Using Prosody and Phonotactics in Arabic Dialect Identification

Fadi Biadsy, Julia Hirschberg

Department of Computer Science, Columbia University, New York, NY, 10027
{fadi,julia}@cs.columbia.edu

Abstract

While Modern Standard Arabic is the formal spoken and written language of the Arab world, dialects are the major communication mode for everyday life; identifying a speaker's dialect is thus critical to speech processing tasks such as automatic speech recognition, as well as speaker identification. We examine the role of prosodic features (intonation and rhythm) across four Arabic dialects: Gulf, Iraqi, Levantine, and Egyptian, for the purpose of automatic dialect identification. We show that prosodic features can significantly improve identification, over a purely phonotactic-based approach, with an identification accuracy of 86.33% for 2m utterances.

1. Introduction

In past years considerable attention has been paid to the automatic classification of languages using acoustic information alone. That is, how can we identify the language a speaker is speaking from the acoustic signal alone? In recent years, the identification of regional accents and dialects has attracted interest from the speech community. Can a speaker's regional origin or regional dialect within a given language group be determined given a small sample of his or her speech?

Our goal is to identify the dialect of a speaker from among the four colloquial Arabic dialects: Iraqi, Gulf, Levantine, and Egyptian in aid of improving Automatic Speech Recognition (ASR). Since speakers with different dialects often pronounce some words differently, consistently altering certain phones and even morphemes, identifying regional dialect prior to ASR allows for the use of a more restricted pronunciation dictionary in decoding, resulting in a reduced search space and lower perplexity. Moreover, identifying the dialect first will enable the ASR system to adapt its acoustic, morphological, and language models appropriately. In previous studies we have presented experiments using a phonotactic modeling approach based on the parallel Phone Recognition followed by Language Modeling (PRLM) [1] to distinguish Arabic dialects among themselves and from Modern Standard Arabic (MSA) [2]. In this work, we focus our attention on identifying prosodic differences across the four Arabic dialect to improve this classification.

In Section 2, we describe related work in language and dialect ID. In Section 3, we describe the Arabic dialect corpora employed in our experiments. In Section 4, we describe some global prosodic differences among the four dialects. We model sequential prosodic features in Section 5. We present our system and experimental results in Section 6. Finally, we conclude in Section 7 and identify directions for future research.

2. Related Work

Some of the most successful approaches to language ID have made use of phonotactic variation. For example, the parallel Phone Recognition followed by Language Modeling (parallel PRLM) approach uses phonotactic information to identify languages from the acoustic signal alone [1]. We have used the parallel PRLM using 9 phone recognizers trained on different languages to distinguish among the four Arabic dialects we examine in this work, as well as MSA [2]. Using phonotactic information, we have obtained an accuracy in four-way classification of 78.5% (using 30s test utterances) and 84.0% (using 2 minute utterances). An ergodic HMM was used to model phonetic differences between two Arabic dialects (Gulf and Egyptian Arabic) employing standard MFCC (Mel Frequency Cepstral Coefficients) and delta features. [3]

Intonational cues have been shown to be useful to human subjects asked to identify regional dialects, with subjects able to distinguish between Western and Eastern Arabic dialects significantly above chance based on intonation alone [4]. It has been also showed that rhythmic differences exist between Western and Eastern Arabic.[5] The analysis of these differences was done by comparing percentages of vocalic intervals (%V) and the standard deviation of intervocalic intervals (ΔC) across the two groups. These features are thought to capture the complexity of the syllabic structure of a language/dialect in addition to degree of vowel reduction. Such features appear to correlate with the rhythmic structure of a language or dialect, and thus may be good cues for language and dialect ID [6].

In this work, we extract local prosodic features at the level of *pseudo syllables* similar to [7, 8] for use in dialect ID. Where previous research discretizes prosodic values to short/long for syllable durations and up/down for F0 values, we model prosodic features as continuous values in an HMM to capture subtle sequential prosodic differences of the entire spectrum without the need for explicit thresholding.

