Proceedings of DiSS-LPSS Joint Workshop 2010,
25-26 September 2010, Tokyo, Japan, pp. 97-100.

Analysis of Prosodic Features for End-of-utterance Prediction in Spontaneous Japanese

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

[1]Speech Media Group, National Institute of Informatics, Japan
[2]School of Media Science, Tokyo University of Technology, Japan

ishimoto@nii.ac.jp, menomoto@media.teu.ac.jp

Abstract

In this study, we analyzed prosodic features of accentual phrases and investigated their temporal changes to obtain cues for detecting boundaries at where turn-taking could occur in spontaneous conversations. The acoustic parameters used as prosodic features were the fundamental frequency, sound pressure level, and duration of accentual phrases in long utterance units. The results showed that the fundamental frequency shift between the first and second accentual phrases could be useful for detecting the number of accentual phrases in the long utterance unit. In addition, the results suggested that a rapid decrease in sound pressure and an extended duration of the accentual phrase constitute a cue for detecting the end of the utterance. That is, the acoustic predictor of the utterance length appeared at the beginning of the utterance, and the predictor of the utterance boundary appeared shortly before the end of the utterance.

Index Terms: prosody, turn-taking, accentual phrase, long utterance unit

1. Introduction

The aim of this study is to clarify the acoustic features that affect perception or prediction of the end of an utterance and that could lead to turn-taking in spontaneous conversations.

In spontaneous conversations, we can smoothly maintain transfers of speakership without gaps. This means that we unconsciously predict the ends of utterances in some way. The turn-taking system proposed by Sacks et al.[1] employed a turn constructional unit (TCU) as an utterance unit about turn-taking. A turn is composed of one or more TCUs in this system. 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]. We suppose that acoustic features which characterize TRPs or predict their existence are important cues.

Koiso et al.[3] 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 pattern at the final mora of IPUs depended on whether the speaker changed or not 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 onset of the next speaker's utterance. The beginning of the final mora is too late for speech planning of the next speaker. That is, acoustic features prior to the final mora are needed for prediction of the TRP.

Maekawa[4] has analyzed F_0 declination in the utterances consisting of two to five accentual phrases (APs), and notes interesting phenomena as follows: (1) the final AP are always much lower compared to the measurement points in the non-final APs and (2) the F_0 range in the final APs are roughly 100-120 Hz regardless of the number of the APs. These observations mean that the speaker varies the prosodic structure toward the final APs according to the number of APs in the utterance, and suggest the existence of cues for the TRP detection

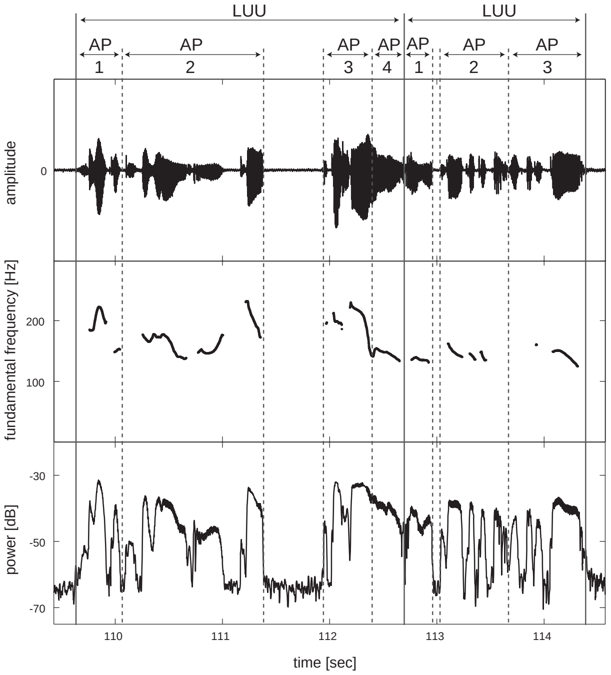

Figure 1: *Relation of APs to LUUs.*

Table 1: *Prosodic features.*

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

Table 2: *p-value using MCMC (Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05).*

Num. of AP	F_0mean	F_0max	F_0min	F_0range	F_0slope
2(N=1329)	0.0001***	0.0001***	0.0001***	0.0038**	0.0044**
3(N=1226)	<0.0001***	<0.0001***	<0.0001***	0.0083**	0.0014**
4(N=954)	<0.0001***	0.0061**	<0.0001***	0.005**	0.1186
5(N=713)	<0.0001***	0.0036**	<0.0001***	0.6784	0.351
6(N=561)	<0.0001***	0.0079**	<0.0001***	0.3848	0.0815

