A Prediction Model for End-of-Utterance Based on Prosodic Features and Phrase-Dependency in Spontaneous Japanese

Yuichi Ishimoto*, Takehiro Teraoka† and Mika Enomoto†
* National Institute for Japanese Language and Linguistics, Tokyo, Japan
E-mail: yishi@ninjal.ac.jp Tel: +81-42-540-4639
† Takushoku University, Tokyo, Japan
E-mail: tteraoka@cs.takushoku-u.ac.jp Tel: +81-42-665-8517
‡ Tokyo University of Technology, Tokyo, Japan
E-mail: menomoto@stf.teu.ac.jp Tel: +81-42-637-2111

***Abstract*—This study aims to reveal a clue for predicting end-of-utterance in spontaneous Japanese speech. In casual everyday conversation, participants must predict the ends of utterances of a speaker to perform smooth turn-taking with small gaps or overlaps. Syntactic and prosodic factors are considered to project the end of utterance of speech, and participants utilize these factors to predict the end-of-utterance. In this paper, we focused on the dependency structure among bunsetsu-phrases as a syntactic feature and F0, intensity, and mora duration for bunsetsu-phrases as prosodic features. We investigated the relationship between the position of a bunsetsu-phrase in an utterance and these features. The results showed that a single feature cannot be an authoritative clue that determines the position of bunsetsu-phrases. Next, we constructed a Bayesian hierarchical model to estimate the bunsetsu-phrase position from the syntactic and prosodic features. The results of the model indicated that prosodic features vary in usefulness according to speakers. This suggests that the different combinations of syntactic and prosodic features for each speaker are relevant to predict the ends of utterances.**

I. Introduction

Conversation is an essential component of everyday human interaction, and turn-taking with minimal gaps or overlaps in the speech of participants ensures smooth communication. Although computer dialogue systems have been in practical use for several years, they cannot make smooth turn-taking between human and computer because they have to detect end-of-utterance by detecting a pause after the utterance. In contrast, participants realize smooth turn-taking by predicting end-of-utterance in spontaneous human-to-human conversations. If the features of utterances that support prediction are clear, the human-system conversation must become more natural.

To explain human turn-taking behavior, Sacks et al. [1] proposed a system that employs a turn constructional unit (TCU) as an utterance unit in turn-taking. According to this system, there is a transition-relevance place (TRP) at the end of each TCU, and turn-taking possibly occurs at the TRP. By this scheme, Ford and Thompson [2] suggested that syntactic, intonational, and pragmatic resources constitute TRPs.

Regarding Japanese prosodic factors, Pierrehumbert and Beckman [3] suggested that the utterance unit is a region in which the fundamental frequencies (F0s) monotonously decline over time, which is called F0 declination, and the F0s fall significantly at the end of the utterance unit, which is called final lowering. According to them, final lowering indicates the end of an utterance. As seen in our previous work, however, final lowering rarely appears in spontaneous speech [4]. Moreover, final lowering cannot be a clue for predicting end-of-utterance because it appears at the end of an utterance [5].

We constructed a generalized linear mixed-model to detect end-of-utterance using syntactic and prosodic features in our previous study [6]. The model achieved high accuracy in estimating final accentual phrases in Japanese utterances. Nevertheless, the final accentual phrase is also at the end of an utterance, and it is unclear whether these syntactic and prosodic features can be used as clues for prediction.

In this paper, we consider different combinations of syntactic and prosodic features to find a clue for predicting end-of-utterance in spontaneous speech. First, we present relationships between a phrase position in an utterance and syntactic and prosodic features. We then construct a model using the syntactic and prosodic features to estimate phrase positions in utterances. Finally, we show the availability of the combination of the features for spontaneous Japanese utterances.

II. Syntactic and Prosodic Features for Bunsetsu-Phrase Positions in Spontaneous Utterances

In this section, we describe syntactic and prosodic features, which are utilized in later sections, and present relationships between the features and their positions in utterances.

TABLE I: Relationship between the difference of unsolved-modifiers and the reverse order of appearance of bunsetsu in the utterances.

