Modeling Japanese boundary pitch movements for speech synthesis

Jennifer J. Venditti, Kazuaki Maeda and Jan P. H. van Santen

ABSTRACT This paper provides a detailed analysis of the intonational form and function of five boundary pitch movements (BPMs) in Tokyo Japanese. The perception study describes the linguistic and paralinguistic dimensions on which meanings of the BPMs are distinguished. The production study details how the F0 heights, rise shapes, segment durations, and (crucially) alignment of the F0 contour with the segments are all used to define the movement types. We suggest a quantitative model of F0 contour alignment which uses observed data to model the entire shape of F0 curves.

1 Background

The accentual phrase in Tokyo Japanese is characterized intonationally by an initial rise and subsequent fall (e.g. [FS71, PB88]). In lexically-accented phrases, the fall is steep (due to the accent) and declines nearly to the bottom of the speaker's range. In lexically-unaccented phrases the fall is more gradual, and does not reach the bottom. Though the rate of the fall differs, in both accent types the canonical accentual phrase shape is a rise followed by a decline to the right phrase edge.

However, there are cases in which the F0 movement at the phrase edge (here, *boundary pitch movement*, or 'BPM') is not a simple decline. Often the speaker's pitch can rise at the edge, or rise and fall again, to convey some type of linguistic or paralinguistic meaning. For example, consider the following sentences containing identical segments, but with different intonations (cued by punctuation here):

sô na no? asking "Is that true?"
sô na no! insisting "It's true!"

A boundary pitch movement which rises to a high level cues a question interpretation, while a sharp rise to a mid level gives the impression that the speaker is insisting. The difference in the shape and extent of the rise at the phrase edge provides a major cue to the meaning of the utterance. In this paper we will describe the intonational form and function of a number of different boundary pitch movements.

Both spoken language understanding and generation systems can benefit from a detailed model of boundary pitch movements. For speech synthesis,

this is important for advanced concept-to-speech systems, in which phrases can be tagged with attributes such as surprise, disbelief, assertion, emphasis, etc., which express pragmatic meaning beyond the meaning of words in the phrase. For text-to-speech synthesis as well, it is crucial to incorporate not only simple question rises, but also prominence-lending rises (occurring at the ends of syntactic units) which help the listener parse the stream of running speech (e.g. [MH94]).

Japanese speech synthesis systems differ in the way they treat boundary pitch movements. This is mainly due to differences in the intonation models which the systems implement. For example, in the superpositional model of Fujisaki [FS71], both rises shown in the sentences above are modeled by accent commands (of different amplitudes) which are truncated by the cessation of voicing at the end of the sentence [FH93, FOO+94]. In contrast, synthesizers using the Pierrehumbert-Beckman model or Japanese ToBI approach [PB88, Ven95] use a H% boundary tone as a target to which the F0 rises at the end of the phrase. The scaling of the H% height can be determined by parameter settings.

The problem with these two approaches is that currently neither pays much attention to the precise timing or shape of the different rises. In Japanese ToBI synthesizers, tone alignment is specified only to the extent that tones are phonologically associated with particular syllables or phrase edges, and are aligned in relation to these locations in a uniform way. Thus, the timing of the H% rise is the same regardless of the meaning that the rise cues. In addition, the shape of the rise is constrained by a uniform smoothing which is applied to the constructed F0 contour. In the Fujisaki model, in contrast, there is more flexibility in the timing of accent commands. Given a model of how the timing differs in each rise type, it would be possible to systematically vary the location of these phrase-final accent commands to produce the desired effect. However, the shape of the rise remains invariant, due to its dependence on the shape of the underlying rectangular accent command.

In this paper, we investigate five different boundary pitch movements in Tokyo Japanese. We start by providing data showing that the meanings of the BPM are perceptually distinct, then we describe in detail the F0 shape and alignment characteristics of each type. We conclude by suggesting a model of F0 contour alignment which is able to quantitatively describe the shape of the entire rise, based on the temporal characteristics of the segments with which the tune is aligned.

2 Intonational Meaning

This study examines the meaning of five boundary pitch movements, listed below. The abbreviations given in parentheses are those which will be used in the remainder of this paper.

- Incredulity question rise (*incredQ*)
- Information question rise (*infoQ*)
- Prominence-lending rise (*prom*)
- Insisting rise (*insist*)
- Rise-fall boundary movement (*rise-fall*)

Figure 1 shows a schematization of the different BPM shapes investigated in this study. The shapes here are based on impressionistic perceptual observations. A detailed description of the actual F0 shapes is presented in Section 3.2.

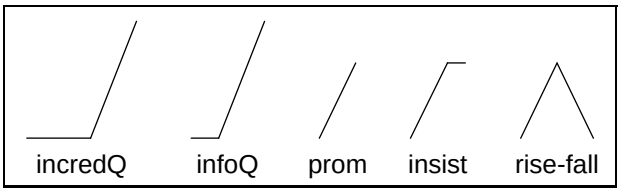

FIGURE 1. Schematic shapes of BPM types investigated in this study.

2.1 Comparison with previous BPM categorization

It is important to note that this is by no means an exhaustive inventory of boundary pitch movements in Tokyo Japanese. These five BPMs are simply those which we have chosen to be most relevant to our speech synthesis efforts, and which seem to be clearly distinct in either meaning or form. However, at the same time, this inventory is not drastically different from or smaller than the 'exhaustive' categorization of BPMs into five types which Kawakami proposes in his seminal work [Kaw95] (originally published in 1963). For the sake of comparison, we will briefly describe his categorization below, and contrast it with our types.

Figure 2 gives the schematizations which Kawakami includes in his characterization of phrase-final rises in Tokyo Japanese. Here, the names of the rise types are approximate translations from Kawakami's own: *futsû no*

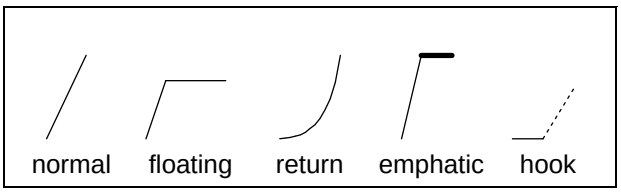

FIGURE 2. Schematic shapes of BPM types in Kawakami's (1963) categorization.

jôshô-chô, *ukiagari-chô*, *hanmon no jôshô-chô*, *tsuyome no jôshô-chô*, and *tsuriage-chô*, respectively.

Kawakami's categorization is based on differences in both surface form and meaning, as is ours, though his data are based on very acute yet impressionistic observations, while our data are from a series of quantitative perception and production experiments. The fact that both studies identify five BPM types is misleading: there is not a one-to-one mapping between categories in the two studies.

