Projectability of Transition-Relevance Places Using Prosodic Features in Japanese Spontaneous Conversation

Yuichi Ishimoto[1], *Mika Enomoto*[2], *Hitoshi Iida*[2]

[1]Speech Media Group, National Institute of Informatics, Tokyo, Japan
[2]School of Media Science, Tokyo University of Technology, Tokyo, Japan
ishimoto@nii.ac.jp, {menomoto, iida}@media.teu.ac.jp

Abstract

In this paper, to clarify acoustic features for predicting the ends of utterances, we investigated prosodic features that project transition relevance places in Japanese spontaneous conversation. Acoustic parameters used as the prosodic features are the fundamental frequency, power, and mora duration of accentual phrases and words. Results showed that the fundamental frequency and power at the beginning of the final accentual phrase indicate whether the utterance includes utterance-final elements, which are the syntactic cue for detecting the end-of-utterance. In addition, the mora duration lengthened in the final accentual phrase. That is, these prosodic features around the beginning of the final accentual phrase showed the characteristic changes that make hearers predict the transition relevance places.

Index Terms: turn-taking, prosody, utterance-final element, accentual phrase

1. Introduction

In spontaneous conversations, we can maintain smooth transfers from one speaker to another without gaps. This means that we unconsciously predict the ends of utterances in some way. Sacks et al.[1] proposed the turn-taking system that employs a turn constructional unit (TCU) as an utterance unit in turn-taking. In this system, a turn is composed of one or more TCUs. There is a transition-relevance place (TRP) at the end of each TCU, and turn-taking could occur at a TRP. It is thought that various factors constitute TRPs[2].

Tanaka[3] has identified certain words indicating the beginning of the TRP as utterance-final elements in Japanese. These elements consist of auxiliary verbs (such as */desu/* and */masu/*), sentence-final particles (such as */ne/* and */yo/*), and so on. These are placed at the end of the sentence as a syntactic factor and project the completion of a TCU. However, we cannot always use the utterance-final elements to find the TRP, because utterances without such utterance-final elements exist. Tanaka observed that rising- or falling-intonation, stresses, and the lengthening of sentence-final morae occur for the *iikiri* form without utterance-final elements, suggesting that these features could be used instead of the utterance-final elements to detect the TRP.

Koiso et al.[4] investigated syntactic and prosodic features appearing at the end of inter-pausal units as points where turn-taking occur. According to their results, prosodic features such as duration and fundamental frequency (F_0) contour patterns at the final mora of inter-pausal units depended on whether or not the speaker changed at the boundaries of the inter-pausal units. However, in spontaneous conversations, we do not distinguish the beginning of a TRP from the final mora of the inter-pausal unit, because the final mora might be uttered after the onset of the next speaker's utterance. The beginning of the final mora is too late for speech planning by the next speaker. That is, acoustic features prior to the final mora are needed for prediction of the TRP.

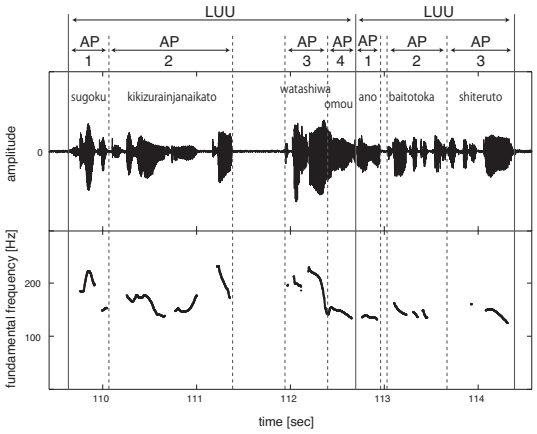

Figure 1: *Example of LUUs and accentual phrases (APs).*

Maekawa[5] analyzed F_0 declination in utterances consisting of two to five accentual phrases (APs), and noted the following interesting phenomena: (1) the final APs are always much lower than the non-final APs, and (2) the F_0 ranges in the final APs are roughly 100–120 Hz regardless of the number of APs. These observations indicate that speakers design the prosodic structure toward the final APs according to the number of APs, and they suggest that TRP signals exist in the prosodic structure.

