Prosodic Changes Pre-announcing a Syntactic Completion Point in Japanese Utterance

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

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

yishi@ninjal.ac.jp, {menomoto,iida}@stf.teu.ac.jp

10.21437/Interspeech.2013-228

Abstract

In this paper we aim to clarify that participants of conversation can predict whether an utterance includes grammatical terminal elements, which have been referred to as the utterance-final elements (UFEs) in Japanese. We carried out perceptual experiments with Japanese utterances missing a part close to their end. The results showed that subjects distinguished whether an UFE follows at the verb before the appearance of the UFE, and they noticed that the end of the utterance arrives sooner or has already arrived even if the last mora is missing when the utterance includes the UFEs. Then, we analyzed the prosodic differences between the utterances with/without the UFE. The results presented the following information. The F0 declines gradually toward the end-of-utterance, the final lowering of the F0s remarkably occurs at the UFE, the power falls at the verb if the utterance does not have the UFE, and the power falls at the UFE if the utterance has the UFEs. That is, the F0 declination towards the end-of-utterance and the power falling at the verb are pre-announcing the syntactic completion point in the utterances to the hearers.

Index Terms: turn-taking, prosody, utterance-final element, perceptual experiment, end-of-utterance

1. Introduction

We can maintain smooth transfers from one speaker to another without gaps in spontaneous conversations. This means that we can somehow predict the ends of utterances. Sacks et al. [1] proposed a turn-taking system that uses 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].

Koiso et al. [3] investigated the 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 the duration and fundamental frequency (F0) contour patterns at the final mora of inter-pausal units depended on whether or not the speaker changed. However, in spontaneous conversations, we do not distinguish the beginning of a TRP from the final mora of the inter-pausal units, because the beginning of the final mora is too late for the speech planning of the next speaker. That is, acoustic features are needed prior to the final mora for prediction of the TRP.

Tanaka [4] has identified certain words indicating the beginning of the TRP as utterance-final elements (UFEs) 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 put to a Japanese utterance at the end of which there appears a conjunctive particle or an inflection form of a verb succeeded by a subordinate clause avoiding the main clause, and act as a syntactic factor and project the completion of a TCU. Enomoto [5] demonstrated that the beginning of the TRP is when hearers recognize the UFE by perceptual experiments. It is possible that the UFE includes not only a syntactic cue of the TRP but also an acoustic cue. However, we cannot always use the UFE to find the TRP, because utterances without such UFEs do exist. Tanaka observed that rising- or falling-intonation, stressing, and sentence-final morae lengthening occur for the *iikiri* (truncated) form characterized by a systematic absence of UFEs, and has suggested that these features could be used instead of the UFEs to detect the TRP.

We previously investigated the prosodic features for predicting the TRP in spontaneous Japanese conversation [6]. In our results, the F0 at the beginning of the final accentual phrase (AP) with the UFEs was lower than that at the beginning of the preceding AP. The power at the beginning of the final AP without UFEs was lower than that at the beginning of the preceding AP. However, there was a question about whether the hearer can recognize the differences in the prosodic features found in the previous work.

We can confirm by conducting two perceptual experiments that hearers have the ability to predict the end-of-utterance using the prosodic features. The experiments are performed by taking the UFEs used as syntactic features into consideration. Then, we investigate which prosodic feature can be used for predicting the UFEs, which become the syntactic completion point, or the end-of-utterance under the syntactic classifications.

2. Perceptual experiments for predicting end-of-utterance

We wonder if hearers can estimate the ends of the utterances even without features just before the end-of-utterance. For this section, we investigate the influence on cognition of an end-of-utterance for hearers by conducting perceptual experiments using utterances missing a part close to their end.

2.1. Experiment 1: Comparison between non-UFE and eliminated-UFE

2.1.1. Method

In this experiment, the subjects listen to utterances without the UFE. As shown in Figure 1, there are two kinds of stimuli:

1-A utterance that does not originally include the UFE, and
1-B utterance eliminated the UFE.

