge University Press, 1999.

[73] C. Gussenhoven. *A semantic analysis of the nuclear tones of English*, pages 193–265. Dordrecht, 1983.

[74] M. Hasegawa-Johnson, J. Cole, C. Shih, K. Chen, A. Cohen, S. Chavarria, H. Kim, T. Yoon, S. Borys, and J.-Y. Choi. Speech recognition models of the interdependence among syntax, prosody and segmental acoustics. In *HLT-NAACL*, 2004.

[75] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[76] N. Hedberg. The prosody of contrastive topic and focus in spoken english. In *Workshop on information structure in context*, 2003.

[77] P. A. Heeman. Modeling speech repairs and intonational phrasing to improve speech recognition. In *In Automatic Speech Recognition and Understanding Workshop*, pages 2–273, 1999.

[78] M. Heldner. Spectral emphasis as an additional source of information in accent detection. In *Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 57–60, 2001.

[79] M. Heldner and B. Megyesi. Exploring the prosody-syntax interface in conversations. In *In Proceedings ICPhS 2003*, pages 2501–2504, 2003.

[80] M. Heldner, E. Stragert, and T. Deschamps. Focus detection using overall intensity and high frequency emphasis. In *ICPhS*, 1999.

[81] J. Hirschberg. Pitch accent in context: Predicting intonational prominence from text. *Artificial Intelligence*, 63(1-2):305–340, 1993.

[82] J. Hirschberg. Communication and prosody: Functional aspects of prosody. *Speech Communication*, 36(1-2):31 – 43, 2002.

[83] J. Hirschberg. The pragmatics of intonational meaning. In *Speech Prosody*, 2002.

[84] J. Hirschberg. *Pragmatics and Intonation*, chapter 14. Blackwell Publishing, 2006.

[85] J. Hirschberg and M. Beckman. The tobi annotation conventions, 1994.

[86] J. Hirschberg and C. Nakatani. A prosodic analysis of discourse segments in direction-giving monologues. In *Proc. of the 34th conference on Association for Computational Linguistics*, pages 286–293, 1996.

[87] J. Hirschberg and C. Nakatani. Acoustic indicators of topic segmentation. In *Proc. of ICSLP*, volume 4, pages 1255–1258, 1998.

[88] J. Hirschberg and C. Nakatani. Using machine learning to identify intonational segmetns. Technical Report SS-98-01, AAAI, 1998.

[89] J. Hirschberg and J. Pierrehumbert. The intonational structuring of discourse. In *ACL*, pages 136–144, 1986.

[90] J. Hirschberg and P. Prieto. Training intonational phrasing rules automatically for english and spanish text-to-speech. *Speech Communication*, 18(3):281–290, 1996.

[91] J. Hirschberg and O. Rambow. Learning prosodic features using a tree representation. In *Eurospeech*, 2001.

[92] J. Hirschberg and G. Ward. The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in english. *Journal of Phonetics*, 20:241–251, 1992.

[93] J. Hirschberg and G. Ward. The interpretation of the high-rise question contour in english. *Journal of Pragmatics*, 1995.

[94] C. Hori, S. Furui, R. Malkin, H. Yu, and A. Waibel. Automatic speech summarization applied to english broadcast news. In *ICASSP*, 2002.

[95] G.-P. Hu, B.-F. Chen, M. Fan, and R. Wang. The contribution of mutual information in the intonational phrase prediction in chinese text. In *Natural Lanugage Processing and Knowledge Engineering*, pages 407–412, 2003.

[96] C. T. Ishi. Analysis of autocorrelation-based parameters for creaky voice detection analysis of autocorrelation-based parameters for creaky voice detection. In *Speech Prosody*, 2004.

[97] C. T. Ishi. Perceptually-related f0 parameters for automatic classification of phrase final tones. *IEICE Transactions on Information and Systems*, E88-D(3):481–488, 2005.

[98] K. Iwano and K. Hirose. Representing prosodic words using statistical models of moraic transition of fundamental frequency contours of japanese. In *ICSLP*, 1998.

