Acoustic-prosodic analysis of phrase finals in Expressive Speech

Carlos Toshinori Ishi & Nick Campbell
JST/CREST at ATR/Human Information Science Labs.

1. Introduction

The prosody of phrase finals in Japanese utterances carries a lot of functional information. For example, it carries grammatical information like the modality of the sentences (declarative vs. interrogative), focus, punctuation of phrase boundaries, indication of continuity of the sentence, etc. It also carries paralinguistic information, like manner and attitude of the speaker.

In the research field of Linguistics and Phonetics, there are a lot of researches proposing the categorization of the sentence final intonation [1,2,3]. However, there are few researches regarding automatic categorization.

Also, the variation of the prosody in phrase finals is higher in spontaneous speech rather than in read speech. In this sense, a labeling method called X-JToBI [3] had been proposed to describe these variations, but automatic labeling is still not possible.

The current research has as a goal the automatic prosodic labeling of a large database of spontaneous and expressive speech collected in the CREST/ESP Project [4]. We focused on the description and automatic categorization of phrase final prosody, and analyzed the relationship between the categories perceived by humans and the acoustic-prosodic features measured from the speech signal.

2. Analysis unit and definition of phrase finals

As speech database for analysis, the natural conversation speech data recorded in the CREST/ESP Project was used. As utterance unit, we used prosodic phrase. The prosodic phrases were segmented semi-automatically, when pauses or pitch resets were evident in the phrase boundaries. 404 phrases included in natural conversations with family and company were used in the analysis.

In the current paper, "phrase finals" are 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 of the phrase without the initial consonant. This definition is based on the evidence that the perceptual rhythm beat position (Perceptual Center, or P-Center) is close to the vowel onset instant [5].

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

3. Categorization and Labeling of the Phrase finals

In [2], the tone types of the sentence final particle are categorized as follows:

1a	Low	Ex: na⌉i ne
1b	Low + Falling tone	Ex: na⌉i ne⌉-
2a	High	Ex: na⌉i⌈ne
2b	High + Lengthened	Ex: na⌉i⌈ne-
2c	Low + Rising tone	Ex: na⌉i ne⌈ -
3	High + Falling tone	Ex: na⌉i⌈ne⌉-

In X-JToBI, the following labels are proposed for phrase boundary pitch movements: {L% (= 1a), L%+H% (= 2a, 2b), L%+HL%(= 3), L%+LH%(= 2c), L%+HLH%}. The codes within the parenthesis are the corresponding tones as proposed in [2]. In X-JToBI, we can see 1b is not described, and information about stretching of the phrase final is not completely represented.

In the present research, the above references were considered to propose the following label set:

- Phrase final length: Short (S), Long (L), Very Long (VL), Extremely Long (EL).
- Phrase final tone: Flat-Rise (FtRs), Rise (Rs), Flat (Ft), Fall (Fa), Flat-Fall (FtFa), Fall-Rise (FaRs).
- Pitch reset: Reset, No reset.

Although most of researches related to intonation are based only on F0 information, without considering phonation types, the non-modal phonation types (like creaky, harsh and whispery) are very frequent in natural speech. Further, the reliability of the estimated F0 values is smaller especially in these non-modal phonation type regions than in modal phonation type ones. So, we took special cares in the F0 estimation (section 4), and we also decided to annotate phonation type labels in the present research. The following label set was proposed.

- Phrase final phonation type: Modal (M), Creaky (C), Whispery (W) (aspiration when speaking laughing), Devoiced/ Deleted (D), Low energy (L) (when the airflow is increasingly lowering).

One native speaker of Japanese labeled these categories for the 404 phrase finals.

4. F0 estimation

In this section, we focused on the problems in voiced/unvoiced decision and selection of important F0 values for pitch perception.