3. Corpora

When training a system to classify languages or dialects, it is important to use training and testing corpora recorded under similar acoustic conditions. We use corpora of spontaneous telephone conversations from the Linguistic Data Consortium (LDC) with similar recording conditions produced by native speakers of the dialects, speaking with family members, friends, and unrelated individuals, sometimes about predetermined topics for Gulf Arabic, Iraqi Arabic, Egyptian Arabic, and Levantine Arabic. Although, the data have been annotated phonetically and/or orthographically by LDC, we do not make use of these annotations for our work.

We use speech from 965 speakers ($\sim$41.02h) from the Gulf Arabic Conversational Telephone Speech corpus [9], holding out out 150 speakers for testing. We use 475 speakers ($\sim$25.73h) from the Iraqi Arabic Conversational Telephone Speech database [9] the Iraqi dialect, again holding out 150 speakers for testing. Our Levantine data consists of 1258 speakers from the Arabic CTS Levantine Fisher Training Data Set 1-3 ($\sim$79h) [10], with 150 speakers from Set 1 held out for testing.

For our Egyptian data, we use CallHome Egyptian and its Supplement [11] and CallFriend Egyptian [12]. We use 398 speakers from these corpora (∼75.7h), holding out 150 speakers for testing. For all our experiments in this paper we use the first 2m of speech from each held-out speaker.

4. Prosodic Differences Across Dialects

In this section, we identify *global prosodic features* that differ significantly across our four Arabic dialects. We randomly select 398 speakers from each dialect corpus and examine the first 2m of speech from each speaker. We first segment the speech files based on silence. We assume that each non-silent segment is a valid *speech segment*; inspection of a random sample of the output of this process shows this assumption to be reasonable. Since several of our prosodic features are calculated at the syllable level, we next syllabify the speech segments. Since, to our knowledge, there are no automatic syllabification systems for Arabic dialects that require only acoustic information, dialects, we employ a pseudo-syllabification approach which has been employed in previous work [7, 8]. We define a pseudo syllable as a cluster of optional consonants followed by a single vowel (i.e., C*V). To identify vowels and consonants, we run an open-loop phone recognizer trained on MSA, mapping all six MSA vowels to V and all other phones to C [2]. Note that we have time boundaries of the syllables from our phone recognizer.

4.1. Pitch Features Across Dialects

To test whether dialects differ in their pitch variation, we compute pitch range for each speaker by first Z-normalizing the entire F0 contour and then computing the average of the F0 maxima in all the speaker's segments. Using the normalized F0 contour, we also compute the pitch register across dialects; it is computed as the average of the difference between the F0 maximum and F0 minimum over all the speech segments of the speaker. Similarly, we extract the average of the F0 minimum of all speech segments of the speaker. We also compute the standard deviation of the entire (unnormalized) F0 contours of the speaker to test if one dialect employes more dynamic intonational contours than other dialects.

Previous work has suggested that H peaks may align earlier in Egyptian formal Arabic (within the stressed syllable) than in Egyptian colloquial Arabic [13]. To test whether Arabic dialects differ in the alignment of the pitch peaks to syllables, we compute the mean of the location of pitch maxima in all syllables. (Currently, we do not attempt to distinguish stressed syllables from unstressed.) All location values are from the onset of the syllable, so the location is a value between 0 to 1.

We then compare these prosodic features for each pair of dialects, using Welch's t tests. Table 1 shows the differences we have observed in the data. We see from these results that Levantine and Iraqi speakers tend to speak with higher pitch range and more expanded pitch register than Egyptian and Gulf speakers. In addition, Gulf speakers tend to use a more compressed pitch register than Egyptian speakers. Moreover, Iraqi and Gulf intonation shows more variation than Egyptian and Levantine. Nonetheless, the intonational contours of Levantine speakers vary significantly more than that of Egyptian speakers. Pitch peaks within pseudo-syllables in Egyptian and Iraqi are shifted significantly later than the pitch peaks in Gulf and Levantine. However, Levantine speakers tend to shift their pitch peaks earlier in syllables than do Gulf speakers.

