Perceptually-related Acoustic-Prosodic Features of Phrase Finals in Spontaneous Speech

Carlos Toshinori Ishi, Parham Mokhtari & Nick Campbell

JST/CREST at ATR/Human Information Science Labs

[carlos,parham,nick]@atr.co.jp

Abstract

With the aim of automatically categorizing phrase final tones, investigations are conducted on the relationship between acoustic-prosodic parameters and perceptual tone categories. Three types of acoustic parameters are proposed: one related to pitch movement within the phrase final, one related to pitch reset prior to the phrase final, and one related to the length of the phrase final. A classification tree is used to evaluate automatic categorization of phrase final tone types, resulting in 76% correct classification for the best combination among the proposed acoustic parameters. Experiments are also conducted to verify the perceived degree of pitch change within a phrase final, and the perceived degree of pitch reset. While a good relationship is found between the perceptual scores and some of the acoustic parameters, our results also advocate a continuous rather than a categorical relationship between some of the phrase final tone-types considered.

1. Introduction

Phrase finals in Japanese utterances convey both linguistic and paralinguistic information. For example, they convey grammatical information such as modality (declarative vs. interrogative), focus, punctuation of phrase boundaries, and continuity of the sentence. They also convey important paralinguistic information such as the manner and attitude of the speaker.

In linguistics and phonetics research, there have been many proposals for categorizing sentence final intonation [1,2,3]. However, such methods are usually based on auditory perception and rarely extend to an automatic categorization of intonation types.

Phrase finals usually have greater prosodic variability in spontaneous speech than in read speech. While the X-JToBI [4] labeling method was proposed in order to more adequately describe such variability, automatic labeling is still not possible.

A goal of the present research is therefore automatic prosodic labeling of phrase finals in a large database of spontaneous, expressive speech collected in the JST/CREST ESP Project [5]. In this paper we focus on the description and automatic classification of phrase final prosody, specifically by analyzing the relationship between tone categories perceived by humans and acoustic-prosodic features extracted from the speech signal.

2. Analysis unit and definition of phrase finals

Our speech database consists of natural daily conversations recorded as part of the JST/CREST ESP Project. We used the prosodic phrase as the utterance unit for analysis. The prosodic phrases were segmented semi-automatically, boundaries being placed at evident pauses or pitch resets. For analysis, we used 404 phrases taken from three natural conversations with family members and with business people.

In this paper, ***phrase final*** is defined as the V (vowel) portion, or the VN (vowel + syllable-final nasal) portion of the last syllable of the phrase, i.e., the last syllable excluding the initial consonant. This definition is compatible with the perceptual rhythmic beat position (Perceptual Center, or P-Center) which is considered to be close to the vowel onset [6]. These rhythm properties have also been reported for Japanese speech [7,8].

The segmentation of the phrase finals was realized semi-automatically, using power and periodicity properties of the speech signal.

3. Categorization and labeling of phrase finals

Table 1 shows relations between the categorizations of sentence final particle tone types proposed by [3], and boundary pitch movement (BPM) labels proposed in the X-JToBI framework [4]. In the examples, ⌉ indicates a pitch fall, ⌈ indicates a pitch reset, and ⌒ indicates a rising pitch movement.

Table 1. Categorization of phrase final tones and corresponding phrase boundary pitch movement (BPM) labels.

Tone Type[3]	Perceptual Properties	Example	X-JToBI BPM [4]
1a	Low	na⌉i ne	L%
1b	Low + Falling tone	na⌉i ne⌉e	L%
2a	High	na⌉i⌈ne	L%+H%
2b	High + Lengthened	na⌉i⌈nee	L%+H%>
2c	Low + Rising tone	na⌉i ne⌒	L%+LH%
3	High + Falling tone	na⌉i⌈ne⌉e	L%+HL%
5	High + Fall-Rise tone	na⌉i⌈ne⌉e⌒	L%+HLH%

The categories shown in Table 1 are perceived as distinct tone-types that convey distinct paralinguistic functions in Japanese [1,2], but in this paper we focus on the problem of tone-type categorization, disregarding their functional properties.

In particular, we adopt the tone category labels proposed by [3] (cf. Table 1). In addition to this basic set of labels, we found it necessary to include a tag "*E*" (Extended) after the tone category label to mark situations when the phrase final is very lengthened; the functional properties of such phrase final lengthening are also pointed out in [1]. Furthermore, a tag "*S*" (Short) was added for the *2c* category, when the rising curvilinearity perceived in the pitch movement within the phrase final is particularly short and fast; the short versus long distinction within the *2c* category is also expressed with

different symbols in [3], albeit along a continuum.