Kawakami's *floating* and *hook* rises do not occur in our inventory, while our *rise-fall* is not mentioned at all in his study. Kawakami states that the *floating* rise has shown itself in Tokyo Japanese only since the 1960s, and appears to have a limited distribution in terms of which words it can co-occur with. The *hook* rise is even more limited, in that it is restricted to the professional speech of female announcers, and occurs only with /desu/ and /-masu/ verbal endings. For these reasons, we do not consider these two types as central to our investigation. On the other hand, Kawakami does not identify the *rise-fall* pattern, which we include in our study. This BPM is admittedly a recent phenomenon in Tokyo Japanese, and is most often associated with the casual speech of young people. However, it is becoming increasingly widespread, and can often be observed in the speech of older men and women, even in professional settings. Therefore, we have decided to include it in the current study.

As for the other rise types, there is some overlap in the distinctions made by each study. Kawakami describes the *return* rise as very similar to our *incredQ* rise, in that it feels like a more scooping rise in which the steep part of the rise is late to begin, and it has associated with it a sense of "deep questioning". This appears to be the only rise that has a direct mapping between the two studies. Kawakami's *normal* and *emphatic* rises appear to overlap with our *infoQ*, *prom*, and *insist* categories. His *normal* rise is similar to *infoQ* in that it can cue simple clarification, though he also uses the same rise to describe sentence-medial bunsetsu-by-bunsetsu style reading, which would be more similar to our *prom* rise. His *emphatic* rise seems very similar to our *insist* rise, though many of his examples can also fit with our *prom* type as well.

It is difficult to make comparisons between the two studies based on only impressionistic and anecdotal observations of the form and function of each rise type. What is needed is a careful and controlled examination of these rises using perceptual judgments and production data from a number of native speakers. We now turn to such an investigation.

2.2 Perception stimuli

The stimuli used in the perception study were noun phrases consisting of a sequence of an unaccented modifier, accented/unaccented noun, and case particle, such as *chairo-no gorira-ni?* "to the brown gorilla?". There

were 5 initially-accented and 5 unaccented nouns, and 2 case particles (dative/locative *ni* and genitive *no*). Each phrase was uttered by a male native speaker of Tokyo Japanese (second author) using one of the 5 BPMs listed above, resulting in a total of 100 stimuli.

Stimuli were arranged in 5 blocks of 20 phrases; each block containing all nouns with both particles. BPMs were distributed across the blocks such that a given noun phrase did not have the same intonation pattern within a block, nor did its *ni/no* match have that intonation. Phrases in each block were in pseudo-random order, with at least 2 phrases separating phrases with the same BPM, and at least one phrase separating phrases with the same base noun.

2.3 Semantic scales

Ten native listeners participated in the perception study; all were unaware of the purpose of the experiment. Subjects were asked to rate each stimulus on a set of 8 'semantic scales' ([GGH+97], inter alia). The scales were designed in an attempt to tease apart subtle differences in meaning among the BPM types, and were based on our impressionistic observations about what the BPMs might mean, and how they might differ. Approximate English translations of the scales are shown below, along with abbreviations used in this paper as a shorthand to refer to each scale.[1]

The speaker ...

1: will continue talking. (continue)
2: is insisting the information. (insisting)
3: is simply confirming information. (confirming)
4: cannot believe the information. (disbelief)
5: is emphasizing the phrase itself. (emph–info)
6: is emph. the function of the phrase in the discourse. (emph-funct)
7: is explaining the information. (explaining)
8: sounds irritated. (irritated)

Each scale was written in Japanese kanji/kana orthography, along with a sequence of 5 dots representing degrees on that scale. Subjects were instructed to circle the appropriate dot indicating the degree to which they AGREE (1) or DISAGREE (5) with the particular characterization of the scale. A total of fifteen seconds was given for subjects to circle their response on all 8 scales, and a beep followed by a 1-second silent interval preceded each trial. A practice session of 4 trials was also included.

[1]The Japanese for each semantic scale reads: *washa-wa ...* 1: *hatsuwa-o tsuzukeru*, 2: *jibun-no hatsuwa naiyô-ni koshitsu shiteiru*, 3: *jôhô-o kikite-ni tanni kakunin shiteiru*, 4: *jôhô-ga shinjirarenai*, 5: *jibun-no tsutaeteiru jôhô jitai-o kyôchô shiteiru*, 6: *kaiwa-no naka-deno sono furêzu-no yakuwari-o kyôchô shiteiru*, 7: *jôhô-o setsumeishiteiru*, 8: *iradatteiru*.

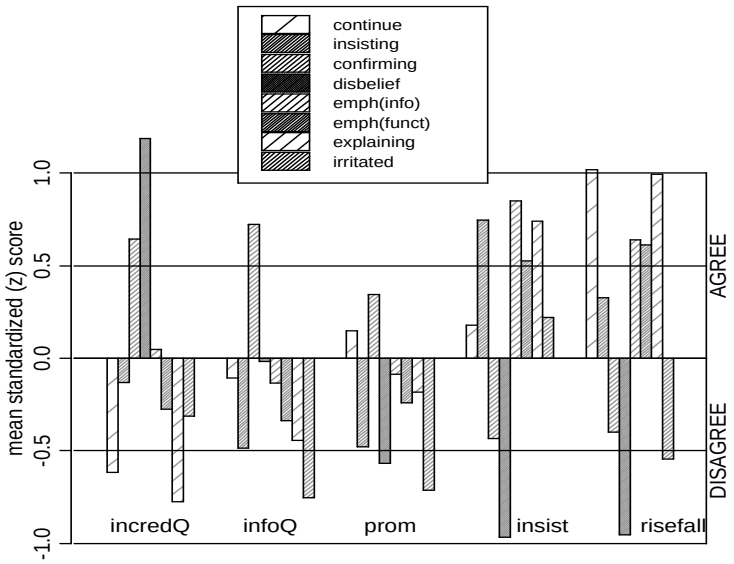

FIGURE 3. Mean standardized (z) scores for each of the BPM types, on 8 semantic scales.

2.4 Perception results

Figure 3 shows mean standardized scores for each of the BPM types, pooled across all subjects. One-way ANOVAs show main effects of BPM type for all 8 dependent variables. Post-hoc tests confirm the following major points: the *rise-fall* BPM was judged to cue a strong sense of continuation, while the *incredQ* was judged to end a turn. As expected, *insist* gave the strongest sense of insisting. Both *questions* were judged highest on the confirming scale, while the *insist* and *rise-fall* BPM did not have this sense. The *incredQ* was the only type to cue a sense of disbelief, as expected. The other types showed negative disbelief judgments, especially the *insist* and *rise-fall* BPM. These two types were also judged to be strongly emphatic (in both types of emphasis), in contrast to the negative judgments for the other types. In addition, *insist* and *rise-fall* were judged to be strongly explanatory, while the other types (especially *questions*) were not. And finally, only *insist* received a positive judgment for being irritating.