Num. of AP	Pmean	Pmax	Prange	D	Dmora
2(N=1329)	0.0001***	0.0001***	0.1472	0.0001***	0.0001***
3(N=1226)	<0.0001***	<0.0001***	0.0651	<0.0001***	<0.0001***
4(N=954)	<0.0001***	<0.0001***	0.4184	<0.0001***	<0.0001***
5(N=713)	<0.0001***	<0.0001***	0.9408	<0.0001***	<0.0001***
6(N=561)	<0.0001***	<0.0001***	0.7132	<0.0001***	0.036*

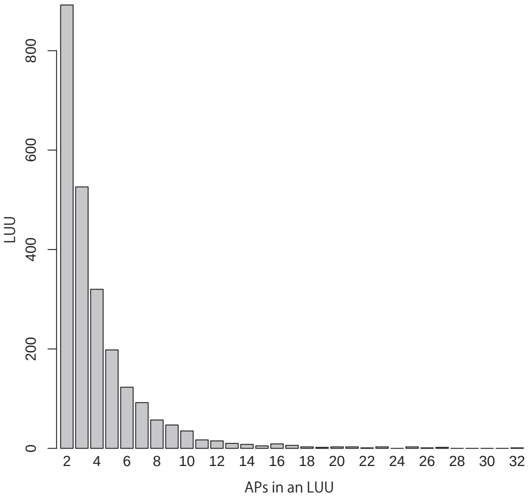

Figure 2: *LUUs for the number of APs in an LUU.*

in the prosodic structure.

Characteristics of the TRP should not suddenly appear in the end of the TCU because speech components such as contour of F_0 and intensity form a smooth trajectory. Therefore, we analyzed prosodic features and investigated their temporal changes to obtain cues for their detection.

2. Analysis of prosodic features

2.1. Data

Twelve dialogues from the Chiba three-party conversation corpus[5] were used for this study. These dialogues were annotated with words and tone structures such as boundary tones and break indices, by using the X-JToBI scheme[6]. In addition, utterance units including inter-pausal units, intonation units, clause units, pragmatic units were annotated.

We focused on long utterance units (LUUs)[7] with boundaries at which turn-taking could occur. The LUUs were designed to segment dialogues by syntactic and pragmatic disjuncture. The dialogues are split into LUUs by the above annotations, because each LUU boundary is expressed by a clause boundary, a linguistic modality, or a turn-completing token. Den et al.[7] reported that the timing of turn-taking was localized at the LUU boundaries. We therefore substituted LUUs for TCUs.

Furthermore, we chose to use APs as a unit of analysis to extract prosodic features, and we split the LUUs into the APs based on break index 2 or 3, both of which mark an accentual phrase break. Figure 1 illustrates relation of APs to LUUs. The upper panel of figure 1 shows a part of Japanese speech wave. The middle and lower panels show F_0 and power contour of the speech, respectively. As illustrated in section 2.2, we analyze F_0 and power as prosodic features in the APs. The LUU has one or more APs. Pauses between APs were excluded from the analysis.

The number of LUUs for each AP number in an LUU is shown in figure 2. In the following analysis, we adopted the AP number with which there were 100 and more LUUs. As a result, we obtained 892 LUUs for two APs in an LUU, 526 LUUs for three APs, 320 LUUs for four APs, 198 LUUs for five APs, and 123 LUUs for six APs.

2.2. Prosodic features

Table 1 lists the acoustic parameters used as prosodic features. These parameters were extracted for each AP.

The F_0s were extracted by SWIPE[8] at 1 ms intervals. To avoid influence from gender and individual differences, the logarithmic F_0s in the APs 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 a linear regression line calculated from the F_0 contour for each AP.

By the same token, 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 by using the sound pressure of the LUU as a reference sound pressure. Pmean is the sound pressure level obtained from the effective value of the sound pressure of the AP. 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 background noise. Duration (D) is the time from the onset to the offset of the AP, and Dmora is the average duration per mora for each AP.

