INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

End-of-Utterance Prediction by Prosodic Features and Phrase-Dependency Structure in Spontaneous Japanese Speech

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

[1]Center for Corpus Development, National Institute for Japanese Language and Linguistics, Japan
[2]School of Media Science, Tokyo University of Technology, Japan

yishi@ninjal.ac.jp, teraokatkhr@stf.teu.ac.jp, menomoto@stf.teu.ac.jp

Abstract

This study is aimed at uncovering a way that participants in conversation predict end-of-utterance for spontaneous Japanese speech. In spontaneous everyday conversation, the participants must predict the ends of utterances of a speaker to perform smooth turn-taking without too much gap. We consider that they utilize not only syntactic factors but also prosodic factors for the end-of-utterance prediction because of the difficulty of prediction of a syntactic completion point in spontaneous Japanese. In previous studies, we found that prosodic features changed significantly in the final accentual phrase. However, it is not clear what prosodic features support the prediction. In this paper, we focused on dependency structure among bunsetsu-phrases as the syntactic factor, and investigated the relation between the phrase-dependency and prosodic features. The results showed that the average fundamental frequency and the average intensity for accentual phrases did not decline until the modified phrase appeared. Next, to predict the end of utterance from the syntactic and prosodic features, we constructed a generalized linear mixed model. The model provided higher accuracy than using the prosodic features only. These suggest the possibility that prosodic changes and phrase-dependency relations inform the hearer that the utterance is approaching its end.

Index Terms: turn-taking, prediction model, prosody, phrase dependency, utterance-final element

1. Introduction

Conversation is an important component of everyday human interaction, and a smooth conversation is one in which there is turn-taking with minimal gap or overlap in the speech of the participants in the conversation. One needs to instinctively predict the ends of utterances to realize smoothness in spontaneous conversations; however, the features of utterances that support prediction are not clear.

Sacks et al. [1] proposed a turn-taking system that employs a turn constructional unit (TCU) as an utterance unit in turn-taking. According to 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 possibly occurs at the TRP. In accordance with this framework, Ford and Thompson [2] suggest that syntactic, intonational, and pragmatic resources constitute TRPs.

For a syntactic factor of the TRP in Japanese, Tanaka [3] identified certain words frequently placed at the ends of utterances as utterance-final elements (UFEs). However, one cannot always identify end-of-utterance using the UFE because there are instances of utterances ending without the UFEs As prosodic factors, Pierrehumbert and Beckman [4] defined the unit that utterance is placed above intonation phrases (IPs) and accentual phrases (APs) in the prosodic hierarchy. The utterance unit is a region in which the fundamental frequencies (F0s) monotonously decline over time, which is called the F0 declination, and the F0s fall significantly at the end of the utterance unit, which is called the final lowering. According to them, the final lowering indicates the end of the utterance, but the final lowering rarely appears in dialog as can be seen in our previous work [5]. This implies that it is difficult to predict end of utterance by the mere presence or absence of the final lowering.

In this paper, we consider a combination of syntactic and prosodic features to predict the end of utterance in spontaneous speech. We focus on the dependency structure among bunsetsu-phrases as a syntactic factor and investigate the relationship between phrase dependency and prosodic features. We then construct a model for predicting end of utterance using the syntactic and prosodic features and show the availability of the combination for spontaneous Japanese utterance.

2. Features characterizing the end of utterance

In this section, we describe the features characterizing the end of utterance, which are utilized in later sections.

2.1. Prosodic features

As mentioned in the introduction, the final lowering is probably not the absolute indicator of the end of utterance, but it may be one of the indicators. In previous work [6], we observed prosodic changes at the final AP in spontaneous Japanese utterance: F0 attained its lowest value in the utterance, intensity decreased significantly, and the average mora duration lengthened. In addition, Maekawa [7] points out that the domain of the final lowering was the final AP in Japanese. Therefore, we consider F0, intensity, and mora duration of each AP as the prosodic features in this study.