		Difference of unsolved-modifiers							
		-6	-5	-4	-3	-2	-1	0	1
	1	1	5	14	60	179	282	113	6
	2	0	0	2	3	11	57	385	216
	3	0	1	0	2	7	38	451	166
Bunsetsu	4	0	0	0	3	1	34	276	118
order	5	0	0	0	2	3	18	165	89
from	6	0	0	0	1	1	9	119	69
the end	7	0	0	0	0	0	10	83	51
	8	0	0	0	1	1	5	68	35
	9	0	0	1	1	1	5	45	32
	10	0	0	0	1	0	3	52	17
	...					...			

A. Data

As spontaneous utterances, we used twelve dialogs from the Chiba three-party conversation corpus (Chiba3Party) [7]. This corpus consists of casual Japanese conversations in 10 minutes on different themes by twelve groups of three people, and the total amount of data is about 120 minutes with 36 speakers. It is difficult to identify the utterance boundaries of spontaneous speech because there is no punctuation when speaking. We adopt the long utterance unit [8] for utterance unit identification.

In our previous study, we showed that syntactic words referred to as utterance final elements (UFEs) could be used for end-of-utterance detection [6]. UFEs are frequently used in a Japanese utterance at the end of which appears a conjunctive particle or the inflection form of a verb succeeded by a subordinate clause avoiding the main clause, and they project the completion of a TCU [9]. These elements consist of auxiliary verbs (such as /desu/, /masu/, and /da/), sentence-final particles (such as /ne/, /yo/, and /ka/), and so on. Enomoto [10] demonstrated that the beginning of the TRP is when hearers recognize the UFE in perceptual experiments. It does not, however, become a reliable indicator of the end of an utterance, because there are utterances without UFEs in spontaneous speech. In order to focus on the features that are not related to UFEs in this study, we excluded utterances with UFEs from the subject of this investigation.

B. Syntactic features

In this study, we adopt the syntactic feature relating to the dependency structure of utterances. A Japanese sentence consists of a sequence of phrasal units called "bunsetsu." The bunsetsu-phrase is a part of a phrase that cannot be divided further in the Japanese language. A bunsetsu depends on another bunsetsu that appears after it in a sentence. In a sentence-reading experience using cutting sentences, Takanashi [11] suggested that for predicting the end of the sentence, Japanese readers use the number of modifiers whose modifying bunsetsu-phrases have not yet appeared in the middle of a sentence. We call the modifiers "unsolved-modifiers" [6]. Though the number of unsolved-modifiers may be regarded as a syntactic factor for predicting the end of an utterance, the modified bunsetsu sometimes does not appear until the very end in spontaneous utterances. Therefore, the number of unsolved-modifiers is not an absolute indicator of the end of an utterance in spontaneous speech. Hence, we utilized the difference in the number of unsolved-modifiers between a bunsetsu and the preceding bunsetsu as a syntactic index for predicting the end of an utterance.

Because the Chiba3Party corpus does not include the dependency structure of utterances, we analyzed the dependency structure from the transcription using CaboCha [12], a Japanese dependency parser to segment the bunsetsus and to manually correct errors. We also excluded short utterances that consist of one or two bunsetsus to deal with utterances with sufficient length for predicting the end of an utterance. We then calculated the number of unsolved-modifiers and the difference between a bunsetsu and the one preceding it.

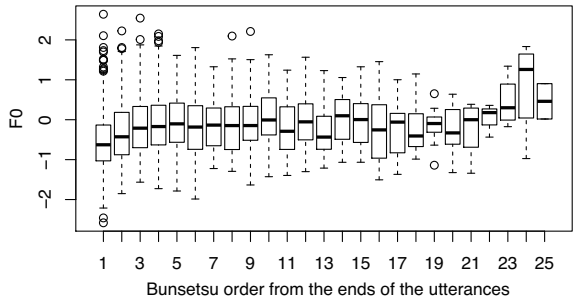
Table I shows the relationship between the difference of unsolved-modifiers and the reverse order of appearance of bunsetsu in the utterances. The number of the order means that the smaller the number, the closer the bunsetsu is to the end of the utterance. Order 1 indicates the final bunsetsu of the utterance. As shown in Table I, if the order from the end of the utterance is large, then the differences of unsolved-modifiers are barely negative values. If small, that is, the bunsetsu is close the end of the utterance, many of the differences of unsolved-modifiers become less than zero. In other words, the difference value can be used as a syntactic index for predicting the end of an utterance.