As for F0 estimation, we used a method based on auto-correlation function. Specifically, first the residual signal obtained from LPC inverse filter of the speech signal is low-pass filtered to calculate its 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 voiced/unvoiced 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 proper to define a fixed threshold for all F0 candidates in the voiced/unvoiced decision. Here, we used

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

as normalization. This normalization minimizes the effects of reduction of *Rxx(i)* as i increases, leading to a more suitable voiced/unvoiced decision.

The following steps were proposed for post-processing (removal of unreliable values) of F0.

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

With these restrictions, a more perceptually relevant F0 values are obtained.

5. Acoustic-prosodic features

- Phrase final duration (*dur*)
- F0 slope: slope obtained from the reliable F0 values within the phrase final, by first order regression analysis (*F0slope1*). For long phrase finals (more than 120 ms), divide the segment in 2 halves and calculate the slopes of each half (*F0slope2a* and *F0slope2b*). Slopes were computed only when 3 or more (non-zero) F0 values were present in the segment.
- F0 movement: difference of the target F0 values obtained in the two halves after splitting the phrase final by two segments (*F0diff*). The target F0 value of a segment is estimated as the average of the F0 values of the final portion of the segment, as proposed in [7].
- F0 reset: difference between the target F0 of the segment right previous to the phrase final and the target F0 of the first half of the phrase final (*F0reset*).

6. Analysis results

The acoustic parameters were arranged according to the labeled categories. Fig. 1 shows the histograms of each acoustic parameter (*F0slope1, F0slope2a, F0slope2b*).