[99] F. Jensen. *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[100] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[101] Y. jun Kim and Y. hwan Oh. Prediction of prosodic phrase boundaries considering variable speaking rate. In *in Proceedings of the International Conference on Spoken Language Processing*, pages 1505–1508, 1996.

[102] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological Bulletin*, 129(5), 2003.

[103] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustic Society of America*, 118(2):1038–1054, August 2005.

[104] P. Koehn, S. Abney, J. Hirschberg, and M. Collins. Improving intonational phrasing with syntactic information. In *ICASSP*, 2000.

[105] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den. An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. *Language and Speech*, 41(3–4):295–321, 1998.

[106] B. Kolluru, Y. Gotoh, and H. Christensen. Multistage compaction approach to broadcast news summarization. In *Interspeech*, 2005.

[107] H. Kozima. Text segmentation based on similarity between words. In *31st Annual Meeting of the ACL*, pages 286–288, 1993.

[108] K. Kryszczuk and A. Drygajlo. On combining evidence for reliability estimation in face verification. In *Proc. of the 14th European Conference on Signal Processing (EUSIPCO)*, Florence, Italy, 2006.

[109] R. Ladd. *The structure of intonational meaning*. Indiana University Press, 1980.

[110] R. Ladd. *Intonational Phonology*. Cambridge University Press, 1996.

[111] R. Ladd and R. Morton. The perception of intonational emphasis: continuous or categorical. *Journal of Phonetics*, 25(3):313–342, July 1997.

[112] R. Lakoff. *Language and Woman's Place: Text adn Commentaries*. Oxford University Press, 1975.

[113] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[114] G. A. Levow. Assessing prosodic and text features for segmentation of mandarin broadcast news. In *HLT-NAACL 2004*, 2004.

[115] G.-A. Levow. Context in multi-lingual tone and pitch accent recognition. In *Interspeech*, 2005.

[116] G.-A. Levow. Unsupervised and semi-supervised learning of tone and pitch accent. In *HLT-NAACL*, pages 224–231, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[117] G.-A. Levow. Automatic prosodic labeling with conditional random fields and rich acoustic features. In *IJCNLP*, 2008.

[118] A. Li. Chinese prosody and prosodic labeling of spontaneous speech. In *Speech Prosody*, pages 39–46, 2002.

[119] J.-F. Li, G.-P. Hu, and R. Wang. Chinese prosody phrase break prediction based on maximum entropy model. In *Interspeech*, 2004.

[120] J.-F. Li, G.-P. Hu, R. Wang, and L.-R. Dai. Sliding window smoothing for maximum entropy based intonational phrase prediction in chinese. In *ICASSP*, pages 285–288, 2005.

[121] X. Li, J. Tao, M. T. Johnson, J. Soltis, A. Savage, K. M. Leong, and J. D. Newman. Stress and emotion classification using jitter and shimmer features. In *ICASSP*, 2007.

[122] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop of Text Summarization*, 2004.

[123] J. Liscombe. *Prosody and Speaker State: Paralinguistics, Pragmatics and Proficiency*. PhD thesis, Columbia University, 2007.

[124] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. P. Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1526–1540, 2006.

[125] P. A. Luce and J. Charles-Luce. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, 78(1949–1957), 1985.

[126] M. Maragoudakis, P. Zervas, N. Fakotakis, and G. Kokkinakis. A data-driven framework for intonational phrase break prediction. In *TSD*, 2003.

[127] M. Marcus, M. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[128] E. Marsi, M. Reynaert, A. van den Bosch, W. Daelmans, and V. Hoste. Learning to predict pitch accents and prosodic boundaries in dutch. In *ACL*, 2003.

[129] S. Maskey and J. Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Eurospeech*, 2005.

[130] S. Maskey and J. Hirschberg. Summarizing speech without text using hidden markov models. In *HLT-NAACL*, 2006.

[131] S. Maskey, A. Rosenberg, and J. Hirschberg. Intonational phrases for speech summarization. In *Interspeech*, 2008.

[132] K. Mckeown and S. Pan. Prosody modeling in concept-to-speech generation: Methodological issues. *Philosophical Transactions of the Royal Society*, 358:1419–1431, 2000.