4.2. Durational and Rhythmic Features Across Dialects

We compare dialects' timing features using [6]'s rhythmic features (see Section 2), (ΔC, %V, and ΔV). We also want to test the effect of speaking rate on distinguishing our Arabic dialects. Speaking rate is computed here as the number of pseudo syllables per second. Again, we use Welch's t test to indicate significant differences in features between each dialect pair. Table 1 again shows our results. Assuming that our automatically obtained pseudo-syllables are good approximation of the true syllables, we may conclude that Gulf and Iraqi dialects tend to have more complex syllabic structure. Egyptian and Levantine tend to have more variation in their vocalic intervals, which correlates to the existence of vowel reduction. These features suggest that some of these dialects do in fact differ in their rhythmic structure, empirical confirmation of previous phonological hypotheses. We also see that Egyptian speakers are the fastest speakers followed by Gulf speakers. Iraqi and Levantine are the slowest speakers, with comparable rates.

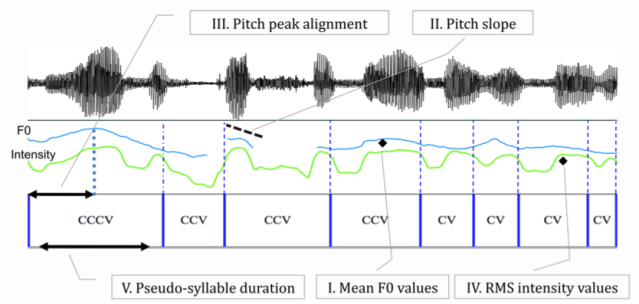

Figure 1: *Local prosodic features extracted from the pseodo-syllables in a speech segment*

5. Modeling Prosodic Patterns

Although we have found major differences between dialects in prosodic and rhythmic variation, we suspect that the global features described above are not enough for modeling aspects of the prosodic structure of a dialect. These features do not, for example, capture specific contextual, segmental and sequential patterns, such as the shape of intonational contours and the distribution of different contour types in a dialect. We believe that modeling sequences of local prosodic features using sequential models, such as Hidden Markov models (HMMs), may be more effective in modeling the prosodic patters of a dialect.

5.1. Sequential Prosodic Feature Modeling

To model sequential prosodic structure, we extract *five* different sequences from each speech segment in our training data: mean F0, pitch slope, pitch peak alignment, RMS intensity, and duration. Each sequence consists of two-dimensional feature vectors extracted from prosodic data within pseudo syllables. These features are illustrated in Figure 1 and described below.

To test whether dialects differ in their characteristics of their intonational contours, we extract three types of sequences from the Z-Normalized F0 contour. We calculate the mean of the F0 values within each pseudo syllable and compute the *deltas* of these means to approximate the first derivative of the F0 contour (this feature is denoted as **I** in Figure 1); we define delta here as the difference between each two consecutive values. To model pitch slope, we fit a linear regression given the values of the Z-normalized F0 contour in each pseudo syllable, and extract the angle of the regression line (denoted as **II** in the figure). We also add the deltas of these angles. For pitch peak alignment, we extract the location in time (starting from the onset of the syllable) of the F0 peak within pseudo syllables (denoted as **III**). The val-

Table 1: *Comparing global prosodic features between dialect pairs. X* indicates that dialect X has a greater mean for that feature than does the other dialect, with significance level of 0.05, ** with 0.01 and *** with 0.001*

Dialect 1	Dialect 2	Pitch Register	Pitch Range	Pitch Min	Pitch Sdv	Pitch Peak Alignment	ΔC	ΔV	%V	Speaking Rate
Gulf	Iraqi	I***	I***	G*	0.12	I***	0.24	G***	G***	G***
Gulf	Levantine	L***	L***	0.52	G .07	G**	G*	0.34	0.66	G**
Gulf	Egyptian	E***	0.49	G***	G***	E***	G***	E**	E***	E**
Iraqi	Levantine	0.64	L .067	I*	I***	I***	I***	L***	L***	0.16
Iraqi	Egyptian	I***	I***	0.056	I***	0.2	I***	I***	E***	E***
Levantine	Egyptian	L***	L***	E***	L***	E***	L***	0.10	E***	E***

ues of these features are between 0 and 1. We also compute the delta of these locations.