It is possible to generalize the perception results and identify three groups in which the BPMs are perceived similarly on certain scales:

QUESTIONS The two *question* BPMs pattern similarly on most scales: they are both the only types to be strongly perceived as seeking confirmation, while they are not perceived (negative judgments) as being insisting, emphatic, explanatory, or irritating. The two question types are distinct

mainly on the disbelief scale, which is consistent with the name given to the *incredulity question* BPM.

ASSERTIVES The *insist* and *rise-fall* contours were judged similar to each other in meaning, though their shapes are drastically different. They are both perceived as highly emphatic and explanatory, and receive negative judgments on the confirming and disbelief scales. While patterning similarly on these four scales, the two assertives are distinct from one another in some ways. Post-hoc tests indicate that *insist* is more insisting, while *rise-fall* is more likely to cue continuation, and sound far less irritating. It is also perceived as being marginally more explanatory, which leads us to refer to the *rise-fall* BPM also as *explan* in the remainder of this paper.

PROMINENCE-LENDING From the schematizations in Figure 1, *prom* and *insist* seem to be very similar in form. Results from the perception study indicate that the two BPM are in fact quite distinct perceptually. The *prom* BPM is far less insisting, and is not as emphatic or explanatory as *insist* is. It is also far less irritating. In addition, there is also similarity in judgments of the *prom* and *infoQ* BPMs. On most scales the two types were perceived similarly, though the *infoQ* BPM was found to have a strong sense of confirming, keeping in line with the fact that it cues a question. It is also important to note also that *prom* patterns with all but *rise-fall* in that it doesn't cue continuation, indicating that this is not a continuation rise, as one might like to interpret it based on English intonation.

We originally hypothesized that the function of *prom* is somehow emphatic; it lends prominence to the phrase whose edge it marks. This is also confirmed by Kawakami's description of his *emphatic* rise, in which he states that the rise on the phrase-final particle is not emphasizing the particle itself, but rather it makes the whole word to which it attaches prominent/emphasized ([Kaw95], p. 285). The two emphasis scales used in the current perception study were constructed to help elucidate this distinction. However, results show that *prom* was not judged emphatic on either scale, contrary to our expectations. This can be interpreted two ways: either *prom* is indeed not emphatic (maybe it serves more of a 'chunking' rather than emphatic function), or it is possible that the (rather obscure) wording of these semantic scales was not straightforward enough for the subjects to internalize. It is often easier for subjects to grasp the meaning of scales with simple, accessible wording, like "The speaker sounds irritated", while we suspect it may be more difficult to internalize and fully understand scales with convoluted wording, such as "The speaker is emphasizing the function of the phrase in the discourse". Therefore, further investigation is necessary before we can interpret this null effect as a negative effect.

The current perception test examines intonational meaning from several angles, both linguistic and paralinguistic. Results show that on a single

semantic scale the BPM types examined here may not be distinct from one another, though when considering all the attributes simultaneously (as listeners certainly do), the BPM types do in fact have distinct meanings.

3 Intonational form

3.1 Production data

A series of production studies were conducted in order to document the precise shape and timing characteristics of the five boundary pitch movements.[2] Four speakers (different from those involved in the perception study) participated in the production studies. Data from speaker KF will be emphasized here, though limited data from other speakers will be included as well.

Mini-discourses were constructed in which the target phrases consisted of a noun plus the following dative case particle *ni*. The noun was either the lexically-accented name *Na'oya* or the unaccented name *Manami*.[3] The different boundary rises on each target phrase were elicited by a context and the previous discourse turn (speaker Q), and were modeled when necessary using an audio prompt containing entirely different words. Utterances both with and without a pause following the target phrase were included in the experiments. In both pause conditions, the target mora was immediately followed by the mora /wa/ in the next word. The following is an example of the context and mini-discourse eliciting the *incredulity question* BPM.

Q-san-ga Yûko-no taisetsu-na yubiwa-o Naoya-ni machigaete watashite shimaimashita. "Q mistakenly gave Naoya Yuko's precious ring." *Q-san-wa sono koto-o A-san-ni oshiete imasu.* "She is telling this to A." *A-san-wa Q-san-ga sonna machigai-o shita koto-o shinjirarenai yôsu desu.* "A can't believe that Q made such a mistake."

Q: *Yûko-no yubiwa-o machigaete Naoya-ni watashichatta.*
"I mistakenly gave Yuko's ring to Naoya."

A: *É* <pause> **Naoya-ni???!!** <pause> *watashitatte???*
"Oh my ... To Naoya???!! You said you gave it to him???"

[2] A total of 4 production studies were conducted over a number of months. The experiments involved different speakers, and differed only slightly in design and analysis. Experiments 1 and 2 examined productions of two speakers (YO and KM), and are reported in [MV98]. Experiment 3 added an additional speaker (YO2) and more data from KM, and is reported in [VMvS98]. Experiment 4 added yet one more speaker (KF), and is the main focus of discussion in this paper. In general, the 4 speakers behaved remarkably similarly throughout the experiments.

[3] In the transcriptions, accented words contain an apostrophe after the vowel with which the accentual fall is associated; unaccented words lack such a marking.

In addition to the 5 BPM types listed above, two other types of rises were included in the production corpus for comparison:

- Unaccented word (*unacc*): An emphasized monomoraic unaccented word (e.g. *to'chi-no MI* "buckeye nut"). The accentual phrase-initial rise on *mi* "nut" is realized when the word is emphasized.
- Accented particle (*part*): Compound particle (e.g. *ni'-wa* (DAT-TOP)).[4]

The mini-discourses were presented in Japanese orthography, and target utterances were recorded in a sound-attenuated room. Sound files were digitized at 16 KHz (16 bit resolution) on SUN and SGI workstations, and analyzed using Entropic Research Laboratories *Waves+* software. Seven repetitions of each type were recorded by speaker KF (other speakers recorded more repetitions), and a few were later excluded from the analysis due to pitch-tracking errors or disfluencies. Speaker KF produced each contour at three speaking rates in sequence: first normal, then slow, then fast (other speakers produced only normal rate). Labels were placed by a trained phonetician at key points in the F0 contours and at segmental landmarks.

3.2 Production results

Contours

Figures 4 (pause following target) and 5 (no pause following target) show F0 contours of the target phrase. Raw data of all tokens of each BPM type produced by KF at a normal speech rate are shown. The left column shows contours with a preceding accented noun (*Na'oya-ni*), and the right column a preceding unaccented noun (*Manami-ni*). The contoured lines trace the F0 values of each frame from the start of the phrase to the end of the particle *ni* (pause condition) or the phonological low associated with the mora /wa/ after the fall (no-pause condition). The solid vertical line marks the onset of the target mora *ni* (all contours are time-aligned by this point), and the dashed horizontal line marks a fixed arbitrary reference F0 height. F0 contours for other speakers are given in [MV98] and [VMvS98]. Although here we show contours for both accented and unaccented preceding nouns, this paper discusses only the accented condition.