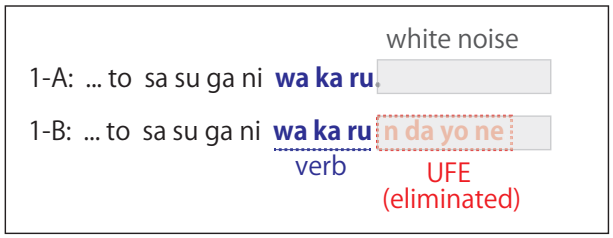


Figure 1: *Examples of stimuli in Experiment 1.*

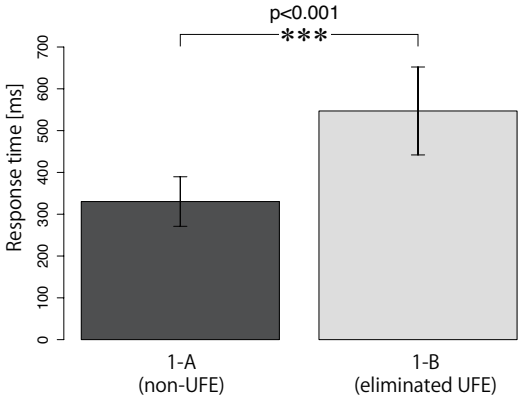


Figure 2: *Response time from end of speech in Experiment 1.*

White noise is added to the stimuli, which the SNR is about 30 dB, from the end of the speech sound onward. The subjects are randomly presented the stimuli, and rapidly push a button when they perceive/predict the end-of-utterance. The response times from the end of speech, which is the beginning of the white noise, are measured.

To create the stimuli, we selected the utterances with the predicates consisting of verbs expressed by base form (non-conjugated). In cases where the verb is a base form, the subjects cannot morphologically distinguish Stimulus 1-A from Stimulus 1-B. If the response time to Stimulus 1-A and 1-B differ, this denotes that the acoustic features are used for the perception/prediction of the end-of-utterance.

Twelve dialogs from the Chiba three-party conversation corpus [7], which are casual Japanese conversations according to some themes by twelve groups of three people, were used for this study. We substituted long utterance units [8] with boundaries at which turn-taking could occur for the TCUs, due to difficulty in identifying the TCU. The utterances with strong emotions and intentions were removed from the stimuli.

The listening experiments were performed by seven male and seven female subjects. The stimuli comprised 15 utterances for Stimuli 1-A and 25 utterances for Stimuli 1-B.

2.1.2. *Results and discussion*

Figure 2 shows the averages and standard deviations of the response times from the end of the speech sound for the stimuli. The result of the two-tailed t-test showed a significant difference ($t(32)=-8.54$, $p<0.01$). The result indicates that the response of Stimulus 1-B is obviously late and the subjects responded after completely passing the end of speech. This means that when the hearers recognize the verb before the appearance of the UFE, they must know that the utterance has the following UFE. That

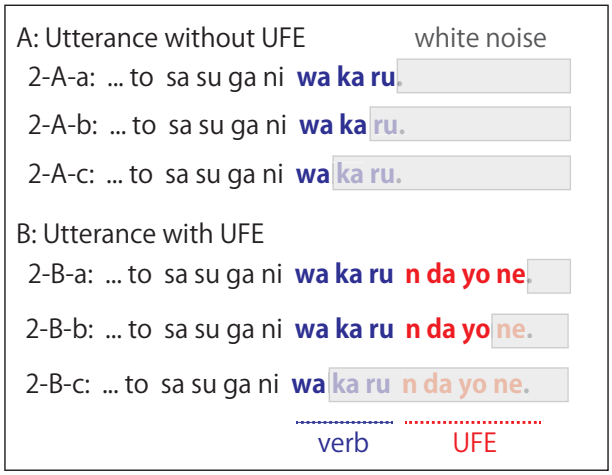


Figure 3: *Examples of stimuli in Experiment 2.*

is, the hearers can distinguish whether there is the following UFE at the point of the verb, and therefore, may adjust their timing to start speaking as the next speaker.

2.2. Experiment 2: Comparison between available features from nearby end of the utterance

2.2.1. *Method*

In this experiment, we examine the part of the utterance that hearers use to predict the end-of-utterance. The subjects randomly listen to the following utterances missing a part of each one close to the end.

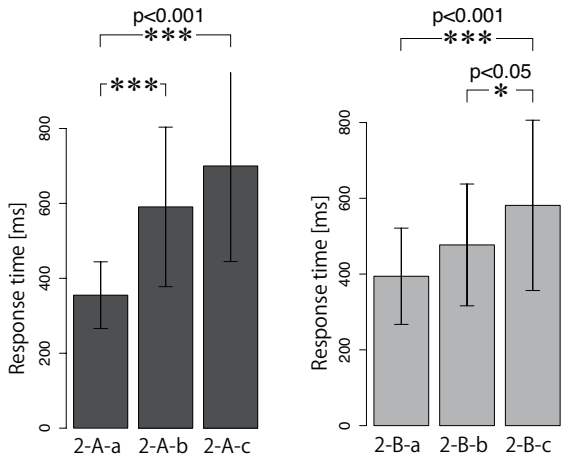
2-X-a Utterance without elimination (original)
2-X-b Utterance with last mora missing
2-X-c Utterance missing after first mora of verb,

where the "X" indicates an index for the two kinds of utterances in which "A" is an utterance without the UFE and "B" is one with it, as shown in Figure 3. White noise was added to the stimuli instead of the eliminated speech. The stimuli were from the same utterances used in Experiment 1. The subjects, which were the same as in Experiment 1, pushed the button at the moment they felt the end-of-utterance, and then the response time from the end of speech were measured.