Intensity features play an important role in prosodic events [14]. Therefore, for each speech segment, we first Z-normalize the intensity contour and then extract the RMS (Root Mean Square) of the intensity values within pseudo-syllables (denoted as **IV** in the figure). We also add the deltas of these RMS intensity features.

As mentioned above, Arabic dialects have been shown to differ in their rhythmic structure. We approximate the rhythm of a dialect by modeling the sequence of the log of the duration of each pseudo syllable (denoted as **V**). Similar to the other sequences, the delta of these log durations is included in the feature vector. This modeling of rhythm is somewhat similar to [8], but that work modeled rhythm using a joint multinomial distribution of two consecutive durations instead of an HMM of log durations and deltas.

5.2. HMM Settings

For each dialect, we model each of the five sequence types mentioned above using a continues HMM with Gaussian mixture state observation densities with diagonal covariance matrices for all Gaussian components. The state transition matrix (A) and initial state distributions (Pi) in all HMMs are initialized uniformly, and the Gaussian mixture components of all the states are initialized by running k-mean clustering first, using the training data described in Section 3. The number of states and number of Gaussians are determined empirically. For all the F0 HMMs (I–III), we use four hidden states with one Gaussian per state. For the intensity HMMs (IV), we use six states and two Gaussian components per state, and for the durational HMMs (V), we use 3 states and one Gaussian per state. We have an HMM for each pair of dialect and sequence type. Since we analyze four dialects in this paper and five sequence types, we have 20 HMMs in total. All HMMs are trained using the Baum-Welch algorithm on the training data in Section 3. We use the HMM Matlab toolkit [15] for training and decoding.

6. Dialect Identification Results

We have shown in previous work that the same four Arabic dialects we analyze in this work can be identified using a phonotactic approach with considerable success, particularly using the parallel Phone Recognition followed by Language Modeling approach (parallel PRLM) [1, 2]. In this section, we describe a system for identifying the four Arabic dialects using the global and sequential prosodic features described above which we then compare to our parallel PRLM system. Finally, we combine these two systems to see if prosodic features provide information that phonotactics does not.

We first evaluate the effectiveness of the global features described in Section 4 in dialect identification. We use 150 speakers form each dialect, to train a logistic classifier that uses only the nine global features. Four-way 10-fold cross-validation classification shows that, with these features, we obtain an accuracy of 54.83%. F-Measures of the classes are shown in Table 2; the chance baseline is 25%. (Note that our parallel PRLM approach 10-cross validation accuracy is slightly different than that in our previous work (84.0%), since we use a different random permutation in our cross-validation experiment here.)

We have also observed that different dialects lengthen certain vowels more than others, so we include the mean and standard deviation of the durations of each vowel type from a speaker as features in our classifier. When we analyzed the errors of our phone recognizer, we also observed that glottal stops and vowels are often confused, so we also include the duration and standard deviation of glottal stop durations as well. Thus, we have fourteen additional features: the mean and standard deviation of 6 vowels and the glottal stop phone. When we add these duration features we obtain a significant increase in accuracy 60%. The F-measures also show some increase, as shown in Table 2. It should be noted that the vowel duration features do not perform well alone; the accuracy of the dialect identification system using the fourteen features alone is only 44.16%.

To test the usefulness of our sequential prosodic features on dialect identification, we extract the feature-vector sequences of each sequence type from each dialect and train an HMM on the training corpus for each of our dialects. In total, we have 20 (4 dialects x 5 sequence types) HMMs. Given a speaker's utterance, we extract each sequence type and compute the likelihood of this sequence given each of the five corresponding HMMs. We normalize these likelihoods by the sequence length.