Accented phrases display an initial sharp rise, then their signature fall due to the accent in the noun *Na'oya*. The F0 reaches a low point, then begins to rise again at the phrase edge for the BPM. In the *rise-fall* (aka.

[4] In addition to these types, Experiment 1 (YO) included an accented word *ni'wa* "two birds", which was shown to behave similarly to the accented particle. Experiment 3 (YO2 and KM) included a low-ending contour without a boundary pitch movement, which showed the canonical accentual phrase shape. These contours are not discussed in detail here.

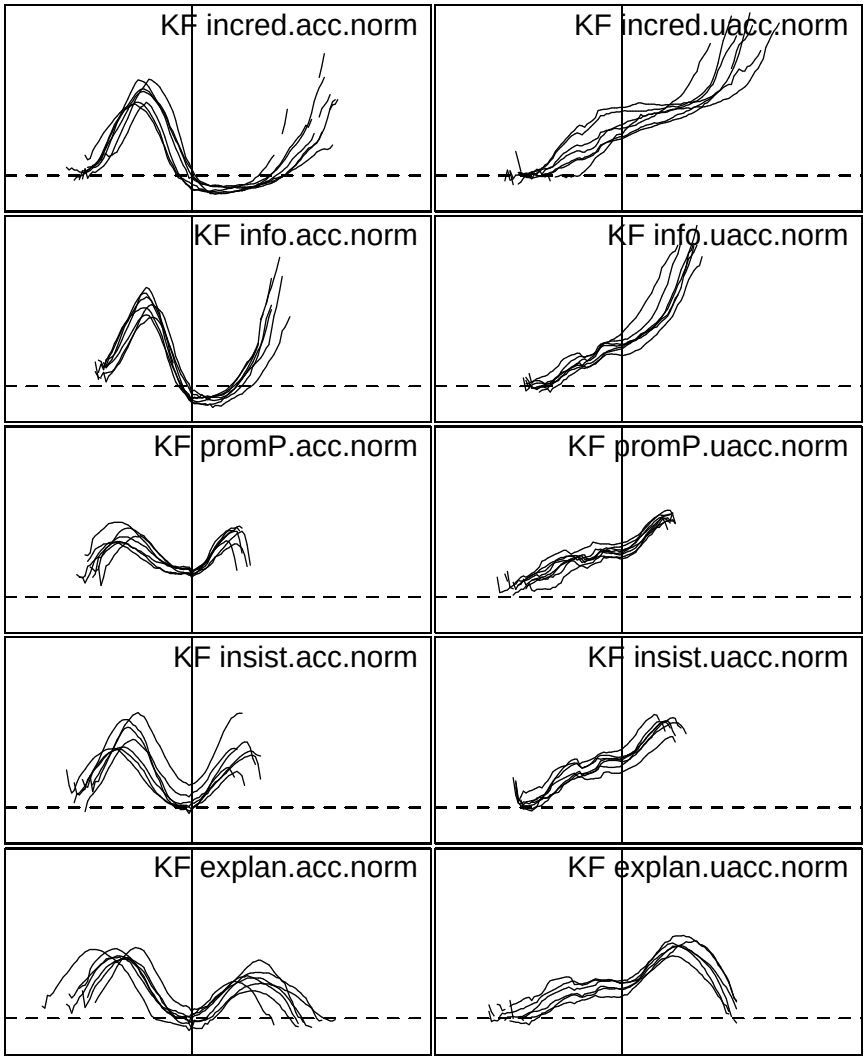

FIGURE 4. F0 contours of the target phrase for speaker KF. Normal rate, pause condition. The left column shows contours with accented nouns, and the right column shows unaccented nouns. Rows show the 5 BPM types.

explan) BPM, the contour falls yet again. This also resembles the shape of the other BPMs when not followed by a pause (Figure 5).[5] The Figures

[5]There is a slight downward movement of the contour at the end of some of the rises in the *prom* and *insist* types (pause condition, Figure 4, left panel). This is an artifact of the pitch tracking; the final portion of the vowel in these types is produced in a pressed voice, and cannot be tracked properly. There is no

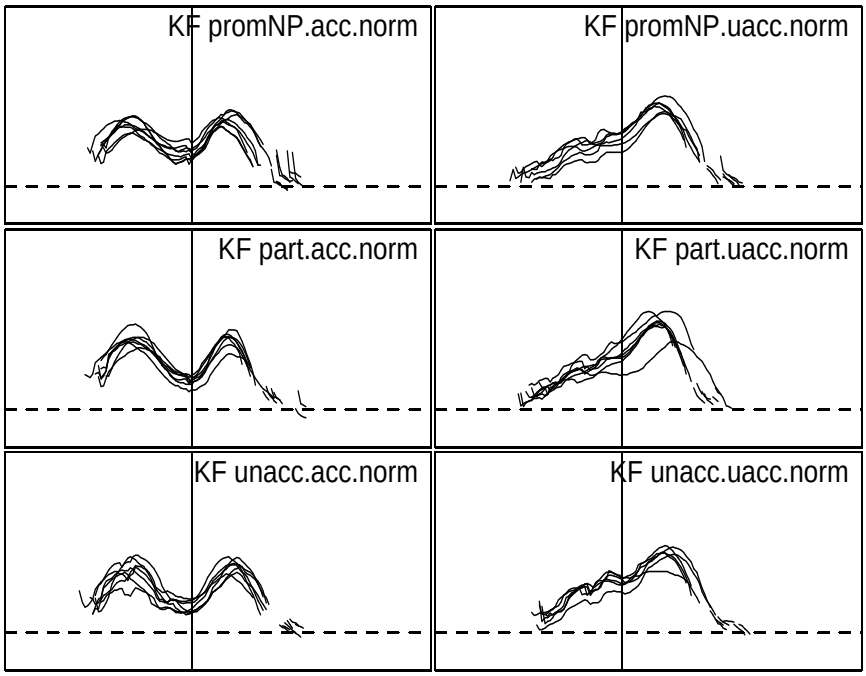

FIGURE 5. F0 contours for speaker KF, no-pause condition. Left and right columns show accented and unaccented nouns, respectively.

show that the F0 heights of the low/high points, the overall shape of the rise, as well as the time-course of the rises differ between BPM types. In the following, we will give a detailed analysis of these differences.

F0 heights

Figure 6 shows mean F0 values of the accented peak in the noun, the low valley, the peak at the end of the rise, and the subsequent fall to low (only in the *explan* BPM). Data shown in the Figure are taken from the productions of *Na'oya-ni* followed by a pause, for all speakers.[6] The no-pause condition patterns similarly. The Figure shows some degree of variability across speakers, though a few important generalizations can be made.