The aim of this study is to clarify acoustic features for predicting the ends of utterances. In analysis 1, by comparing prosodic features of different APs, we investigate which parts of utterances exhibit acoustic changes relating to the end-of-utterance. In analysis 2, we show that the changes in the prosodic features differ depending on whether or not the utterance-final elements are present.

2. Data

Twelve dialogues from the Chiba three-party conversation corpus[6], which are casual conversations on different themes, are used for this study. These dialogues are annotated with words and tone structures such as boundary tones and break indices, by using the X-JToBI scheme[7].

We focus on long utterance units (LUUs)[8] with boundaries at which turn-taking can occur. The LUUs are designed to

Table 1: *Prosodic features.*

Sign	Explanation
F_0mean	Mean value of fundamental frequencies (F_0s)
F_0max	Maximum value of F_0s
F_0min	Minimum value of F_0s
F_0range	Range of F_0s
F_0slope	Slope of F_0 contour
Pmean	RMS power
Pmax	Maximum value of short-term power
Prange	Range of short-term power
D	Duration
Dmora	Average mora duration

segment dialogs by syntactic and pragmatic disjunctures such as clause boundaries, linguistic modalities, or turn-completing tokens. Den et al.[8] reported that the timing of turn-taking was localized at the LUU boundaries. We therefore substitute LUUs for TCUs. Furthermore, we split the LUUs into APs for the analysis performed in the next section. Figure 1 illustrates the relationship of the APs and the LUUs. The top panel shows part of a Japanese speech wave, and the bottom panel shows the F_0 contour of the speech. The LUU contains one or more APs. We obtained 2,252 LUUs consisting of two to ten APs.

3. Analysis 1: Changes of prosodic features between accentual phrases

Maekawa[5] reported that F_0s decline toward the end-of-utterance at every AP and their ranges are 100–120 Hz in the final APs. If other prosodic features also indicate distinctive changes toward the end-of-utterance, hearers may perceive the changes to predict the end-of-utterance. In this section, we analyze temporal changes in prosodic features between adjacent APs, and we investigate these features as acoustic clues for predicting the end-of-utterance.

3.1. Method

Table 1 lists the acoustic parameters used as prosodic features. The F_0s were extracted at 1-ms intervals. To avoid influences from gender and individual differences, the logarithmic F_0s were normalized by the mean value of the F_0s for each LUU. F_0mean, F_0max and F_0min are the mean, maximum, and minimum values of the F_0s for each AP, respectively. F_0range is the difference between F_0max, and F_0min. F_0slope is the slope of the linear regression line calculated from the F_0 contour for each AP. The sound pressure of the APs was normalized by the sound pressure of each LUU. This means that the relative sound-pressure level was calculated using the sound pressure of the LUU as a reference pressure. Pmean is the sound pressure level obtained from the effective value of the sound pressure of the AP (i.e. the RMS power). Pmax is the maximum value of the short-term sound-pressure levels calculated in 10-ms window lengths at 1-ms intervals. Prange is the difference between Pmax and the minimum sound-pressure levels, which are equal to the level of the background noise. Duration (D) is the time from the start to the end of the AP. Dmora is the average duration normalized by the number of morae in the AP, which is called the average mora duration.

To see the differences among the AP positions in the LUUs, we applied linear mixed-effects models and obtained p-values using Markov chain Monte Carlo (MCMC) sampling.