Instead of identifying the dialect of a held-out utterance by, for example, simply identifying the dialect associated with HMMs which produce the highest average likelihood over it, we make use of the normalized likelihoods of all HMMs by treating them as a feature vector (4 dialects x 5 HMMs = 20 features). Using a four-way logistic classifier to identify dialect, we report 10-fold cross-validation results over the 600 speakers held out from HMM training. Using this back-end classifier significantly improves identification accuracy over simpler methods of combining likelihoods. Using the sequential prosodic features alone we obtain an accuracy of 64.33% compared to relying on average likelihoods, which produces a classification accuracyof 38.0%. When we add the global prosodic features, we obtain an accuracy of 72% (Table 2) – a significant increase.

6.1. Dialect ID with Phonotactic Features

As noted earlier, we have previously shown that the parallel PRLM approach [1] is effective in identifying Arabic dialects [2]. We used 9 phone recognizers trained on different languages to produce 9 phone streams for the training data of each dialect. We then trained a trigram model for each stream for each dialect. During testing, we ran all phone recognizers on the test utterance and computed the perplexity of each trigram model on the corresponding output phone sequence. Finally, the perplexities were fed to a back-end logistic classifier to determine the hypothesized dialect. The results of identifying the four Arabic dialects are shown in Table 2; the accuracy is 83.5%.

Table 2: *The four-way 10-fold cross-validation dialect-ID results for our 600 speakers, with different feature sets; F_1 is the F-Measure*

Feature Type	Accuracy (%)	Gulf (F_1)	Iraqi (F_1)	Levantine (F_1)	Egyptian (F_1)
Chance baseline	25.0	-	-	-	-
Nine global prosodic features	54.8	41.2	53.6	56.5	65.3
+ Vowel duration mean & sdv.	60.0	52.7	57.1	62.8	66.9
+ Sequential prosodic modeling	72.0	68.9	66.4	72.9	79.2
Phonotactic classifier (Parallel PRLM only)	83.5	74.7	75.7	88.4	95.2
Phonotactic & prosodic features (one classifier)	81.5	74.1	74.6	86.3	90.2
Combining phonotactic & prosodic classifiers	86.3	79.5	81.5	89.5	94.9

6.2. Combining the Phonotactics and Prosodic Features

So we see that prosodic features when used alone are valuable features for identifying Arabic dialects. We also observe that phonotactic features are superior at distinguishing dialects. Now we examine whether prosodic features add new information that may improve dialect classification. If so, how can we best combine phonotactic and prosodic information?

Recall that we have two back-end logistic classifiers, one for the phonotactic approach and another for the prosodic approach. If, instead of training the two separately, we train a single classifier that includes both phonotactic and prosodic information, we obtain an accuracy of 81.5% – somewhat lower than the accuracy of the phonotactic classifier alone. (We speculate that the reason for this lower performance may be a data sparsity issue, since we increase the feature dimensionality but still perform 10-fold cross-validation on 600 instances.) So, instead of training one classifier that combines all features, we instead combine the posterior probability distribution of the two classifiers. We combine these posteriors by multiplying the posterior probabilities and then returning the class with the maximum score; this approach outperforms the sum and max combinationa strategies [16]. Using this approach, we obtain a significant (p=.022) increase in accuracy (86.33%) over the phonotactic approach alone (Table 2). We have also validated this statistical significance using 15, 25 and 50 -fold cross-validation.

Note that the percentage of instances that are *in*correctly classified by the phonotactic classifier but *correctly* classified by the prosodic classifier is 9.5%. Thus, the upper bound accuracy that could be obtained by using the phonotactic and the prosodic classifiers together would be 93% (9.5 + 83.5). In future work we will explore additional methods of classifier combination. We also observe that the most distinguishable dialect among our four dialects is Egyptian, followed by Levantine, and the most confusable dialect pairs are Iraqi and Gulf Arabic – not surprising since some scholars classify Iraqi Arabic as a subdialect of Gulf.