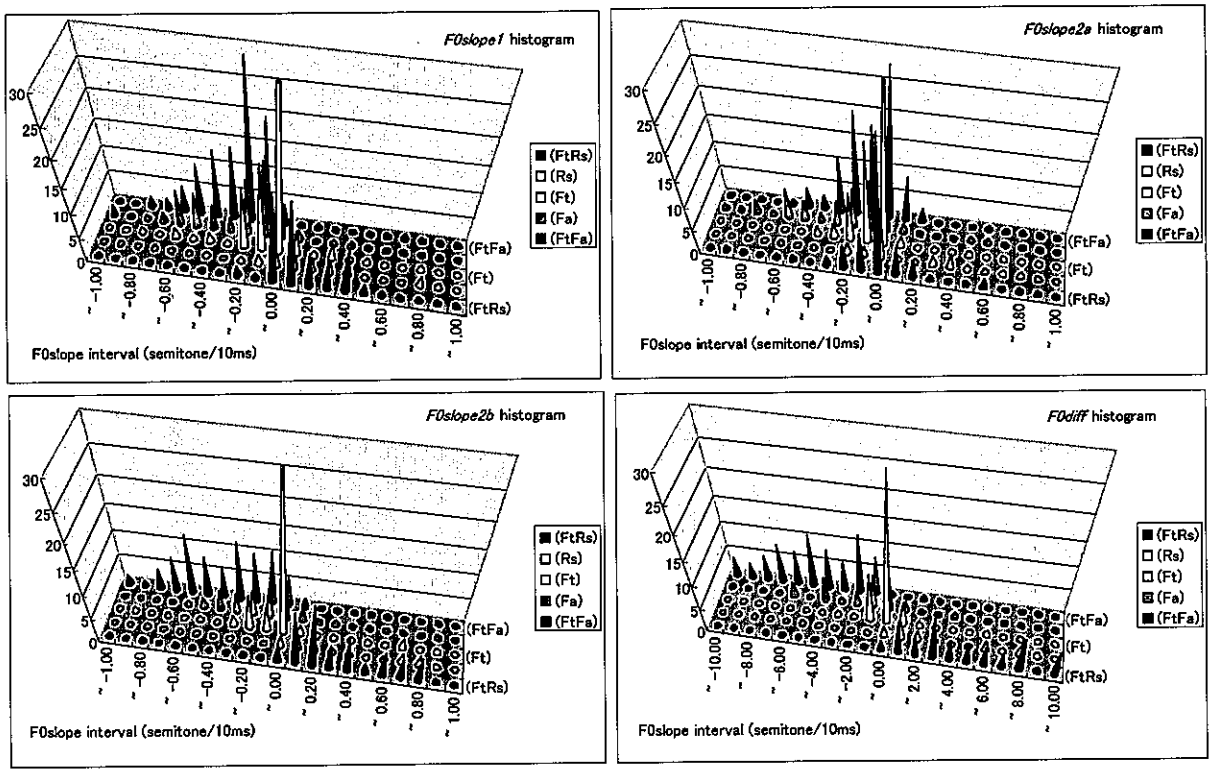

Fig. 1. Histograms of each tone category for each acoustic parameter

From the figure, we can see a better (visual) discrimination between the categories in *F0slope2b* rather than in *F0slope1*, especially around the slope 0 region. This indicates the slope of the second half is more proper than the slope of the whole segment to discriminate the categories. In *F0slope2a*, all categories concentrate data around the slope 0 region, such that no clear discrimination is remarkable. This is because of the F0 co-articulation between syllables, such that the first half of the phrase final syllable is influenced by the F0 of the previous syllable. *F0diff* shows similar tendencies in the histogram as *F0slope2b*. This point will be discussed in the automatic classification (section 7).

The separation between *FtRs* and *Rs*, and between *FtFa* and *Fa* is not clear even for *F0slope2b* (and *F0diff*), so that the identification of these categories were realized using the phrase final duration.

7. Automatic classification of phrase finals

The thresholds of *F0slope2b* and *F0diff* obtained from analysis were used to evaluate the automatic classification of phrase finals. As results, 61% of correct identification was obtained when using *F0slope2b*, and 63% when using *F0diff*. This close identification rate may be because the calculation of both parameters are based on first regression analysis. However, *F0diff* represents a range, while *F0slope2b* represents a slope, and we are not sure which of these features (or maybe both) humans perceive to classify phrase final tones.

In the automatic categorization, confusions were obtained between *Fa* and *Ft*. These results are in agreement with the labeler's impression of difficulty in the discrimination between these categories. From this perspective, we can say that the perceptual discrimination between these categories is difficult, and therefore, a fusion of categories could be convenient.

Independent experiments were conducted for pitch reset identification, after set a threshold for *F0reset* parameter. Results indicated 83% of correct identification.

Problems in F0 estimation occurred mainly in the phrase finals with {C,L,D} phonation type labels, such that a proper *F0slope* calculation was not possible in these segments. However, it was noted that almost all samples with {C,L,D} phonation type labels were labeled as *Ft* or *Fa* for F0 movement.

8. Conclusion

In order to describe prosodic categories of phrase finals, we investigated the relationship between acoustic features quantifying the F0 movement and the labels related to perceived pitch movement, length and phonation types of phrase finals. Using the analysis results, thresholds were set for each acoustic parameter in order to evaluate automatic categorization of the phrase final tones. The results of automatic categorization were not so good, probably because of the segments with low reliability in F0 estimation, and perceptual confusions between some of the categories. These problems are now being investigated. Analysis will also be conducted for automatic detection of phonation type.

Acknowledgements

We would like to thank Masaya Hanazono and Chikako Oura, both of NAIST, to contribute in the analysis for automatic segmentation and classification of phrase finals. We also thank all members of the CREST/ESP group, especially Parham Mokhtari for the valuable discussions and advices, and Minako Kimura for helping the labeling task.

References

[1] 土岐哲「発音・聴解」外国人のための日本語例文・問題シリーズ 12, 荒竹出版, 37-40. (1987)

[2] 服部匡「終助詞の音調について」同志社女子大学日本語日本文学, 第 14 号, 1-16. (2002)

[3] 菊地、五十嵐、米山、前川「X-JToBI リファレンスマニュアル ver.1.3」11-42. (2002)

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

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

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

[7] Ishi, Hirose, Minematsu. "Investigations on a quantified representation of pitch movements in syllable units," Proc. of *Acoustic Society of Japan Spring* Meeting 2002, Vol. I, 419-420. (2002)