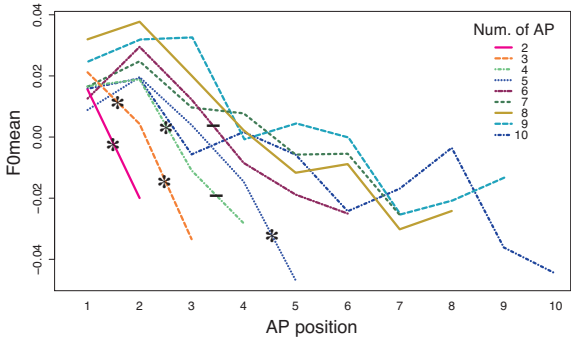

Figure 2: *F_0mean for each AP position.*

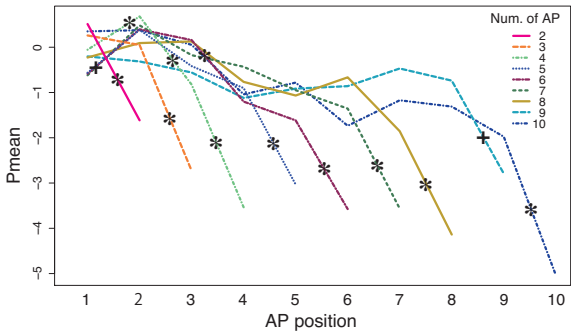

Figure 3: *Pmean for each AP position.*

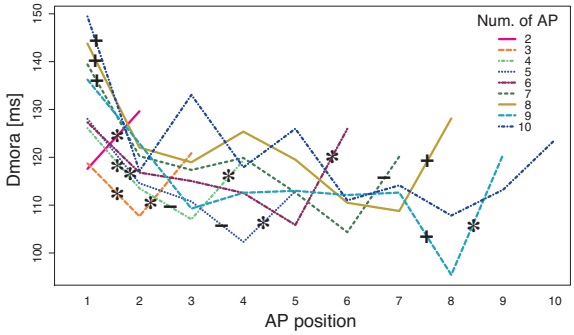

Figure 4: *Dmora for each AP position.*

3.2. Results

The MCMC results indicate that the main effects of F_0mean, F_0max, F_0min, Pmean, Pmax, D, and Dmora were significant for LUUs consisting of less than seven APs ($p<.05$).

Figure 2 shows the F_0mean for each AP position. Here "*" and "+" indicate a significant difference between adjacent APs: "*" indicates a significance level of 1%, and "+" is a level of 5%. Whereas F_0 showed a steep decline everywhere for LUUs with few APs, it changed only slightly between adjacent APs for LUUs with many APs. Figure 3 shows the Pmean for each AP position. The power decreased significantly in the final AP. In LUUs with many APs, the power fell markedly in the final AP, although it decreased only slightly in comparison with the previous AP. Figure 4 shows the Dmora for each AP position. The mora duration of the final AP increased remarkably.

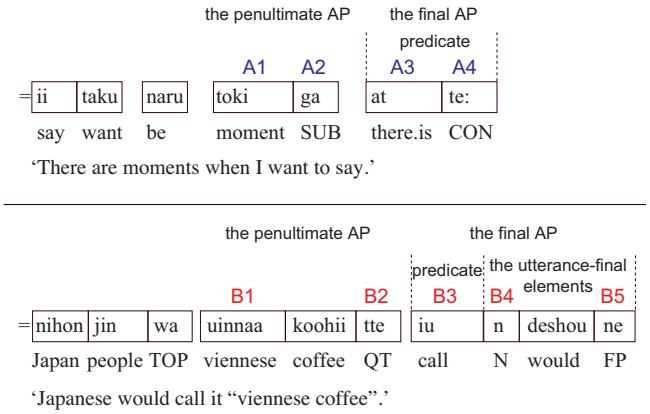

Figure 5: Analysis points in the final AP and the penultimate AP, where SUB denotes the subject marker; CON, conjunction; TOP, topic marker; QT, quotation marker; N, nominalizer; and FP, final particle.

3.3. Discussion

The changes in the prosodic features between the APs can be summarized as follows:

- When there are few APs, F_0 drops rapidly at the second or third AP.
- F_0 reaches its lowest value at the final AP.
- The power decreases significantly in the final AP.
- The average mora duration increases in the final AP.

The F0 shift between the first and second APs is useful for predicting whether or not there are three or more APs in the LUU. A rapid decrease in power and an extended duration of the AP constitute a cue for detecting that the AP is the final one in the LUU.

4. Analysis 2: Changes in prosodic features in final accentual phrase

We have shown in analysis 1 that the prosodic features in the final AP mark existence of the end-of-utterance. However, if the changes in the prosodic features appear at the end of the utterance, for example in the final mora, they are not useful to the next speaker because he/she needs to begin speech preparation long before this point.

Tanaka[3] have suggested that the utterance-final elements indicate the beginning of a TRP and that turn-taking occurs after the emergence of these elements. If so, in the case of utterances without such elements, the next speaker has to begin without waiting for the emergence of the utterance-final elements. That is, the next speaker predicts whether or not the utterance-final elements will occur, and sets a starting time. To enable this prediction, there must be some change in the prosodic features near the beginning of the final AP.