C. Prosodic features

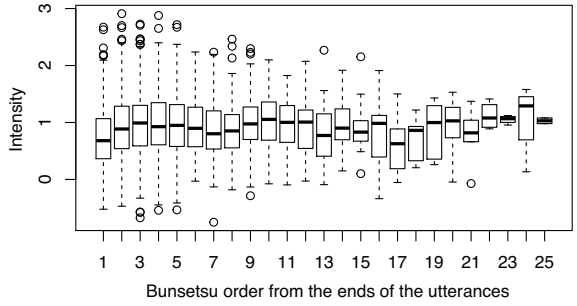
In our previous study, we observed prosodic changes at the final accentual phrase in spontaneous Japanese utterances [13]. The results showed that F0 attained its lowest value in the utterance and intensity decreased significantly. Moreover, mora duration lengthened; in other words, speech rate slowed down. We also focus on F0, intensity, and mora duration as prosodic features in this study and summarize the features not by the accentual phrase but the bunsetsu.

The average of the logarithmic F0s, intensity, and mora duration were calculated for each bunsetsu. To avoid the influence of gender and individual differences, we converted these prosodic features to z-values for each speaker.

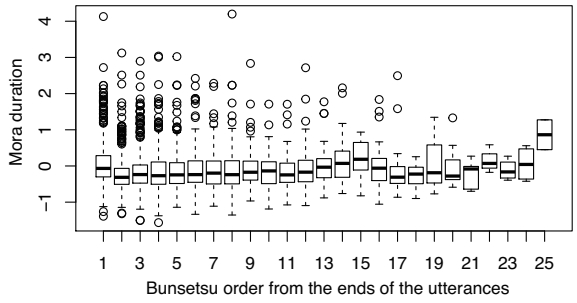
Figure 1 shows the relationship between the prosodic features and the reverse order of appearance of bunsetsu in the utterances. As shown in Figure 1(a), the F0s tend to decline at the final bunsetsu and the penultimate bunsetsu so that the bunsetsu order is 1 and 2. However, it is not a sharp drop. This result is because the subject of this study is the utterances without UFEs, and a sharp F0 drop readily occurs at UFEs [14]. As shown in Figure 1(b), the intensity slightly declines at the final bunsetsu, but



(a) F0



(b) Intensity



(c) Mora duration

Fig. 1: Relationship between the prosodic features and the reverse order of appearance of bunsetsu in the utterances.

it seems unclear for predicting the end of an utterance. Similarly, as shown in Figure 1(c), the mora duration tends to increase at the final bunsetsu, but it is not obvious.

Thus, it is uncertain whether a single prosodic feature is useful for predicting the end of an utterance. On the other hand, in our previous study, we found that the combination of syntactic and prosodic features is useful for end-of-utterance detection [6]. Considering this previous

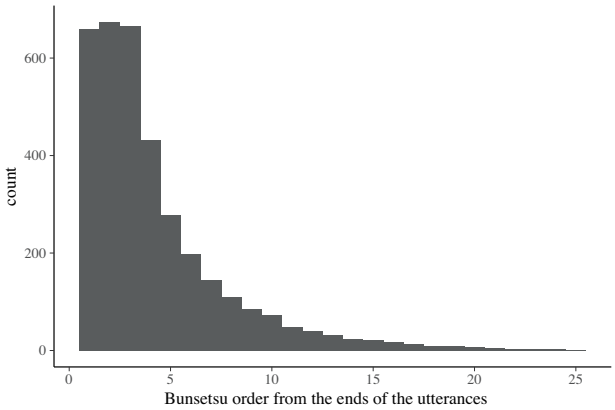


Fig. 2: The number of bunsetsus indexed in reverse order (from the ends of the utterances).

study, we investigated the effect of the combination of the features in predicting end-of-utterance.

III. A model for predicting the end-of-utterance

In this section, we construct a statistical model that estimates the position of a bunsetsu in an utterance to clarify the syntactic and prosodic features that are useful for predicting end-of-utterance.