The peak height of the preceding noun is highest in the *incredQ* BPM, while it is generally lowest for *prom*. This suggests differences in overall pitch range.[7] The height of the low valley also differs among types: the

perceptible low or falling tone in these cases.

[6]Data from *Na'oya-no* are also included for speakers KM and YO2.

[7]Differences in pitch range have been shown to provide systematic cues to

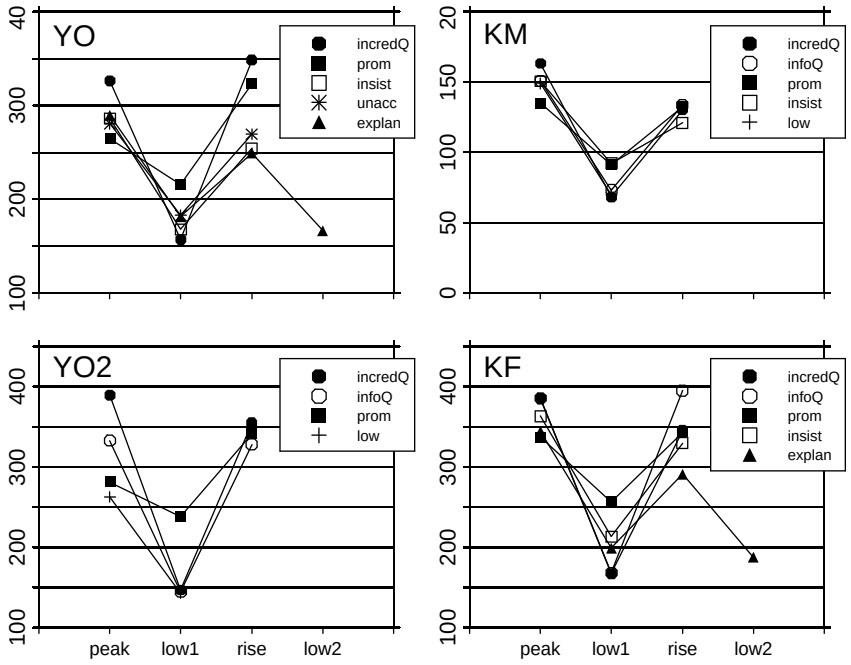

FIGURE 6. Mean F0 values (Hz) of preceding accented noun peak, low valley, peak of rise, and optional fall to low. All four speakers, pause condition.

intervening low in *prom* is consistently higher than the other types, and the two *question* types have the lowest F0 value, which is in fact not different from the final F0 value in the *low* (no rising BPM) type. This difference in the height of the low valley is consistent with an expansion of range in *questions* and a compression for *prom*. However, the height of the final rise itself does not support this hypothesis: the *question* and *prom* types pattern together in that they rise higher than the other BPMs. While the F0 heights in *questions* can be accounted for by an expansion of range, there probably is not an overall compression for *prom*. We suspect that that the mid-range values in the noun accent peak and intervening low in the *prom* type are due to an intentional compression in order to enhance the prominence of the high rise at the phrase edge. This would mean a local compression then expansion or some other kind of boost at the edge. However, more research is necessary (particularly perception studies using synthesis/re-synthesis to vary the F0 heights) in order to fully understand the F0 height characteristics of the *prominence-lending* rise.

---

question interpretation in Korean [JO96] and incredulity vs. uncertainty readings of the L*+H L- H% contour in English [WH88, HW92].

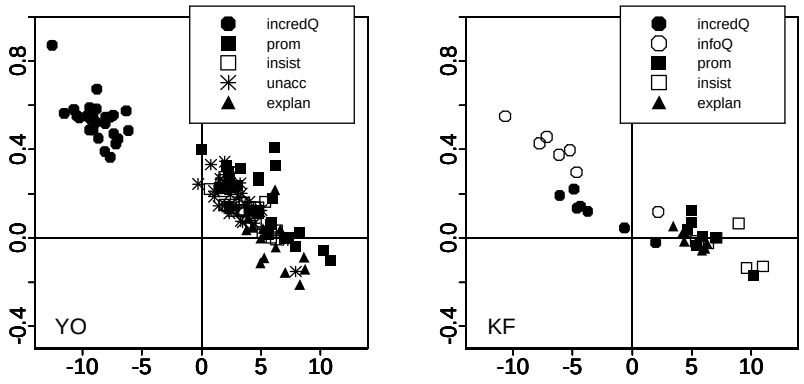

FIGURE 7. Linear and quadratic components of curves fit to rise portion. Pause condition, accented noun, speakers YO and KF.

Rise shapes

We now turn to a quantitative analysis of the overall shape of the BPM rises. A brief examination of the raw F0 contours in the accented targets in Figures 4 and 5 shows that some BPM types appear more curved (concave) than others. In order to quantify this curvature, it is possible to fit a quadratic function to the rise portion of the contours (from the onset of the target particle *ni* to the end of the rise). Figure 7 shows scatter plots of the linear and quadratic components of such a fit, for speakers YO and KF. The value of the quadratic component gives an idea of the curvature of the rise (positive=concave, negative=convex, zero=linear). This method of quantifying shape shows that the *questions* pattern apart from the other types: the shape of both *incredQ* and *infoQ* are more concave. Points for the other types lie near (or slightly above for YO) the zero line of the quadratic component, indicating that these shapes can be well-approximated by a linear function. In the absence of a significant quadratic component, the value of the linear component indicates the steepness of the linear rise.

Duration and timing

Figure 8 shows the durations of the segments in the target mora *ni* for all speakers, and the time-course of the rise with respect to these durations. The left panel (pause condition) shows the durations of /n/ and /i/, and the right panel (no-pause condition) also shows the duration of the following /wa/. Vertical bars mark mean locations of rise start ('elbow') and rise peak, respectively.