Although most previous research related to intonation is based only on F0 information without considering phonation types, it is well known that non-modal phonation (e.g., creaky, harsh and whispery) appears quite frequently in natural speech. Furthermore, F0 measurements are known to be less reliable especially in these segments containing non-modal types of phonation. We therefore took special care in F0 estimation (Section 4.1), and decided to also annotate phonation types in the present research. The following set of labels is proposed:

- Phrase final phonation type: Modal (*M*), Rough (*R*), Creaky (*C*), Aspirated (*A*) (aspiration when speaking while laughing), Devoiced/Deleted (*D*).

Two native speakers of Japanese with experience in prosodic labeling annotated the 404 phrase finals according to the categories of tone and phonation types described above. These labels were then checked by three other native speakers in order to resolve any disagreements.

4. Acoustic-prosodic features

4.1. F0 estimation

In this section, we focus on the problems of voicing decisions for discriminating between voiced and unvoiced portions of the speech signal, and on selection of F0 values relevant to pitch perception.

For F0 estimation, we used a method based on the auto-correlation function. Specifically, the residual signal obtained from the LPC inverse filter of the speech signal is low-pass filtered before calculating the auto-correlation function (*Rxx*). The peaks in the autocorrelation function are detected and treated as candidates for F0.

The autocorrelation function is usually normalized as *Rxx(i)/Rxx(0)*, and a threshold is determined for the voicing decision. However, as *Rxx(i)* is calculated as a summation of $N - i$ multiplications and *Rxx(0)* is calculated as a summation of N multiplications, the more i increases, the smaller the *Rxx(i)/Rxx(0)* value. Thus, it is not appropriate to define a fixed threshold for all F0 candidates. Here, we normalized *Rxx* according to the following expression:

$$\frac{N}{N-i}\frac{Rxx(i)}{Rxx(0)} \qquad (1)$$

This normalizes the effects of reduction of *Rxx(i)* as i increases, leading to a more suitable decision.

The following steps were proposed for post-processing of F0 in order to remove unreliable values:

- Removal of points where the normalized autocorrelation coefficients are smaller than a threshold value.
- Removal of isolated points.
- Removal of the points where the power decreases more than 6 dB in an interval of 50 ms, taking into account the effects of perceptual masking [9].

These constraints also have the effect of reducing microprosodic variations, and therefore lead to perceptually more relevant F0 values. Also in this context, F0 values are converted to a musical scale (semitone intervals) prior to further processing.

4.2. Segment representative F0 parameters

Based on our previous work [10] regarding perceptually-related representative F0 values for syllable units, the following parameters are proposed:

- F0 average: average F0 value of the first (*F0avg2a*) and second (*F0avg2b*) halves of the phrase final; average of the final portion of the syllable immediately preceding the phrase final (*F0avg_p*). *F0avg_p* was estimated from four reliable F0 values obtained by back-tracking and searching from the phrase final start point.
- F0 target: target F0 value of the first (*F0tgt2a*) and second (*F0tgt2b*) halves of the phrase final. The target value of a segment is defined as the extrapolated value at the end of the segment, of a first order regression analysis of F0 values within the segment, as proposed in [10].
- F0 minimum (*F0min*) and F0 maximum (*F0max*).

4.3. F0 movement parameters

In order to quantify the pitch movement within the phrase final, we define parameters related to F0 slope (yielding information on how fast pitch changes) and range of F0 movement (yielding information on how much the pitch changes).

- *F0slope*: the slope obtained from the reliable F0 values within the phrase final, by first order regression analysis (*F0slope1*); slopes obtained from each of the segments after splitting the phrase final into two parts corresponding to first half and second half (*F0slope2a* and *F0slope2b*). Slopes were computed only when 3 or more (non-zero) F0 values were detected in the segment.
- *F0move*: the range of F0 movement within the phrase final, computed as the difference of representative F0 values between the second and first parts of the phrase final.

  F0avg2b – F0avg2a *F0tgt2b – F0tgt2a*
  F0tgt2b – F0avg2a *F0tgt2b – F0max* (for falling tones)
  F0tgt2b – F0min (for rising tones)

  Here, falling tones and rising tones were automatically identified using *F0slope2b* values.

Another important factor in categorizing the phrase final tones is the presence or absence of pitch reset between the phrase final and the preceding syllable. Taking this into account, we also define the following parameter for F0 reset.