A. Method

Figure 2 shows the distribution of the number of bunsetsus in the utterances in reverse order. Because of the utterances with more than three bunsetsus, the distribution peaks in the range of 1–3 and has a right-skewed curve. Therefore, we assumed that the distribution follows a log-normal distribution. We defined a model to estimate the bunsetsu order from the syntactic and prosodic features, which is described below.

$$
\begin{aligned}
y[n] &\sim \text{LogNormal}(\mu[n], \sigma), && (1)\\
\mu[n] &= a[S[n]] + b_1[S[n]] \times \Delta Mod[n] \\
&+ b_2[S[n]] \times F0[n] + b_3[S[n]] \times Int[n] \\
&+ b_4[S[n]] \times Dur[n], && (2)\\
a[k] &= a0 + a_k[k], && (3)\\
b_i[k] &= b0_i + b_{ik}[k], && (4)\\
a0 &\sim \text{Normal}(0, \sigma_{a0}), && (5)\\
a_k[k] &\sim \text{Normal}(0, \sigma_{ak}), && (6)\\
b0_i &\sim \text{Normal}(0, \sigma_{b0i}), && (7)\\
b_{ik}[k] &\sim \text{Normal}(0, \sigma_{bik}), && (8)
\end{aligned}
$$

where n is an index of bunsetsus ($n = 1, \ldots, 3561$); $y[n]$ s the position number of the bunsetsu from the end of the utterance, for example, 1 for the final bunsetsu of the utterance and 2 for the penultimate bunsetsu; k is the index of the speakers ($k = 1, \ldots, 36$) and $S[n]$ is the

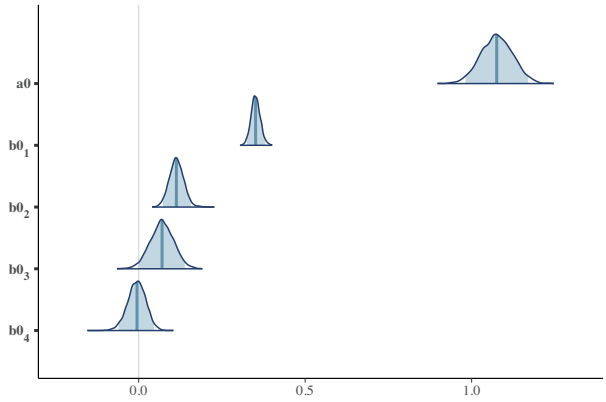


Fig. 3: Posterior distribution and confidence intervals estimated from MCMC samples for the parameters. The colored section in the distribution means 95% confidence interval.

speaker number k for the nth sample; i s the index of the parameters for syntactic and prosodic features, namely, 1 for the difference of unsolved-modifiers, 2 for F0, 3 for intensity, and 4 for mora duration.

We employed four features as explanatory variables: the difference in the number of unsolved-modifiers ($\Delta Mod[n]$), the means of F0s ($F0[n]$), intensity ($Int[n]$), and mora duration ($Dur[n]$), which are mentioned in Section II. σ, σ_{a0}, σ_{ak}, σ_{b0i}, and σ_{bik} were set to a non-informative prior. This Bayesian hierarchical model has speaker-independent parameters ($a0$ and b_{0i}) and speaker-dependent parameters ($a_k[k]$ and $b_{ik}[k]$). Accordingly, the model can reveal the effect of interindividual differences in the features. The parameters of the model were iteratively estimated using the Markov chain Monte Carlo (MCMC) method.

B. Results

Figure 3 shows posterior distributions and confidence intervals estimated from MCMC samples for the parameters. The 95% confidence interval in the posterior of $b0_4$ includes zero. This means that the speaker-independent coefficient for the mora duration is likely to be zero in the model, that is, mora duration does not contribute to predicting end-of-utterance. In other words, the parameters for the difference of unsolved-modifiers, F0, and intensity are significant in predicting end-of-utterance. Accordingly, the combination of syntactic and prosodic features is effective in predicting end-of-utterance.

However, considering the speaker-dependent coefficients, the significance of the parameters $a[k]$, $b_1[k]$, $b_2[k]$, $b_3[k]$ and $b_4[k]$ may differ from the result of the speaker-independent parameters. Table II shows the significant parameters for each speaker. The results indicate two things. First, the difference of unsolved-modifiers is significant for all speakers. In other words, the syntactic feature corresponds to the prediction of end-of-utterance. Second, the prosodic features used for the prediction vary from speaker to speaker. The methods of utilizing prosodic features are divided into the following four groups:

1) both F0 and intensity (4 speakers),
2) only F0 (24 speakers),
3) only intensity (1 speaker),
4) none of the prosodic features (10 speakers).