2.2. Syntactic features

2.2.1. Utterance final elements

The 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 project the completion of a TCU [3]. 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 [8] demonstrates that the beginning of the TRP is when hearers recognize the UFE in perceptual experiments. Therefore, UFEs are a potent factor used by participants for predicting the end of utterance.

http://dx.doi.org/10.21437/Interspeech.2017-837

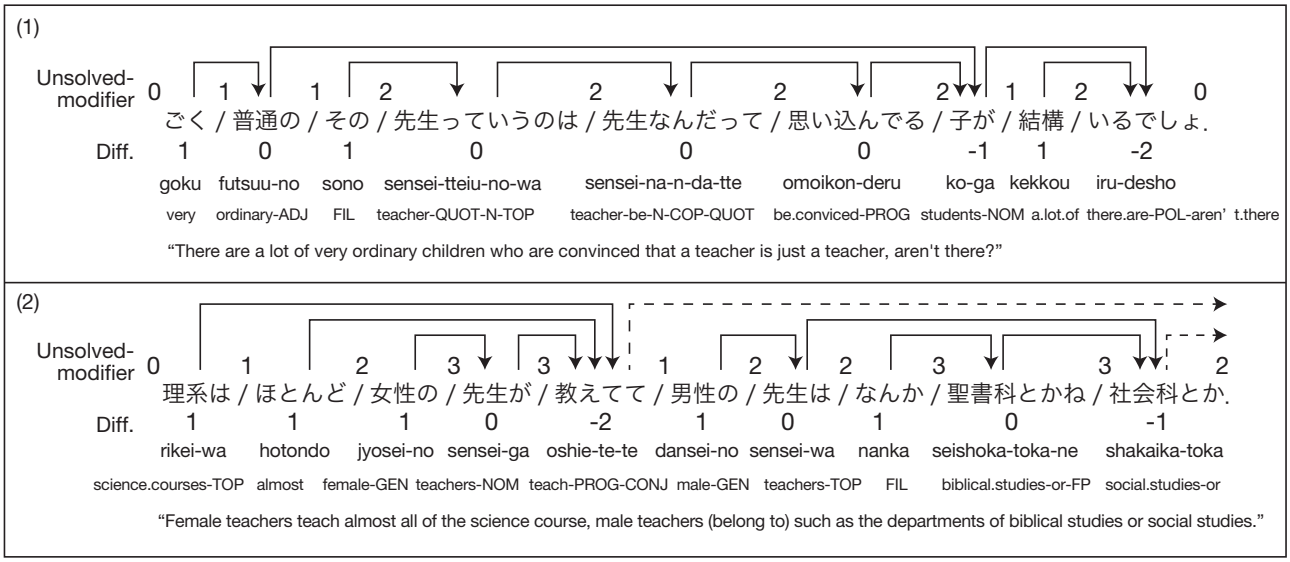


Figure 1: *Examples of dependency structure among bunsetsu-phrases. "/" denotes a boundary of bunsetsu-phrases. Sentence (1) describes an utterance in which there is no unsolved-modifier at the end, and sentence (2) describes an utterance in which an unsolved-modifier is present at the end of the sentence.*

2.2.2. Dependency structure among bunsetsu-phrases

A Japanese sentence consists of a sequence of phrasal units called "bunsetsu." A bunsetsu-phrase is a part of a phrase that cannot be divided further in Japanese language, and it depends on another bunsetsu to its right. Takanashi [9] suggested that readers use the number of modifiers whose modifying bunsetsu-phrases have not yet appeared in the middle of a sentence (hereinafter referred to as "unsolved-modifier") for predicting the end of the sentence, and he examined this statement by using cutting sentences. We adopt the number of unsolved-modifiers as one of syntactic factors for predicting the end of utterance.

Figure 1 illustrates the dependency structure among bunsetsu-phrases and the number of the unsolved-modifiers. In sentence (1), for example, at the point of the penultimate bunsetsu-phrase /kekkou/, its modified bunsetsu-phrase does not exist yet. Moreover, the bunsetsu-phrase modified by the preceding /koga/ has also not appeared yet. Therefore, the number of unsolved-modifiers is 2 toward the end of /kekkou/. Then, the final bunsetsu-phrase /irudesho/ that /koga/ and /kekkou/ modify appears, and the number of unsolved-modifiers is 0 by the end of the sentence.