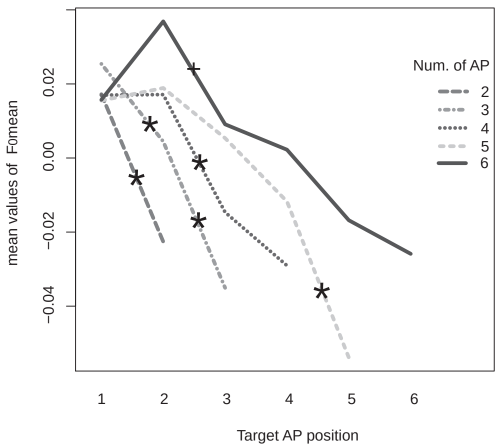

Figure 3: *Mean values of F_0mean for each AP position.*

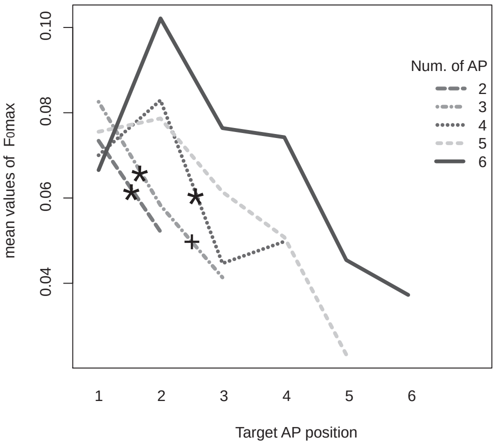

Figure 4: *Mean values of F_0max for each AP position.*

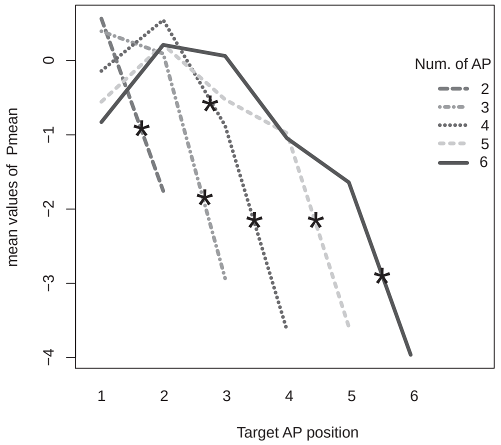

Figure 5: *Mean values of Pmean for each AP position.*

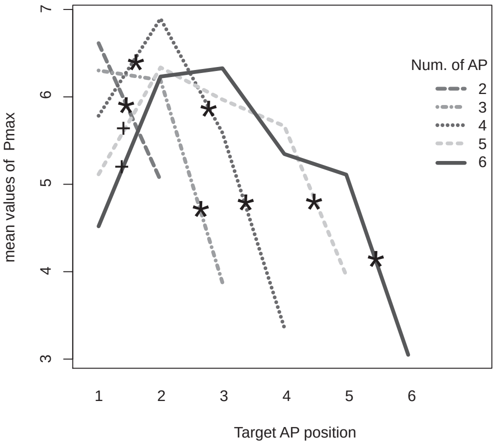

Figure 6: *Mean values of Pmax for each AP position.*

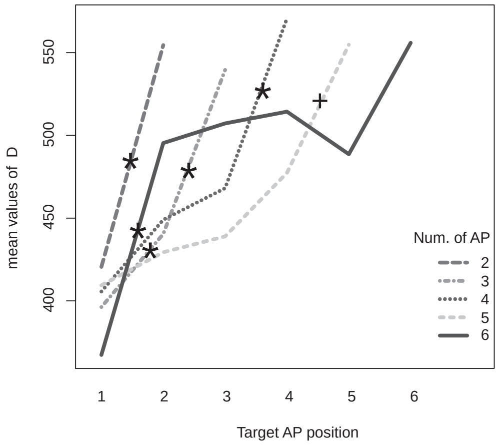

Figure 7: *Mean duration for each AP position.*

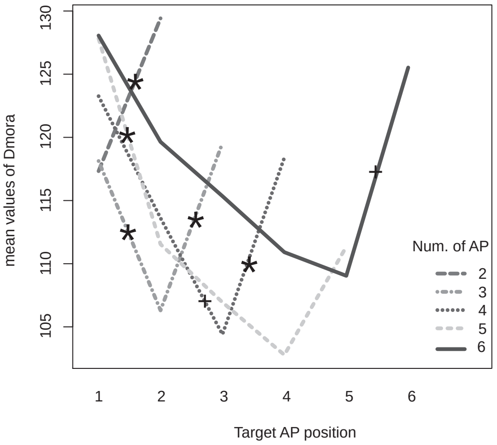

Figure 8: *Mean duration normalized by the number of mora for each AP position.*

3. Results

In order to see the difference among levels, we applied liner mixed-effects models and obtained p-values using Markov Chain Monte Carlo (MCMC) sampling. We used the `lmer()` and `pvals.fnc()` functions from packages `lme4` and `languageR` of the R software environment[9]. Table 2 shows the MCMC results, where N is the total amount of APs except the APs that the feature values were unavailable by distorted speech sounds. The p-values indicate whether the effects of the AP position in the LUU are significant. In particular, the main effects of F_0mean, F_0max, F_0min, Pmean, Pmax, D, and Dmora were significant at all AP numbers.