This means that there are no prosodic features that contribute to the prediction of end-of-utterance for all speakers.

C. Discussion

The above results suggest that listeners utilize different prosodic features of speakers' speech to predict the ends of utterances. Considering the various ways in which the utterance act can be performed, it may be natural that no authoritative feature exists for such prediction. Therefore, it is possible that listeners adaptively learn the useful features of speakers' speech in conversation.

On the other hand, there were ten speakers for whom the prosodic features were not practical. Do their utterances have no prosodic characteristics that can help predict end-of-utterance? We think that the prosodic features used in this study did not fit the model. For example, we should adopt not only the average of the features for a bunsetsu but also the difference from the preceding bunsetsu to introduce time-dependent changes of prosodic features.

In future research, we will improve the model and investigate other features to precisely explain the effects of syntactic and prosodic features in predicting end-of-utterance.

IV. Conclusions

To clarify the information necessary for predicting the ends of utterances in spontaneous speech, we investigated useful syntactic and prosodic features. As a syntactic feature that can help predict the end of Japanese utterance, we adopted the difference in the number of modifiers whose modifying bunsetsus do not appear in mid-utterance. As prosodic features, we adopted the average of F0, intensity, and mora duration for bunsetsus. We constructed a statistical model that estimates the position of a bunsetsu in an utterance from the features. The results suggest that the syntactic feature is always significant, whereas the prosodic features vary in usefulness according to the speakers. We conclude that the combinations of syntactic and prosodic features for each speaker are relevant in predicting the ends of utterances.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers 15K00390 and 18K11514.

TABLE II: Significant parameters for each speaker in the model. "*" denotes the significance at 95% confidence level.

	Speaker No.																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Intercept ($a[k]$)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ΔMod ($b_1[k]$)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
F0 ($b_2[k]$)	*	*	*	*	*	*		*		*	*		*				*	
Intensity ($b_3[k]$)		*	*															*
Mora duration ($b_4[k]$)																		

	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
Intercept ($a[k]$)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
ΔMod ($b_1[k]$)	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
F0 ($b_2[k]$)		*	*	*		*	*	*	*			*	*	*	*	*	*	
Intensity ($b_3[k]$)	*													*	*			
Mora duration ($b_4[k]$)																		

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. Schegloff, and S. A. Thompson, Eds. Cambridge University Press, pp. 134–184, 1996.

[3] J. B. Pierrehumbert and M. E. Beckman, *Japanese tone structure*, Cambridge: MIT Press, 1988.

[4] Y. Ishimoto and H. Koiso, "Utterance-final F0 changes in Japanese monologs and dialogs," *Proc. Oriental COCOSDA 2014*, pp. 255–260, 2014.

[5] Y. Ishimoto and M. Enomoto, "Experimental investigation of end-of-utterance perception by final lowering in spontaneous Japanese," *Proc. Oriental COCOSDA 2016*, pp. 205–209, 2016.

[6] Y. Ishimoto, T. Teraoka, and M. Enomoto, "End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous Japanese speech," *Proc. Interspeech2017*, pp. 1681–1685, 2017.

[7] 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, pp. 207–330, 2007.

[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," *Proc. LREC2010*, pp. 2103–2110, 2010.

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

[10] M. Enomoto, "The cognitive mechanism of the completion of turn-constructional units in Japanese conversation," *The Japanese Journal of Language in Society (in Japanese)*, Vol. 9, No. 2, pp. 17–29, 2007.

[11] K. Takanashi, "Elucidation of incremental prediction mechanism in human sentence processing," in *Sentences and utterances in time (in Japanese)*, S. Kushida, T. Sadanobu, and Y. Den, Eds. Hituji Shobo, pp. 159–202, 2007.

[12] T. Kudo and Y. Matsumoto, "Japanese dependency analysis using cascaded chunking," *Proceedings of the 6th Conference on Natural Language Learning 2002*, pp. 63–69, 2002.

[13] Y. Ishimoto, M. Enomoto, and H. Iida, "Projectability of transition-relevance places using prosodic features in Japanese spontaneous conversation," *Proc. Interspeech2011*, pp. 2061–2064, 2011.

[14] Y. Ishimoto, M. Enomoto, and H. Iida, "Prosodic changes pre-announcing a syntactic completion point in Japanese utterance," *Proc. Interspeech2013*, pp. 788–791, 2013.