- *F0reset*: the range of F0 reset between the phrase final and the syllable immediately preceding the phrase final, computed as the difference of representative F0 values between the first part of the phrase final and the last portion of the syllable previous to the phrase final.

  F0avg2a – F0avg_p *F0tgt2a – F0avg_p*
  F0max – F0avg_p

5. Relationship between acoustic parameters and perceptual tone categories

5.1. Qualitative analysis

In 44 of the 404 phrase finals, no reliable F0 values could be obtained, so for these samples, the calculation of the parameters was not possible. All the subsequent results were therefore limited to the 360 phrase finals where reliable F0 values could be obtained.

Fig. 1 shows the distributions of four of the acoustic parameters representing respectively F0 slope (*F0slope2b*), movement (*F0tgt2b-F0avg2a*), reset (*F0avg2a-F0avg_prev*), and segment duration (*Dur*), which were found by visual inspection to yield the best separation of the tone categories. In all four panels, the vertical dashed bars separate the samples of each category, and the horizontal dashed lines indicate thresholds that may be used to discriminate the tone categories of the phrase finals.

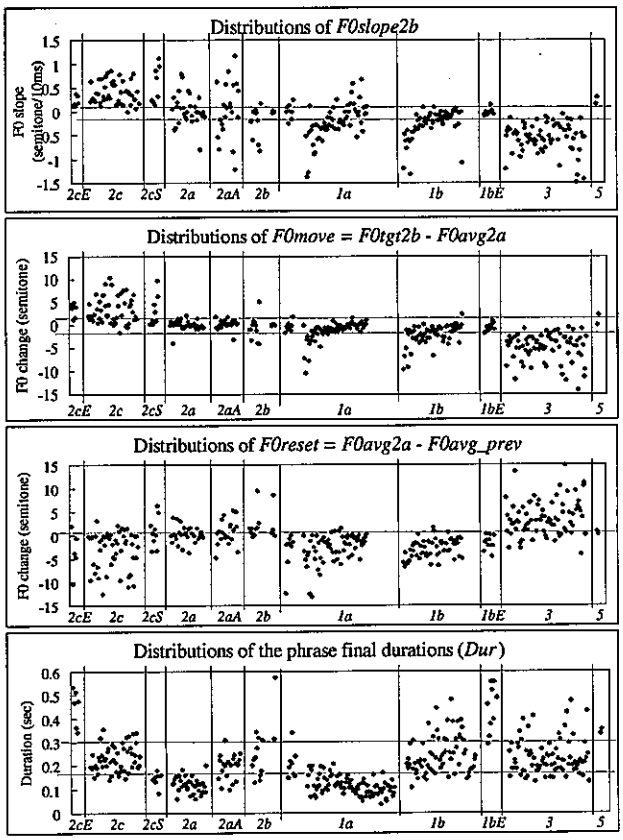

Figure 1. Distributions of several acoustic parameters per each category: (a) *F0slope2b*, (b) *F0move = F0tgt2b – F0avg2a*, (c) *F0reset = F0avg2a – F0avg_p*, (d) *Duration*.

A comparison of the top two panels of Fig. 1 shows similar tendencies for the distributions of *F0slope2b* and *F0move*. However, *F0slope2b* exhibits a relatively larger spread for categories *1a* and *2a*, perhaps owing to the fact that the number of reliable F0 values in some of these short phrase finals may not have been sufficient to reliably estimate slope values. Another possible reason is that perceiving a pitch change is more difficult in shorter segments, even if there is an acoustic F0 movement within the segment [11]. The tighter distributions for *F0move* shown in Fig. 1(b) suggest thresholds which approximately distinguish the following three groups of categories: {*3*}, {*1a, 2a, 2b*} and {*2c*}.

Similarly, a threshold for the *F0reset* parameter could be used to distinguish {*3, 2b, 2a*} from the other tone types. However, while the threshold shown in Fig. 1(c) seems reasonable for categories *3* and *2b*, discrimination of category *2a* is less clear. This may imply that perceptual identification of category *2a* could be influenced by acoustic features other than *F0reset* alone, such as the change in power. More investigation is needed to clarify how human perception is used to identify this category.

Finally, the thresholds of duration shown in Fig. 1(d) can be used to distinguish between length-distinctive categories {*1a, 2a, 2cS*}, {*1b, 2b, 2c, 3*} and {*1bE, 2cE, 5*}. Note also that aspirated phrase finals tend to be longer than their non-aspirated counterparts, as indicated in Fig. 1(d) by the distributions for the *2a* and *2aA* categories. A duration measurement disregarding the aspiration portion would shift the distribution of category *2aA* toward shorter durations falling below the threshold.