The panels in the Figure show that while the onset (/n/) duration stays relatively consonant across BPM types within each speaker, it is the vowel

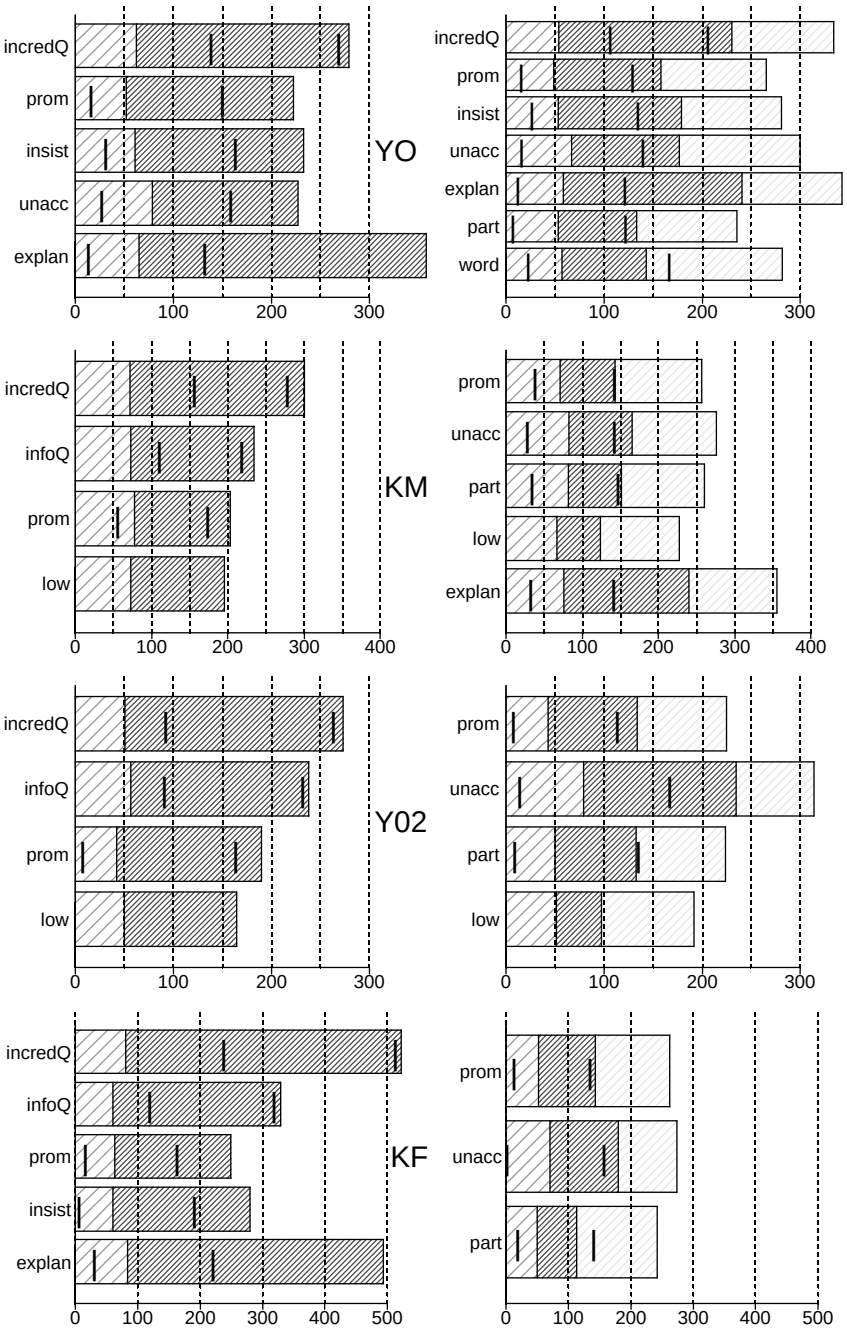

FIGURE 8. Mean durations (ms) of segments in target particle *ni* in the accented noun phrase *Na'oya-ni* "to Naoya", for all speakers. The left panel (pause condition) shows the mean durations of /n/ and /i/, and the right panel (no-pause condition) also shows the mean duration of the following /wa/. Vertical bars mark mean locations of rise start ('elbow') and rise peak, respectively.

length that significantly differs among the types. The vowel in both *question* types are lengthened, with *incredQ* being lengthened the most. In addition, the vowel is lengthened to the same or greater degree in the *explan* BPM.

The Figure also shows how the start and end of the BPM rise (the two vertical bars) align with the segments. In the *question* BPMs, the rise (first bar) starts well within the vowel itself, while in the other types it starts soon into the onset. The end of the rise (second bar) occurs near the end of the vowel in nearly all BPM types in both pause conditions, though there is a fair amount of variability in this location. The one notable exception to this is the *explan* BPM, in which the end of the rise occurs well within the lengthened vowel.[8]

Though the timing of the rise start and end differ systematically across the BPM types, there is one striking observation: the duration of the rise itself (the duration between the bars) remains constant throughout. That is, with only a few exceptions, there is no significant effect of BPM type on the duration of the rise portion (within each speaker). Table 1 summarizes the results of a series of one-way ANOVAs and post-hoc tests examining the effect of BPM type on rise duration for all speakers (normal rate).

speaker	pause condition	no-pause condition
YO	n.s.	*word* > all other types
KM	n.s.	n.s.
YO2	n.s.	*unacc* > *prom* only
KF	*incredQ* > *prom* and *insist* only	n.s.

TABLE 1.1. Summary of one-way ANOVA and post-hoc test results on rise duration differences across BPM types. Significance is measured at the 0.01 level.

This apparent invariance in rise duration within a given speaker gives the impression that the rise is somehow 'anchored' by its onset at the beginning of the target mora in non-questions, and by its offset at the end of the mora in questions. However, while the statistical tests show no significant difference in rise duration in most comparisons, there is still more to be told.

Figure 9 gives a plot of speaker KF's mean rise durations in each BPM type, in three different speech rates. There are two important observations here. First, within a given speech rate the mean rise duration is fairly con-

[8]It is common for the rise to end before cessation of voicing in the pause cases, especially in the *prom*, *insist* and *unacc* BPMs. This is because of pitch tracking problems due to the presence of non-modal voice quality at the end of such rises. The final portion of the vowel is uttered with pressed voice, giving the percept that the contour is abruptly cut-off. This is not the case with the *question* types, in which the rise is tracked all the way to the very end of the vowel.

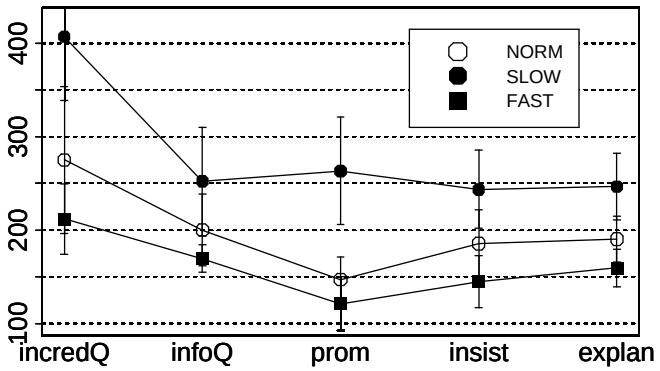

FIGURE 9. Mean rise durations (and standard deviation bars) across BPM types in 3 rates. Speaker KF.

stant across BPM types, though it differs consistently across speech rates, as expected.[9] Second, there is a large amount of variability in the means, as shown by the standard deviation bars surrounding each point. This variability is not adequately represented by the mean rise durations in normal speech rate plotted in Figure 8 and by the accompanying summary of the statistical tests in Table 1. The tests treat variability within each BPM type as noise in their comparison of the mean rise durations. However, this variability could be both meaningful and predictable. It is expected that rise durations should vary in different speech rates, due to the large variability in the segment durations themselves. In the same way, it is probable that the segment durations differ to some degree within a given speech rate as well. If the rise duration is somehow dependent on the duration of the segments to which the BPM attaches, its variability could be explained by the variability in the segment durations.