In this section, we clarify places at which acoustic cues of the end-of-utterance appear, by a comparison of the prosodic features at the beginning and end of the final AP and the penultimate APs for utterances with and without utterance-final elements.

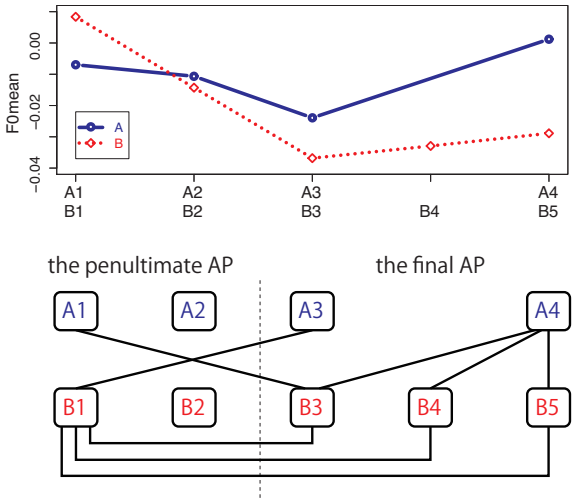

Figure 6: *(Top) F_0mean for each analysis point. (Bottom) Results of multiple comparison. Solid lines indicate significant differences ($p<0.01$).*

4.1. Method

We selected the LUUs for which the final AP begins with a predicate, and we obtained 543 LUUs without utterance-final elements and 752 LUUs with such elements. The F_0mean, Pmean, and Dmora were extracted for each word at the analysis points in the LUUs. As shown in figure 5, in the group without utterance-final elements, the analysis points were the words occurring at:

A1 - the beginning of the penultimate AP;
A2 - the end of the penultimate AP;
A3 - the beginning of the final AP;
A4 - the end of the final AP.

In the group with utterance-final elements, the analysis points were the words occurring at:

B1 - the beginning of the penultimate AP;
B2 - the end of the penultimate AP;
B3 - the beginning of the final AP;
B4 - the beginning of the utterance-final element;
B5 - the end of the final AP.

We carried out the one-way ANOVA between the analysis points.

4.2. Results

The top panels of Figs. 6–8 show F_0mean, Pmean, and Dmora at the analysis points. The ANOVA results indicated that there were significant differences for the main effects of all these features. For F_0mean F(8, 2407)=12.34 and $p<0.001$; for Pmean F(8, 2403)=44.15 and $p<0.001$; and for Dmora F(8, 2378)=113.88 and $p<0.001$. The results of the multiple comparisons are shown in the bottom panels of Figs. 6–8. Solid lines connecting the analysis points indicate significant differences ($p<0.01$).

From Fig. 6 we observe that:

- If there are no utterance-final elements, F_0 is the same at the beginnings of the final AP and the beginning of the penultimate AP (points A1 and A3)

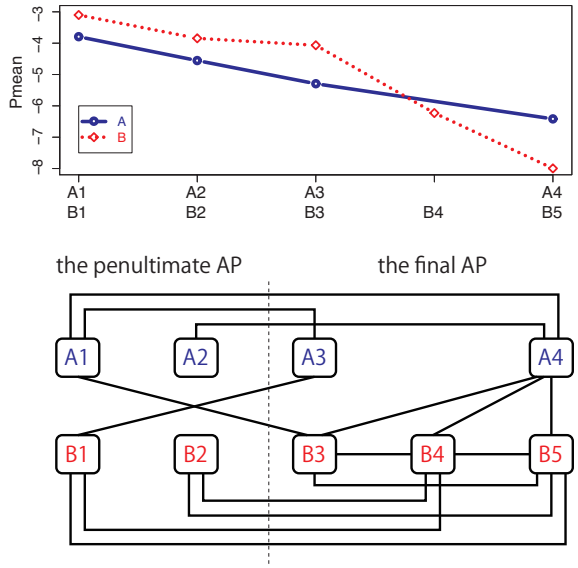

Figure 7: *(Top) Pmean for each analysis point. (Bottom) Results of multiple comparison. Solid lines indicate significant differences ($p<0.01$).*