For spontaneous utterances, however, the modified bunsetsu-phrase sometimes does not appear until the very end. In sentence (2), the last bunsetsu-phrase /shakaikatoka/ is a modifier but its modified bunsetsu-phrase does not exist at the end of the utterance, and the number of unsolved-modifiers is not 0 at the end of the sentence. In other words, the number of unsolved-modifiers is not the absolute indicator of the end-of-utterance in spontaneous speech. Hence, we regarded the difference in the number of unsolved-modifiers between a bunsetsu-phrase and the preceding bunsetsu-phrase as a syntactic index for predicting the end of utterance. This is because a decrease in the number of unsolved-modifiers implies finding the modified bunsetsu-phrase, and a negative value of the difference leads to the possibility of the end of the utterance.

Table 1: *Number of accentual phrases for the differential values of unsolved-modifiers.*

Diff. of Mod.	-3	-2	-1	0	1
APs	212	607	1482	5869	2097

3. Relationship between phrase-dependency and prosodic features

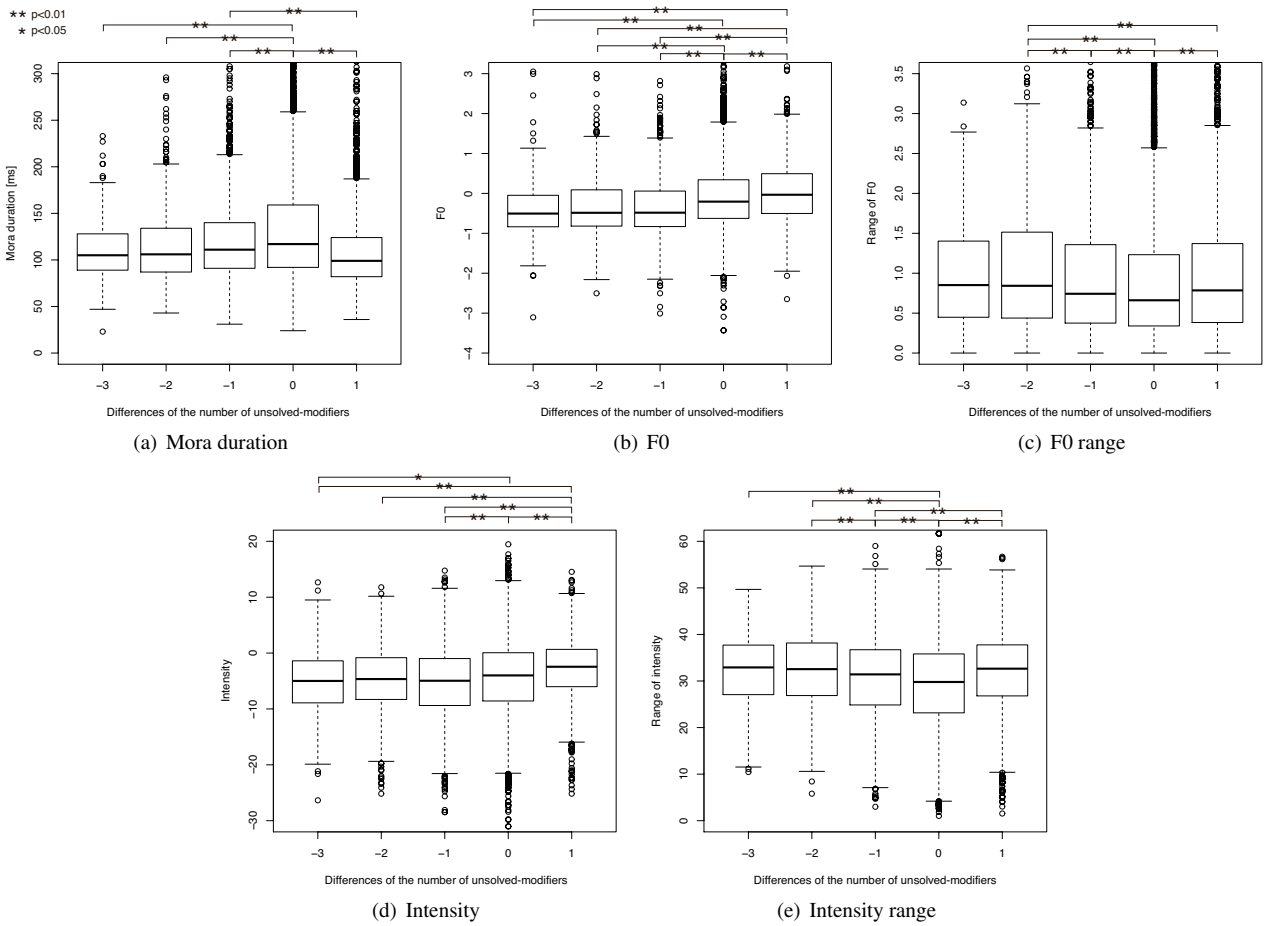
In this section, we investigate the relationship between the number of unsolved-modifiers and the prosodic features in order to examine the prosodic characteristics of phrase-dependency.