Figure 10 shows the correlation between the observed rise durations and those predicted by considering the segmental durations. Only the *incredQ* type from speaker KF (all speech rates pooled) is given here. The correlation of 0.914 indicates that the variability in rise durations (both within and across speech rates) can be systematically and reliably predicted by a model which considers weighted sums of the onset and vowel durations of the target particle.

The rise in each BPM type has a different relationship to the segments, as shown by different weights for each linear regression model, given in Table 2. However, it is the case that the weights of *questions* pattern differently from the other types: the vowel duration contributes more in the *questions*

[9]Note here that KF is the only speaker who shows longer rise durations for the *incredQ* BPM, with other types not significantly different from each other.

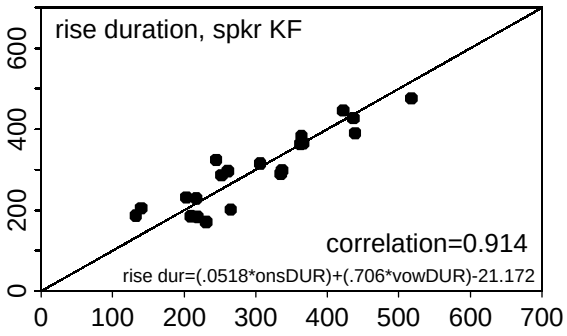

FIGURE 10. Observed and predicted rise duration values (ms) for the *incredQ* BPM, using linear regression. Speaker KF, all speech rates pooled.

and the onset duration contributes more in the other BPMs.

Figure 10 and Table 2 show that the variability in rise durations is not necessarily 'noise', but can be reliably predicted from segmental durations. That is, the timing of the F0 contour is intimately tied to the durations of the segments to which it is attached. Likewise, not only the rise duration, but any point on the curve, such as the start of the rise, or the midpoint of the rise, can also be reliably predicted considering segmental durations (given a model of how the durations are weighted for each BPM type). Taking this even further, it is desirable to predict *all* points on the curve, for implementation in a speech synthesis system. In the next section, we describe possibilities for such a model.

BPM type	corr.	onset weight	vowel weight	intercept
incredQ	.914	.051	.706	-21.172
infoQ	.818	.152	.480	54.270
prom	.958	1.012	.715	-47.633
insist	.942	1.454	.606	-36.566
explan	.884	.916	.174	44.753

TABLE 1.2. Correlations and linear regression coefficients for models of rise duration in each BPM type. Speaker KF, all speech rates pooled.

4 Model of F0 contour alignment

The production data of the current study indicate that not only F0 height and shape, but also the segment durations and (crucially) the timing of the

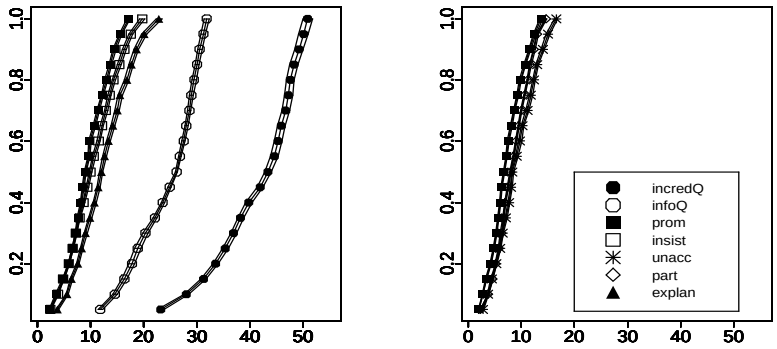

FIGURE 11. Time-course of F0 rises. Mean time values at each F0 anchor point. KF, normal speech rate, particle *ni*. Frames are at 10ms intervals. Standard error bars.

F0 contour with respect to these segments is important for distinguishing the different BPM types.

To implement this in a speech synthesis system, one option would be to use this data to develop a model of alignment of key F0 events (such as the rise start and end) with respect to the segments, and propose a linear interpolation between the points [PB88]. The problem with this is that the rises are not simply straight lines, but curves of various shapes, which should not *a priori* be discarded as unimportant or artifacts of some smoothing mechanism. The curve shapes are shown in Figure 11, which plots the normalized averaged F0 contours for KF's different BPM types. The Figure shows that the *incredQ*, *infoQ* are both concave, while the other types are more convex, or even linear. This is essentially what Figure 7 showed. Another approach to synthesizing these BPMs would be to use the data given here to determine the start time (and amplitude) of a single rise shape which begins at that point [FH93, FOO+94]. Again, the problem is that the shapes of these curves are not identical in their concavity/convexity.

We suggest an alternate approach to modeling these boundary pitch movements which involves modeling the *entire* shape of the rise, as it occurs in the data. As mentioned above, we do not see a need to *a priori* disregard the details of the rise shape, but rather we choose to quantitatively describe the whole shape, in terms of how it aligns with the segments. The approach is described briefly below (but see [vSM97, vSMVS98] for a more detailed discussion).

4.1 Quantization procedure

We take the onset of the case particle *ni* as the starting point of the BPM, (though it is possible that other points may prove more useful). The F0

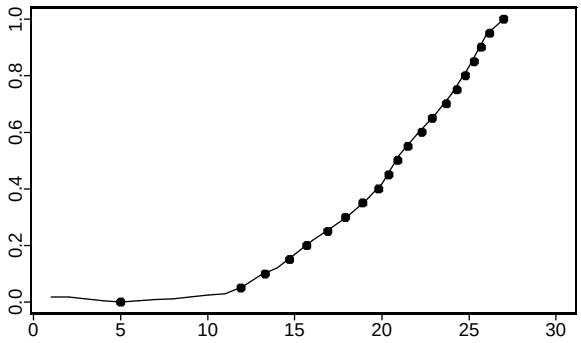

FIGURE 12. Sample *incredQ* contour with dots marking F0 anchor points. Frames are at 10ms intervals.

contour between the onset and the end of the rise is first normalized with respect to the pitch range, so that the F0 values now range from 0 to 1. It is then sampled at locations corresponding to a range of percentages between 0% and 100% of the maximum height. Thus, the 100% point is the peak location, and the 50% point is the time point where the F0 curve is half of the maximum height. We call these time points *anchor points*. Figure 12 gives an example of the sampling (at 5% intervals).[10]