5.2. Automatic classification of phrase finals

As a first step towards automatic identification of the prosodic category of phrase finals, we applied a statistical classification-tree algorithm to various combinations of the acoustic parameters described above. In particular, the task of the classification tree was to discriminate the 11 tone categories listed along the abscissae of Fig. 1. Of the four acoustic parameters whose distributions were shown in that figure, the highest single classification accuracy (49.7%) was obtained using *F0move = F0tgt2b – F0avg2a*. Successively recruiting the remaining parameters in order of increasing overall accuracy, the best combination of two parameters yielded 58.8% (*F0move & F0reset = F0avg2a – F0avg_p*), the best combination of three parameters yielded 70.0% (*F0move & F0reset & Dur*), and the best combination of four parameters yielded 75.9% (*F0move & F0reset & Dur & F0slope2b*).

Table 2. Matrix of confusions between perceived tone types (rows) and the categories yielded by a classification tree (columns) using all four acoustic parameters portrayed in Fig.1.

	2cE	2c	2cS	2a	2aA	2b	1a	1b	1bE	3	5
2cE		8									
2c		44	2			1	1	1			
2cS			6								
2a		7		15			6			1	
2aA	1	3		2	10			3		1	
2b						7	1			5	
1a		5		2			52	7			
1b					1		6	52		4	
1bE			1			2		2	5		
3							2	5		67	
5			2								

The performance of the classification tree using all four acoustic parameters is summarized by the confusion matrix in Table 2. It is interesting to observe that the majority of misclassifications occur mainly between categories which are also perceptually close. For example, some confusions occur amongst categories *1a* and *1b* that are distinguished mainly by their duration, categories *1a* and *2a* that are distinguished mainly by the presence of a pitch reset, and categories *2a* and *2c* that are distinguished mainly by perception of curvilinearity in the F0-rising portion. These types of confusions accord with subjects' impressions about the difficulties in deciding between these pairs of categories in some of the samples, suggesting that some phrase final tones may form a perceptual continuum rather than distinct categories.

In the context of automatic categorization, it is worth remembering that the results reported above exclude the 44 phrase finals with problematic F0 measurements. These problems occurred mainly as a result of *C* (creaky), *R* (rough), and *D* (devoiced) phonation types. It is interesting to note that almost all of those samples with {*C, R, D*} phonation-type labels were of the tone-type *1a* category which is typically short, with no pitch reset, no pitch movement, and low F0.

6. Relationship between acoustic parameters and the degree of perceptual pitch change

As hinted in the previous section, during the process of tone category labeling it was found that the subjects had particular difficulties in discriminating among certain of the perceptually more similar categories. To gain a new perspective on the problem, the subjects were asked to rate the perceptual degree

or extent of pitch movement, whether at the reset or within the phrase final. Here, four subjects rated the degree of pitch movement as follows:

- Degree of pitch change within the phrase final, hearing only the phrase final. (Rising degree for category *2c*, and falling degree for category *3*).
- Degree of pitch reset between the phrase final and the syllable immediately preceding the phrase final, hearing the whole phrase. (Categories *3* and *2a*)

The degree of pitch change was annotated based on an 11-point scale, where 0 denotes no change and 10 denotes a large change. Each subject's scores were normalized by Z-scoring (i.e., subtracting the mean and dividing by the standard deviation). The Z-scores were then averaged across subjects to obtain a perceptual reference score for each phrase final, and these averages were used to evaluate the acoustic parameters.

Table 3. Coefficients of correlation between acoustic parameters (slopes and ranges) and perceptual scores of F0 change within the phrase final, for categories 2c and 3.

	slope1	*slope2a*	*slope2b*	*avg2b – avg2a*	*tgt2b – tgt2a*	*tgt2b – avg2a*	*tgt2b – min*	*tgt2b – max*
Move 2c	0.23	0.04	0.29	0.49	0.54	0.54	0.61	-
Move 3	0.22	-0.23	0.45	0.59	0.68	0.66	-	0.65

As shown in Table 3 which considers only the *2c* and *3* tone-types, all of the parameters based on F0-range appear to be perceptually more relevant than those based on F0-slope. A comparison of the fifth and seventh columns of the table also reveals that in defining the extent of F0 movement, *F0tgt2b* is perceptually more relevant than *F0avg2b*. However, as the highest correlations in the table are not significantly different, more detailed analyses are required to identify the perceptually most relevant combination of acoustic parameters.