[133] P. Mermelstein. Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, 58(4):880–883, 1975.

[134] R. Metusalem and K. Ito. The effect of contrastive accent in discourse construction. In *21th Annual CUNY Conference*, 2008.

[135] A. C. Morris and H. Misra. Confusion matrix based posterior probabilities correction. IDIAP-RR 53, IDIAP, 2002.

[136] K. Murray. A study of automatic pitch tracker doubling/halving "errors". In *SIGDIAL Workshop on Discourse and Dialog*, 2001.

[137] M. Nakai, S. Harald, Y. Sagisaka, and H. Shimodaira. Automatic prosodic segmentation by f0 clustering using superpositional modeling. In *ICASSP*, 1995.

[138] C. Nakatani, J. Hirschberg, and B. Grosz. Discourse structure in spoken language: Studies on speech corpora. In *AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995.

[139] A. Nenkova, J. Brenier, A. Kothari, S. Calhoun, L. Whitton, D. Beaver, and D. Jurafsky. To memorize or to predict: Prominence labeling in conversational speech. In *NAACL-HLT*, 2007.

[140] L. Neumeyer, H. Franco, M. Weintraub, and P. Price. Automatic text-independent pronunciation scoring of foreign language student speech. In *ICSLP*, 1996.

[141] A. Niemistö, V. Lukin, I. Shmulevich, and O. Yli-Harja. Correction of misclassification in recognition of local image features with different classifiers. In *Finnish Signal Processing Symposium*, 2003.

[142] D. Oliver. Deriving pitch accent classes using automatic f0 stylisation and unsupervised clustering techniques. In *HLT*, 2005.

[143] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel. The boston university radio news corpus. Technical Report ECS-95-001, Boston University, March 1995.

[144] M. Ostendorf and K. Ross. Multi-level recognition of intonation labels. In Y. Sagisaka, N. Campbell, and N. Higuchi, editors, *Computing Prosody: Computational Models for Processing Spontaneous Speech*. Springer, 1997.

[145] M. Ostendorf and N. Veilleux. A hierarchical stochastic model for automatic prediction of prosodic boundary location. *Comput. Linguist.*, 20(1):27–54, 1994.

[146] M. Ostendorf, C. Wightman, and N. Veilleux. Parse scoring with prosodic information: An analysis/synthesis approach. *Computer Speech and Language*, pages 193–210, 1993.

[147] D. D. Palmer, M. Reichman, and E. Yaich. Feature selection for trainable multilingual broadcast news segmentation. In *HLT/NAACL 2004*, 2004.

[148] R. J. Passonneau and D. J. Litman. Discourse segmentation by human and automated means. *Computational Liunguistics*, 23(1):103–109, 1997.

[149] L. Pevzner and M. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

[150] T. Pfau and G. Ruske. Estimating the speaking rate by vowel detection. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 945–948, 1998.

[151] J. Pierrehumbert. The perception of fudamental frequency declination. *Journal of the Acoustical Society of America*, 66:363–369, 1979.

[152] J. Pierrehumbert. *The phonology and phonetics of English intonation*. PhD thesis, MIT, 1980.

[153] J. Pierrehumbert. Phonological and phonetic representation. *Journal of Phonetics*, 18:375–394, 1990.

[154] J. Pierrehumbert and J. Hirschberg. The meaning of intonational contours in the interpretation of discourse. In *Intentions in communication*. MIT Press, 1990.

[155] J. Pierrehumbert and S. Steele. How many rise-fall-rise countours? *ISPhS*, 1987.

[156] J. Pitrelli, M. Beckman, and J. Hirschberg. Evaluation of prosodic transcription labeling reliability in the tobi framework. In *ICSLP*, 1994.

[157] J. Platt. Machines using sequential minimal optimization. In B. Schoelkopf and C. Burges, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1998.

[158] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[159] W. J. Poser. *The Phonetics and Phonology of Tone and Intonation in Japanese*. PhD thesis, Department of Linguistics, MIT, 1984.

[160] S. Prevost and M. Steedman. Generating contextually appropriate intonation. In *EACL*, pages 332–340, Utrecht, 1993.