7. Discussion and Future Work

We have shown empirically that four Arabic dialects, Gulf, Iraqi, Levantine, and Egyptian, exhibit significant differences from one another in terms of characteristics of their prosodic structure, including pitch range, register, and pitch dynamics, as well as differences in their rhythmic structure, speaking rate, and vowel durations. We have demonstrated that we can utilize these prosodic features to automatically identify the dialect of a speaker with considerable accuracy. Modeling sequences of local prosodic features at the level of pseudo syllables using HMMs significantly improves accuracy when combined with global prosodic features. This approach can also significantly improve our previous system, which used only phonotactic features, resulting in an accuracy of 86.33% on 2m utterances.

We have observed that dialects also appear to differ in speakers' production of prosodic events, such as phrasing and prominence type. In future work, we will use automatic prosodic event detection techniques [17] to see whether such features can improve dialect identification. We have also observed that the more difficult it is to classify a dialect using the phonotactic approach, the more difficult it is to classify it using only prosodic features. Since phonotactics and prosody use different streams of data, we plan to investigate the relationship between phone sequence distribution and the prosody of Arabic dialects in more detail to see if we can leverage such a relationship for improved dialect identification.

8. Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-06-C-0023 (approved for public release, distribution unlimited). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

9. References

[1] M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions of Speech and Audio Processing*, vol. 4, no. 1, 1996.

[2] F. Biadsy, J. Hirschberg, and N. Habash, "Spoken Arabic Dialect Identification Using Phonotactic Modeling," in *Proceedings of EACL 2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, 2009.

[3] F. S. Alorfi, "PhD Thesis: Automatic Identification Of Arabic Dialects Using Hidden Markov Models," in *University of Pittsburgh*, 2008.

[4] M. Barkat, J. Ohala, and F. Pellegrino, "Prosody as a Distinctive Feature for the Discrimination of Arabic Dialects," in *Proceedings of Eurospeech'99*, 1999.

[5] R. Hamdi, M. Barkat-Defradas, E. Ferragne, and F. Pellegrino, "Speech Timing and Rhythmic Structure in Arabic Dialects: A Comparison of Two Approaches," in *Proceedings of Interspeech'04*, 2004.

[6] F. Ramus, "Acoustic Correlates of Linguistic Rhythm: Perspectives," in *Speech Prosody*, 2002.

[7] J. Rouas, "Automatic Prosodic Variations Modeling for Language and Dialect Discrimination," in *IEEE Transactions On Audio, Speech, and Language Processing*, vol. 15, 2007.

[8] E. Timoshenko and H. Hoge, "Using Speech Rhythm for Acoustic Language Identification," in *Proceedings of Interspeech 2007*, 2007.

[9] Appen Pty Ltd, "Gulf and Iraqi Arabic Conversational Telephone Speech Linguistic Data Consortium, Philadelphia," 2006.

[10] M. Maamouri, "Levantine Arabic QT Training Data Set 5, Speech Linguistic Data Consortium, Philadelphia," 2006.

[11] A. Canavan, G. Zipperlen, and D. Graff, "CALLHOME Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia," 1997.

[12] A. Canavan and G. Zipperlen, "CALLFRIEND Egyptian Arabic Speech Linguistic Data Consortium, Philadelphia," 1996.

[13] S. Hellmuth and D. El Zarka, "Variation in phonetic realization or in phonological categories? Intonational pitch accents in Egyptian Colloquial Arabic and Egyptian Formal Arabic," in *Proceedings of 16th ICPhS*, 2007.

[14] A. Rosenberg and J. Hirschberg, "On the correlation between energy and pitch accent in read english speech," in *Interspeech*, 2006.

[15] K. Murphy, "Hidden markov model (hmm) toolkit for matlab," in *www.cs.ubc.ca/ murphyk/software/hmm/hmm.htm*, 2004.

[16] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On Combining Classifiers," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, 1998.

[17] A. Rosenberg, "PhD Thesis: Automatic Detection and Classification of Prosodic Events (to be published)," in *Columbia University*, May 2009.