Table 4. Coefficients of correlation between the F0 reset-related acoustic parameters and the perceptual degree of pitch reset prior to the phrase final, for categories *3* and *2a*.

	max -prev	*tgt2a-prev*	*avg2a-prev*
Reset 3, 2a	0.78	0.74	0.79

The coefficients of correlation listed in Table 4 reveal a good correspondence between perceptual scores and acoustic parameters for the pitch reset degree of categories *3* and *2a*. Furthermore, scatter-plots for the three acoustic parameters discussed earlier in Section 5 show a reasonable correspondence with the relevant perceptual degrees of pitch rise (*2c*), fall (*3*), and reset (*3* and *2a*), respectively.

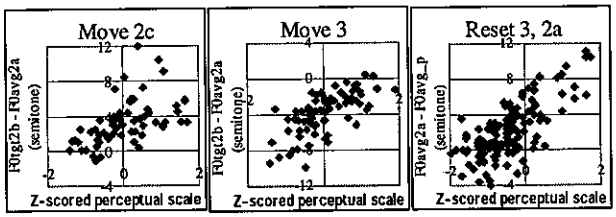

Figure 2. Relationship between acoustic parameters and perceptual scores. (a) *F0move* vs. perceptual rising degree (category *2c*). (b) *F0move* vs. perceptual falling degree (category *3*). (c) *F0reset* vs. perceptual pitch reset degree (categories *3* and *2a*).

7. Conclusion

In this paper we have proposed acoustic parameters that quantify F0 changes in phrase finals, and investigated how they were related with perceptual categories of phrase final tones. A preliminary evaluation of automatic categorization of those tones resulted in 76% correct classification for the best combination of only four parameters. Both qualitative analyses and automatic categorization results showed *F0move* = *F0tgt2b – F0avg2a* as the best parameter for quantifying F0 changes within a phrase final, and *F0reset* = *F0avg2a – F0avg_p* as the best parameter for quantifying the pitch reset prior to the phrase final. Perceptual experiments were also conducted to quantify the degree of pitch movement within the phrase final and of the pitch reset prior to the phrase final. Good correlations were obtained between these perceptual scores and both *F0move* and *F0reset* acoustic parameters. These results also suggest the possibility of regarding some phrase final tones as a continuum rather than as discrete categories. In addition, we note that while the errors in automatic categorization are partly due to perceptual confusions between some of the tone categories, they are also partly due to missing F0 values mainly in segments with non-modal phonation types which occur frequently in our spontaneous speech data. These problems are now being investigated, and experiments are being conducted for automatic detection of phonation types.

Acknowledgements

We would like to thank Masaya Hanazono and Chikako Oura, both of NAIST, for contributing to the automatic segmentation and classification analyses of phrase finals. We also thank all members of the CREST/ESP group, especially Minako Kimura and Kyoko Nakanishi, for the valuable discussions about the labeling task.

References

[1] Toki, T., Murata, M. *Pronunciation & Task Listening - Innovative Workbooks in Japanese*, Aratake Publishers, 37-55. (1987) (in Japanese)

[2] Sugito, M. *Accent, Intonation, Rhythm and Pause*, Sanseido Publishers, 169-202. (1997) (in Japanese)

[3] Hattori, T. "Tones of sentence final particles," Journal of Japanese Language and Literature, Doshisha Women's College, Vol. 14, 1-16. (2002) (in Japanese)

[4] Kikuchi, H., Igarashi, Y., Yoneyama, S. and Maekawa, K. "X-JToBI Reference Manual ver.1.3" 11-42. (2002) (in Japanese)

[5] The JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp

[6] Scott, S. "P-Centres in speech – an acoustic analysis," PhD thesis, Univ. College London. (1993)

[7] Sato, H. "Temporal characteristics of spoken words in Japanese," *JASA*, Vol. 64, Sup. No. 1, S113. (1978)

[8] Ishi, C., Hirose, K., and Minematsu, N. "A study on isochronal mora timing of Japanese," Proc. *Acoust. Soc. Japan*, Vol.I, 199-200. (Sep. 2000) (in Japanese)

[9] Zwicker, E. "Calculating loudness of temporally variable sounds," *JASA*, Vol. 62, No. 3, 675-682. (1977)

[10] Ishi, et al. "Investigation on perceived pitch and observed F0 features to represent Japanese pitch accent patterns." Proc. *Int. Conf. Speech Process.*, Vol. 1, 437-442. (2001)

[11] Nabelek, I., Nabelek, A.; Hirsh, I. "Pitch of Tone Bursts of Changing Frequency," *JASA*, Vol 48, No.2, 536-553. (1970)