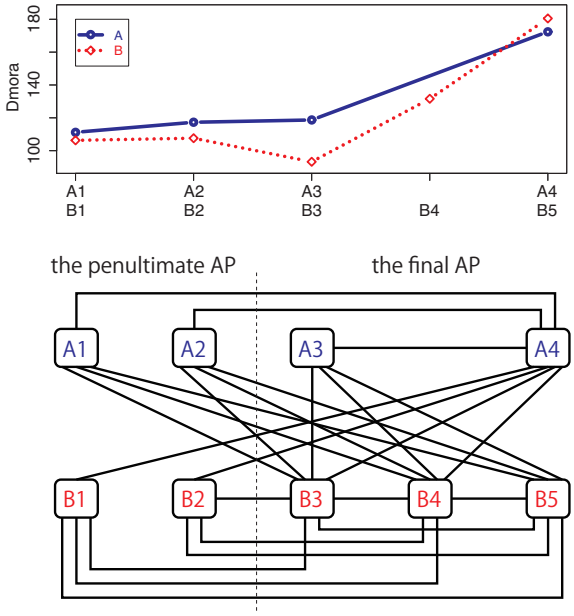

Figure 8: *(Top) Dmora for each analysis point. (Bottom) Results of multiple comparison. Solid lines indicate significant differences ($p<0.01$).*

- If there are such elements, F_0 is lower at the beginning of the final AP (points B1 and B3).

From Fig. 7 we observe that:

- If there are no utterance-final elements, the power is lower at the beginning of the final AP than at the beginning of the penultimate AP (points A1 and A3).
- If there are such elements, the power falls not at the beginning of the final AP but at the utterance-final elements (points B1 and B4).

From Fig. 8 we observe that:

- The mora duration begins to increase in the final AP (points A3 and A4; B3 and B5) and is almost the same at the end of the utterance (points A4 and B5)
- If there are no utterance-final elements, the mora duration is longest at the end of the predicate (point A4).
- If there are such elements, the mora duration still increases from the end of the predicate to the end of the utterance (points B4 and B5).

4.3. Discussion

These results show that we can distinguish the presence or absence of utterance-final elements in the AP using observations of F_0 and the power at the beginning of the AP. That is, hearers can decide if they detect the TRP by waiting for the emergence of the utterance-final elements, or they can decide during the predicate. Furthermore, an increase in the mora duration in the AP indicates that this is the final AP. This may support the results detected using F_0 and the power. However, in the final AP without utterance-final elements it is not clear yet where the mora duration starts to increase.

5. Conclusions

We investigated prosodic features for predicting the TRP in spontaneous Japanese conversation. We showed that with utterance-final elements the F_0 at the beginning of the final AP was lower than that at the beginning of the penultimate AP; without such elements the power at the beginning of the final AP was lower than that at the beginning of the penultimate AP; and the mora duration increased in the final AP. That is, the changes in the prosodic features that project the ends of utterances appeared near the beginning of the final AP. We believe that hearers can use the prosodic features to judge whether to use utterance-final elements or the predicate part as the TRP.

6. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] C. E. Ford and S. A. Thompson, "Interaction units in conversations: Syntactic, intonational, and pragmatic resources for the management of turns," in *Interaction and grammar*, E. Ocks, E. A. Schegroff, and S. A. Thompson, Eds. Cambridge University Press, 1996, pp. 134–184.

[3] H. Tanaka, *Turn-taking in Japanese conversation: a study in grammar and interaction*. John Benjamins Publishing, 1999.

[4] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[5] K. Maekawa, "Final lowering and boundary pitch movements in spontaneous Japanese," in *Proc. DiSS-LPSS Joint Workshop 2010*, 2010, pp. 47–50.

[6] Y. Den and M. Enomoto, "A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation," in *Conversational informatics: An engineering approach*, T. Nishida, Ed. John Wiley & Sons, 2007, pp. 307–330.

[7] K. Maekawa, H. Kikuchi, Y. Igarashi, and J. Venditti, "X-JToBI: An extended J_ToBI for spontaneous speech," in *Proc. ICSLP2002*, 2002, pp. 1545–1548.

[8] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida, "Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme," in *Proc. LREC2010*, 2010, pp. 2103–2110.