3.1. Data

Twelve dialogs from the Chiba three-party conversation corpus (Chiba3Party) [10], which are casual Japanese conversations on different themes by twelve groups of three people, were used.

We analyzed the dependency structure of the utterances using CaboCha [11], a Japanese dependency parser, and corrected it manually. We then calculated the number of unsolved-modifiers and the difference between a bunsetsu-phrase and its preceding one. Table 1 shows the number of APs for each value of the differences. The bunsetsu-phrases with differences less than -4 were excluded because of the small size of the data sets. Furthermore, bunsetsu-phrases with a rising intonation at the end of the AP were excluded because the rising intonation obviously indicates the end of the utterance. We also excluded fillers and backchannels.

To identify the prosodic features, we focused on the last AP in each bunsetsu-phrase, and extracted five prosodic features for each AP: Average mora duration, average F0, range of F0s, average intensity, and range of intensity. The F0s and intensity values were converted to z-values for each speaker to avoid influences of gender and individual differences.



(a) Mora duration (b) F0 (c) F0 range

(d) Intensity (e) Intensity range

Figure 2: *Prosodic features for the difference in the number of unsolved-modifiers between a bunsetsu-phrase and its preceding one.*

3.2. Results

Figure 2 shows the relationship between the differential values of unsolved-modifiers and prosodic features.

As shown in Figure 2(a), speech rate is slower when the differential value is 0, namely, when a bunsetsu-phrase depends on its adjacent bunsetsu-phrase. As shown in Figures 2(b), (c), the F0s decline when the difference decreases, namely, when the modified bunsetsu-phrase appears. This suggests that F0 declination projects the end of utterance. In comparing Figures 2(b), (c) and (d), (e), it can be seen that intensity has the same tendency as the F0 characteristics, that is, the prosodic features and the differential values of unsolved-modifiers are correlated to the progress of an utterance. From this, we infer that prosodic features change in conformance with the dependency structure, and their combination may be useful in predicting the end of utterance.

4. An end-of-utterance prediction model

In this section, we construct an end-of-utterance prediction model using the syntactic and prosodic features.

4.1. Method

We performed a generalized linear mixed-model analysis to estimate whether the final AP in an utterance signifies the end of utterance. We used the logistic regression model with random intercepts for each speaker. The number of target APs was 10,267 from the Chiba3Party corpus, including 3,505 APs corresponding to end of utterance. We employed seven features as explanatory variables: the five prosodic features mentioned in Section 3.1, the difference in the number of unsolved-modifiers, and the presence of UFEs in the bunsetsu-phrases mentioned in Section 2.2. The prosodic features were normalized to zero-mean and unit standard deviation for the model. The UFE was treated as a binary variable: 0 for absence and 1 for presence in the AP.