[161] P. Price, M. Ostendorf, and C. Wightman. Prosody and parsing. In *DARPA Speech Natural Language Workshop*, 1989.

[162] E. F. Prince. Toward a taxonomy of give-new infomration. In P. Cole, editor, *Radical Pragmatics*, pages 223–255. Academic Press, 1981.

[163] C. Prolo. Handling unlike coordinated phrases in tag by mixing syntactic category and grammatical function. In *Proc. of the 8th International Workshop on Tree Adjoining Grammar and Related Formalisms*, pages 137–140, 2006.

[164] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[165] I. Read and S. Cox. Automatic pitch accent prediction for text-to-speech synthesis. In *Interspeech*, 2007.

[166] I. Read and S. Cox. Stochastic and syntactic techniques for predicting phrase breaks. *Computer Speech & Language*, 21(3):519–542, 2007.

[167] Y. Ren, S.-S. Kim, M. Hasegawa-Johnson, and J. Cole. Speaker-independent automatic detection of pitch accent. In *Speech Prosody*, 2004.

[168] J. Reynar and A. Ratnaparkhi. A maximum entropy approach to identifying sentence boundaries. In *Fifth Conference on Applied Natural Language Processing*, 1997.

[169] A. Rosenberg and J. Hirschberg. On the correlation between energy and pitch accent in read english speech. In *Interspeech*, 2006.

[170] A. Rosenberg and J. Hirschberg. Detecting pitch accent using pitch-corrected energy-based predictors. In *Interspeech*, 2007.

[171] A. Rosenberg and J. Hirschberg. Varying input segmentation for story boundary detection in english, arabic and mandarin broadcast news. In *Interspeech*, 2007.

[172] A. Rosenberg and J. Hirschberg. Detecting pitch accents at the word, syllable and vowel level. In *HLT-NAACL*, 2009.

[173] K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech & Language*, 10(3):155–185, 1996.

[174] J. L. Rouas. Automatic prosodic variations modeling for language and dialect discrimination. *IEEE Transactions on Audio, Speech and Language Processing*, 15(6):1904–1911, August 2007.

[175] D. Ruta and B. Gabrys. An overview of classifier fusion methods. *Computing and Information Systems*, 7(1):1–10, 2000.

[176] R. E. Schapire. The strength of weak learnability. *Foundations of Computer Science*, 1990.

[177] H. Schmid and M. Atterer. New statistical methods for phrase break prediction. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 659, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[178] E. Selkirk. Contrastive focus vs. presentational focus: Prosodic evidence from right node raising in english. In *Speech Prosody*, pages 643–646, 2002.

[179] M. Selting. On the interplay of syntax and prosody in the constitution of turn constructional units and turns in conversation. *Pragmatics*, 6:357–388, 1996.

[180] E. Shriberg, R. Bates, A. Stolcke, P. Taylor, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, 41:443–492, 1998.

[181] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1-2):127–154, 2000.

[182] R. Silipo and S. Greenberg. Prosodic stress revisited: Reassessing the role of fundamental frequency. In *NIST Speecn Transcription Workshop*, 2000.

[183] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi: A standard for labeling english prosody. In *Proc. of the 1992 International Conference on Spoken Language Processing*, volume 2, pages 12–16, 1992.

[184] K. Silverman, A. Kalyanswamy, J. Silverman, S. Basson, and D. Yashchin. Syntehsiser intelligibility in the context of a name-and-address information service. In *Eurospeech*, pages 2169–2172, 1993.

[185] A. M. C. Sluijter and V. J. van Heuven. Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4):2471–2485, 1996.

[186] A. M. C. Sluijter, V. J. van Heuven, and J. J. A. Pacilly. Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1):503–513, 1997.

[187] V. R. Sridhar, S. Bangalore, and S. Narayanan. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. *IEEE Transactions on Audio, Speech & Language Processing*, 16(4):797–811, 2008.

[188] V. R. Sridhar, S. Narayanan, and S. Bangalore. Acoustic-syntactic maximum entropy model for automatic prosody labeling. In *IEEE-ACL Conference on Spoken Language Technology*, 2006.