Figures 3 and 4 show the mean values of F_0mean and F_0max for each AP in the LUU, respectively. The marks of "*" and "+" indicate a significant difference between adjacent APs. "*" means that significant level is 1%, and "+" means the level is 5%. As shown in these figures, while F_0 showed a steep decline everywhere for LUUs with a small number of APs, F_0 changed gently between adjacent APs for LUUs with many APs.

Figures 5 and 6 show the mean values of Pmean and Pmax, respectively. Power decreased significantly in the final AP. In LUUs with many APs, power fell markedly in the final AP, although power decreased more gently in comparison with the previous AP.

Figures 7 and 8 show the duration of AP and average duration normalized by the number of mora in the AP, respectively. In figure 7, the duration of the final AP lengthened remarkably. For LUUs with six APs, the duration of the final AP tended to be longer than that of the previous AP, although no significant difference was found between the APs. In figure 8, the duration per mora also lengthened in the final AP.

4. Discussion

As shown in figure 3, F_0 of the second AP tended to rise once in LUUs consisting of more than three APs. It is thought that by ascertaining whether the F_0 shift between the first and second AP is a rise or fall, we can detect an LUU has more than three APs early in the utterance.

Let us discuss characteristics appearing in the final AP. Figure 5 clearly indicates that power in the final AP fall sharply in all cases. In addition, as is evident in figure 8, the mora duration in the final AP is much longer than that of the previous AP, though the duration in the other APs generally became shorter and shorter toward the end of the utterance. It is considered that we can detect that the AP is the final one in the LUU by recognizing the power drop and the extended duration.

These findings suggest that the acoustic characteristic related to the LUU length is at the beginning of the LUU, and the characteristics related to the LUU boundary are at shortly before the end of the LUU. These acoustic features can be the information which have hearers in conversation predict the end of the utterance. For further discussion, it is necessary to confirm that hearers utilize the above features for predicting the ends of utterances in perceptual experiments.

5. Conclusions

We analyzed the F_0, sound pressure level, and duration of APs in LUUs, which were regarded as turn construction units. The results showed that the F_0 shift between the first and second APs are useful for predicting whether there are less than three APs or more in the LUU. Furthermore, the results suggested that 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. That is, the acoustic predictor of the TCU length appeared at the beginning of the TCU, and the predictor of the TCU boundary appeared shortly before the end of the TCU.

As future work, we plan to analyze of LUUs with seven or more APs. For its part, we show the mean values of Pmean for seven to ten APs in figure 9. Power did not monotonically decrease but slightly increased before the final AP. This result suggests that the speakers prepare to effectively form power drop because the power would approach to a lower limit if it remained to drop.

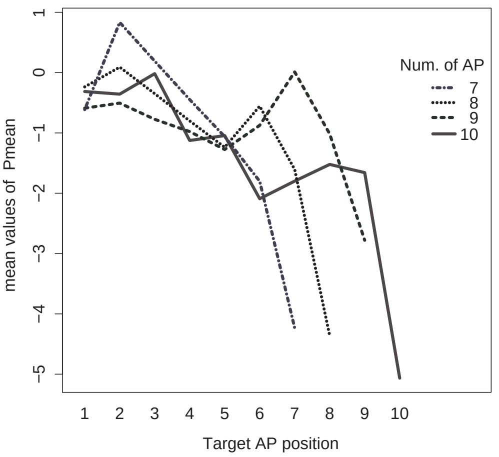

Figure 9: *Mean values of Pmean for seven to ten APs.*

6. Acknowledgments

The authors would like to thank Prof. Yasuharu Den (Chiba University) for suggesting the evaluation by the MCMC.

7. References

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

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

[3] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y., "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs", Language and speech, 41(3–4), 295–321, 1998.

[4] Maekawa, K., "Final lowering and boundary pitch movements in spontaneous Japanese", the 5th Workshop on Disfluency in Spontaneous Speech and the 2nd International Symposium on Linguistic Patterns in Spontaneous Speech (DiSS-LPSS Joint Workshop 2010), 2010.

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

[6] Maekawa, K., Kikuchi, H., Igarashi, Y. and Venditti, J., "X-JToBI: An extended J_ToBI for spontaneous speech", Proc. 7th International Conference on Spoken Language Processing, 1545–1548, 2002.

[7] Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M. and Yoshida, N., "Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme", Proc. the 7th conference on International Language Resources and Evaluation, 2103–2110, 2010.

[8] Camacho, A. and Harris, J. G., "A sawtooth waveform inspired pitch estimator for speech and music", J. Acoust. Soc. Am. 124(3), 1638–1652, 2008.

[9] Baayen, R. H., *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*, Cambridge University Press, 2008.