4.2. Results and discussion

Table 2 shows the estimated fixed effect coefficients after the AIC-based model selection. The model using all the seven features was chosen by the model selection procedure. The result indicates that both prosodic and syntactic features are significant for predicting the end of utterance. The result regarding prosodic features supports the previous study [6] that F0s and intensity of APs decline gradually in utterance and drop significantly just before the end of utterance, and that mora duration lengthens around the end of the utterance. With regard to syntactic features, this result suggests that dependency structure is useful for predicting the end of utterance in the same way as end of sentence is predicted by cognition of dependency, as reported by Takanashi [9].

Table 2: *Estimated fixed-effect coefficients in the model.*

	Estimate	Std. Error	z value		
(Intercept)	-1.553	0.110	-14.09	$p<0.001$	***
Mora Duration	0.495	0.030	16.36	$p<0.001$	***
F0 Mean	0.086	0.034	2.55	$p=0.011$	*
F0 Range	0.131	0.032	4.15	$p<0.001$	***
Intensity Mean	-0.219	0.036	-6.16	$p<0.001$	***
Intensity Range	-0.295	0.035	-8.36	$p<0.001$	***
Diff. of Mod.	-1.514	0.049	-31.13	$p<0.001$	***
UFE	4.747	0.177	26.87	$p<0.001$	***

Table 3: *Results of predicting the final AP of utterance in the model using prosodic and syntactic features.*

		Estimated		
		Non-final	Final AP	Total
Observed	Non-final	6457	305	6762
	Final AP	1158	2347	3505

Table 4: *Results of predicting the final AP of utterance in the model using only prosodic features.*

		Estimated		
		Non-final	Final AP	Total
Observed	Non-final	6141	621	6762
	Final AP	2426	1079	3505

Table 3 shows a confusion matrix that summarizes the results of 10-fold cross validation. For this model, the F-measure of 0.762 (precision=0.885, recall=0.670) was obtained. To compare models, Table 4 shows the confusion matrix for a model using only prosodic features. The F-measure of the prosodic-features-only model was found to be 0.423 (precision=0.635, recall=0.317). From this, it is obvious that the proposed model achieves better performance than the prosodic-only model, and that syntactic features are effective for prediction.

Nearly 90 percent of the bunsetsu-phrases predicted as the final AP by the proposed model were certainly the final bunsetsu-phrases in the utterances. There were, however, many false-positive results, nearly 30 percent of the total of the final APs. The cause for this can be attributed to the difficulty in short utterances comprising very few bunsetsu-phrases. In case of an utterance with few bunsetsu-phrases, the number of unsolved-modifiers cannot provide information about the possibility of the end of utterance. In such a case, prosodic features must compensate for the lack of syntactic information. We believe that prosodic features related to predicting the end of utterance are insufficient in the proposed model.

For example, the F0 declination means that F0 at an AP becomes lower than that at the preceding AP, but this characteristic is not introduced into the proposed model. To introduce this, we need to adopt some features such as the difference in the F0s between the adjacent APs, the slope of the F0 contour in an utterance, and so on. It is also necessary to investigate and improve other features and models.

5. Conclusions

To discover a method by which participants in a conversation predict the end of utterance from spontaneous speech, we investigated the prosodic and syntactic features leading to the end of utterance. Considering syntactic features to relate to the possibility of end of utterance, we introduced the difference in the number of modifiers whose modifying bunsetsu-phrases do not appear in mid-utterance. The results of investigating the relationship between the features showed that prosodic features change in conformance with the dependency structure among bunsetsu-phrases. This suggests that a combination of prosodic and syntactic features is useful for end-of-utterance prediction. Next, the generalized linear mixed model using prosodic and syntactic features was constructed for the prediction. The model achieved high accuracy in spontaneous Japanese utterances. We conclude that participants use the phrase-dependency structure in addition to prosodic features to predict the end of utterance.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 15K00390.

7. 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, 1996, pp. 134–184.

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

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

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

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

[7] K. Maekawa, "Domain of final lowering in spontaneous Japanese," in *The Journal of the Acoustical Society of America*, vol. 135, 2014, p. 2194.

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

[9] 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. Tokyo: Hituji Shobo, 2007, pp. 159–202.

[10] 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.

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