[189] V. R. Sridhar, S. Narayanan, and S. Bangalore. Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework. In *HLT-NAACL*, 2007.

[190] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lei, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, and Q. Zhu. Recent innovations in speech-to-text transcription at sri-icsi-uw. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1729–1744, 2006.

[191] A. Stolcke, E. Shriberg, D. Hakkani-Tür, and G. Tür. Modeling the prosody of hidden events for improved word recognition. In *Eurospeech*, 1999.

[192] A. Stolcke, E. Shriberg, G. Tur, and K. Sonmez. Combining words and speech prosody for automatic topic segmentation. In *In Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, pages 61–64, 1999.

[193] S. Strassel and M. Glenn. Creating the annotated tdt-4 y2003 evaluation corpus. http://www.nist.gov/speech/tests/tdt/tdt2003/papers/ldc.ppt, 2003.

[194] X. Sun. Pitch accent predicting using ensemble machine learning. In *ICSLP*, 2002.

[195] A. Syrdal, J. Hirschberg, J. McGory, and M. Beckman. Automatic tobi prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33(1–2):135–151, January 2001.

[196] A. Syrdal and J. McGory. Inter-transcriber reliability of tobi prosodic labeling. In *ICSLP*, 2000.

[197] F. Tamburini. Automatic prosodic prominence detection in speech using acoustic features: an unsupervised system. In *Eurospeech*, 2003.

[198] F. Tamburini. Prosodic prominence detection in speech. In *Proc. 7th International Symposium on Signal Processing and its Applications = ISSPA2003*, pages 385–388, 2003.

[199] F. Tamburini. Automatic prominence identification and prosodic typology. In *Proc. InterSpeech 2005*, pages 1813–1816, 2005.

[200] P. Taylor. The rise/fall/connection model of intonation. *Speech Commun.*, 15(1-2):169–186, 1994.

[201] P. Taylor. The tilt intonation model. In *ICSLP*, 1998.

[202] P. Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America*, 2000.

[203] P. Taylor, R. Caley, A. Black, and S. King. Edinburgh speech tools library with documentation edition 1.2. Technical report, The University of Edinburgh, 1999.

[204] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sonmez. Evaluation of speaker's degree of nativeness using text-independent prosodic features. In *Proceedings of the Workshopo on Multilingual Speech and Language Processing*, 2001.

[205] C. Teixeira, H. Franco, E. Shriberg, K. Sönmez, and K. Precoda. Prosodic features for automatic text-independent evaluation of nativeness for language learners. In *ICSLP*, 2000.

[206] J. Tepperman, A. Kazemzadeh, and S. Narayanan. A text-free approach to assessing nonnative intonation. In *Interspeech*, 2007.

[207] J. Tepperman and S. Narayanan. Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners. In *Proc. ICASSP*, volume 1, pages 937–940, 2005.

[208] J. Tepperman and S. Narayanan. Tree grammars as models of prosodic structure. In *Interspeech*, 2008.

[209] J. Tepperman, D. Traum, and S. Narayanan. Yeah right: Sarcasm recognition for spoken dialogue systems. In *ICSLP*, 2006.

[210] J. Terken and J. Hirschberg. Deaccentuation of words representing 'given' information: Effects of persistence of grammatical function and surface position. *Language and Speech*, 37(2):125–145, 1994.

[211] S. Toki and M. Murata. *Pronunciation & Task Listening-Innovative Workbooks in Japanese*. Aratake Publishers, 1987.

[212] R. Tong, B. ma, D. Zhu, H. Li, and E. S. Chng. Integrating acoustic, prosodic and phonotactic features for spoken language identification. In *ICASSP*, 2006.

[213] K. Toutanova and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC-2000*, 2000.

[214] G. Tür, D. Hakkani-Tür, A. Stolcke, and E. Shriberg. Integrating prosodic and lexical cues for automatic topic segmentation. *Computational Linguistics*, 27:31–57, 2001.

[215] A. Tyler. Discourse structure and the perception of incoherence in international teaching assistants' spoken discourse. *TESOL Quarterly*, 26(4):713–729, 1992.

[216] C. J. Van Rijsbergen. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.