4.2 Alignment parameter matrix

As mentioned in Section 3.2, we propose that each anchor point on a given BPM can be described as a function of the onset and vowel durations. That is, the location of a given anchor point depends linearly on these two durations: all else equal, if either the onset or vowel of a particular BPM type is lengthened, the location of that anchor point will be shifted to the right to some degree. The relationship between the anchor point and the onset/vowel durations can be determined using a regression analysis, in which the regression coefficients serve to weight the contribution of the onset/vowel lengths. Using this method, a regression analysis is performed on each of the n anchor points for a given BPM type, and the ensemble of resulting regression weights forms what we call an *alignment parameter matrix (APM)*. The APM as a whole characterizes how the F0 contour is

[10]In previous analyses of English using such quantization ([vSM97]), the procedure consisted of subtracting a non-horizontal phrase curve from the observed curve, and then normalizing the resulting difference curve by dividing by the peak value. In the current data, it is not clear how to determine the local shape of this hypothetical phrase curve, so that we opted for this admittedly cruder method of obtaining anchor points.

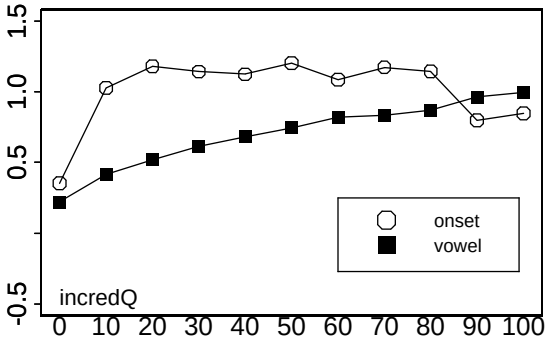

FIGURE 13. Regression weights for each anchor point in KF's *incredQ* BPM type.

aligned with the segments, for a given BPM type. Contours with systematically different alignment properties, such as the *incredQ*, *infoQ*, and *prom* rises described above, will necessarily have different alignment parameters. Therefore, the important fact that these BPMs are phonologically distinct (different heights, shapes and alignment characteristics lead to different perceived meanings) is preserved in this quantitative model of F0 contour alignment.

Figure 13 shows a graphic representation of the alignment parameter matrix for KF's *incredQ* BPM type. The plot characters show the weights (regression coefficients) of a linear model applied to each anchor point in the curve. The Figure shows that the onset duration is more important than the vowel duration in predicting anchor points early on in the BPM, though the contribution of the vowel increases further into the BPM.

This approach has been successfully applied to the modeling of English nuclear accent peaks and boundary rises by van Santen and Möbius [vSM97]. They showed that the alignment of the peak and other F0 points of an accent curve depend systematically on the durations of the onset, rhyme, and remaining portion of the accent group. At present, the application of this model to Japanese BPMs is still underway. The results shown above are encouraging in that the time-course of the rise in the BPMs is clearly dependent in some way on the durations of the segments to which it is associated, and this relationship can be quantified using linear regression models. More research is necessary to explore how far this approach can be taken, and to compare differences among the APMs for all BPM types.

5 Summary

This paper provides a detailed analysis of the intonational form and function of five boundary pitch movements in Tokyo Japanese. The perception study describes the linguistic and paralinguistic dimensions on which meanings of the BPMs are distinguished. The production study details how the F0 heights, durations, and (crucially) alignment of the F0 contour with the segments are all used to define the movement types. We suggest a quantitative model of F0 contour alignment which uses observed data to model the entire shape of the curves.

6 References

[FH93] Hiroya Fujisaki and Keikichi Hirose. Analysis and perception of intonation expressing paralinguistic information in spoken Japanese. In David House and Paul Touati, editors, *Proceedings of the ESCA Workshop on Prosody*, volume 41 of *University of Lund Working Papers in Linguistics*, pages 254–257, 1993.

[FOO+94] Hiroya Fujisaki, Sumio Ohno, Masafumi Osame, Mayumi Sakata, and Keikichi Hirose. Prosodic characteristics of a spoken dialogue for information query. In *International Conference on Spoken Language Processing (ICSLP)*, pages 1103–1106, 1994.

[FS71] Hiroya Fujisaki and H. Sudo. Synthesis by rule of prosodic features of connected Japanese. In *International Congress on Acoustics*, pages 133–136, 1971.

[GGH+97] Esther Grabe, Carlos Gussenhoven, Judith Haan, Erwin Marsi, and Brechtje Post. Preaccentual pitch and speaker attitude in Dutch. *Language and Speech*, 41(1):63–85, 1997.

[HW92] Julia Hirschberg and Gregory 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(2):241–251, 1992.

[JO96] Sun-Ah Jun and Mira Oh. A prosodic analysis of three types of Wh-phrases in Korean. *Language and Speech*, 39:37–61, 1996.

[Kaw95] Shin Kawakami. On phrase-final rising tones. In *A Collection of Papers on Japanese Accent*, pages 274–298. Kyûko Shoin, Tokyo, 1995. [Originally published in 1963] (in Jpns).

[MH94] Toshiko Muranaka and Noriyo Hara. Features of prominent particles in Japanese discourse: Frequency, functions, and acoustic features. In *International Conference on Spoken Language Processing (ICSLP)*, pages 395–398, Yokohama, Japan, 1994.

[MV98] Kazuaki Maeda and Jennifer J. Venditti. Phonetic investigation of boundary pitch movements in Japanese. In *International Conference on Spoken Language Processing (ICSLP)*, pages 631–634, Sydney, Australia, 1998.

[PB88] Janet B. Pierrehumbert and Mary E. Beckman. *Japanese Tone Structure*. MIT Press, 1988.

[Ven95] Jennifer J. Venditti. Japanese ToBI labelling guidelines. [http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_homepage.html], 1995.

[VMvS98] Jennifer J. Venditti, Kazuaki Maeda, and Jan P. H. van Santen. Modeling Japanese boundary pitch movements for speech synthesis. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 317–322, Jenolan Caves, Australia, 1998.

[vSM97] Jan P. H. van Santen and Bernd Möbius. A model of fundamental frequency contour alignment. In *Proceedings of the ESCA Workshop on Intonation*, pages 321–324, Athens, Greece, 1997.

[vSMVS98] Jan P. H. van Santen, Bernd Möbius, Jennifer J. Venditti, and Chilin Shih. Description of the Bell Labs intonation system. In *Proceedings of the 3rd ESCA Workshop on Speech Synthesis*, pages 293–298, Jenolan Caves, Australia, 1998.

[WH88] Gregory Ward and Julia Hirschberg. Intonation and propositional attitude: The pragmatics of L*+H L H%. In *Proceedings of the Eastern States Conference on Linguistics (ESCOL)*, pages 512–522, 1988.