[217] N. M. Veilleux. *Computational models of the prosody/syntax mapping for spoken language systems*. PhD thesis, Boston University, Boston, MA, USA, 1994. Major Professor-Ostendorf,, Mari.

[218] N. M. Veilleux and M. Ostendorf. Probabilistic parse scoring based on prosodic phrasing. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 429–434, Morristown, NJ, USA, 1992. Association for Computational Linguistics.

[219] R. Villing, J. Timoney, T. Ward, and J. Costello. Automatic blind syllable segmentation for continuous speech. In *ISSC*, volume 2004, pages 41–46. IEE, 2004.

[220] T. Vogt, E. André, and N. Bee. Emovoice – a framework for online recognition of emotions from voice. In *PIT '08: Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems*, pages 188–199, Berlin, Heidelberg, 2008. Springer-Verlag.

[221] A. Waibel. Recognition of lexical stress in a continuous speech understanding system - a pattern recognition approach. In *ICASSP*, volume 11, pages 2287–2290, 1986.

[222] M. Q. Wang and J. Hirschberg. Predicting intonational boundaries automatically from text: the atis domain. In *HLT '91: Proceedings of the workshop on Speech and Natural Language*, pages 378–383, Morristown, NJ, USA, 1991. Association for Computational Linguistics.

[223] M. Q. Wang and J. Hirschberg. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6(2):175–196, 1992.

[224] G. Ward and J. Hirschberg. Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, 61:747–776, 1985.

[225] N. Ward and W. Tsukuhara. Prosodic features which cue back-channel responses in english and japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.

[226] P. Warren. Patterns of late rising in new zealand english: intonational variation or intonational change? *Language Variation and Change*, 17:209–230, 2005.

[227] C. L. Wayne. Multilingual topic detection and tracking: Successful research enabled by corpora and evaluation. In *LREC*, pages 1487–1494, 2000.

[228] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg. Using prosodic and lexical information for speaker identification. In *ICASSP*, 2002.

[229] G. M. Weiss and F. Provost. The effect of class distribution on classifier learning. Technical report, Department of Computer Science, Rutgers University, 2001.

[230] C. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processig*, 2(4), October 1994.

[231] C. Wightman, M. Ostendorf, and S. Shattuck-Hufnagel. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustic Society of America*, 1992.

[232] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham. Weka: Practical machine learning tools and techniques with java implementation. In *ICONIP/ANZIIS/ANNES International Workshop: Emerging Knowledge Engineering and Connectionist-Based Information Systems*, pages 192–196, 1999.

[233] C. Wooters, J. Fung, B. Peskin, and X. Anguera. Towards robust speaker segmentation: The icsi-sri fall 2004 diarization system. In *RT-04F Workshop*, November 2004.

[234] G. Xydas, D. Spiliotopoulos, and G. Kouroupetroglou. Modeling prosodic structures in linguistically enriched environments. In *in "Text, Speech and Dialogue", Lecture Notes in Artificial. Intelligence. (LNAI), Springer-Verlag Berlin Heidelberg, Vol 3206*, pages 521–528. Springer, 2004.

[235] R. Yan, Y. Liu, R. Jin, and A. Hauptmann. On predicting rare cases with svm ensembles in scene classification. In *ICASSP*, 2003.

[236] T. Yoon. Predicting prosodic phrasing using linguistic features. In *Speech Prosody*, 2006.

[237] T. Yoon, S. Chavarria, J. Cole, and M. Hasegawa-Johnson. Intertranscriber reliability of prosodic labeling on telephone conversation using tobi. In *ICSLP*, 2004.

[238] K. Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Research and Development in Information Retrieval*, 2001.

[239] P. Zervas, M. Maragoudakis, N. Fakotakis, and G. Kokkinakis. Bayesian induction of intonational phrase breaks. In *Eurospeech*, 2003.

[240] T. Zhang, M. Hasegawa-Johnson, and S. E. Levinson. Automatic detection of contrast for speech understanding. In *Interspeech*, 2004.

[241] X. Zhu and G. Penn. Roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In *HLT-NAACL*, 2006.