2.2.2. *Results and discussion*

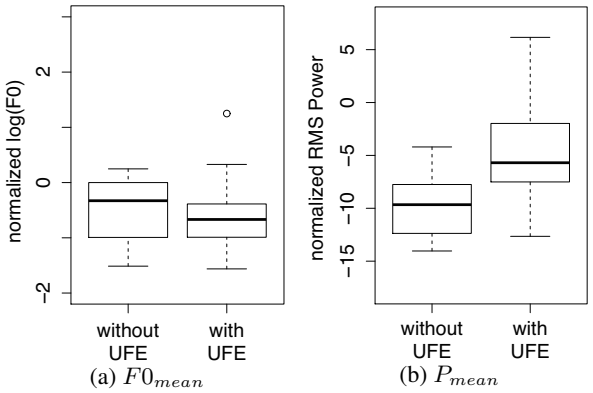
Figure 4 shows the averages and standard deviations of the response times from the end of the speech sound for the stimuli. The two-way ANOVA indicated a significant interaction between the kinds of elimination and with/without UFE ($F(2,238)=5.32$, $p<0.001$). The simple main effect of with/without UFE was significant at 5% for Stimuli 2-X-b. The simple main effect of the kinds of elimination was significant at 1% for both with/without UFE. Moreover, the results from multiple comparisons showed that the response time for stimulus 2-A-a was significantly short compared to the other two without UFE, and the response time for stimulus 2-B-c was significantly long compared to the other two with UFE.

This means that, for cases without UFE as shown in Figure 4(a), the hearer cannot predict the end-of-utterance and respond late when hearers cannot catch the last mora. However, with UFE as shown in Figure 4(b), they respond fast and can predict the end-of-utterance even if they cannot catch the last



(a) without UFE (b) with UFE

Figure 4: *Response time from end of speech in Experiment 2.*



(a) $F0_{mean}$ (b) P_{mean}

Figure 5: Prosodic features for verbs in utterances with/without UFE.

mora. That is, they feel that the end of the utterance arrives soon or has arrived already even if the last mora is missing by using the UFE. On the other hand, they cannot predict the end-of-utterance and delay responding when they listen only for the first mora of the verb. The whole of the verb must contain competent information for prediction of the end of the utterance.

3. Prosodic features in utterances with/without UFE

3.1. Verbs with/without UFE

3.1.1. Method

In Experiment 1, it was made clear that the hearers can predict whether the utterance has the UFE or not before it appears. This section shows the prosodic difference between the verbs with and without the following UFEs. If there is the difference the hearers can use it as an acoustic cue for predicting the UFE.

The logarithmic F0s were extracted from the utterances used as the stimuli in Experiment 1, and were converted to z-score using averages and standard deviations for each speaker. Then, the mean values of the F0 ($F0_{mean}$) were calculated for the verbs in the utterances. The RMS power (P_{mean}) of the verbs was also calculated. The power was normalized for each speaker.

3.1.2. Results and discussion

Figure 5(a) shows $F0_{mean}$ of the verbs with and without the UFEs. The result of the t-test was not significant (t(16.22)=1.45, p=0.166). This is an interesting result considering a well-known phenomenon that the F0s fall significantly at the end of the utterance, which is called the final lowering [9]. Because the verb without the UFEs is the end of the utterance, the F0 should became the lowest in the utterance by the final lowering. Meanwhile, the F0 in the verbs with the following UFEs is not expected to be the lowest because the verbs are situated before the ends of the utterances. In spite of this, the difference of the F0s in the verbs with and without the UFEs was not observed. There is a possibility that the final lowering relates to the appearance of the UFEs.

Figure 5(b) shows P_{mean} of the verbs with and without the UFEs. The result of the t-test indicated a significant difference between with and without the UFEs (t(24.32)=-4.07, p<0.001). This result is expected from the previous observation, because the power decreases significantly in the final AP [6]. We consider that the power of the verbs with the following UFEs is on the way to drop to the ends, and therefore the hearers cannot feel the end-of-utterance from the verbs.

3.2. Prosodic changes in utterances with/without UFE

3.2.1. Method

As shown in the previous section, the final lowering may not occur in the utterances without the UFEs. In previous works [6, 9, 10], the following prosodic changes in utterance have been observed:

- F0s decline toward the end of the utterance (F0 declination).
- F0s fall significantly at the end of the utterance, such as the final AP (final lowering).
- Power drops significantly in the final AP.

We wonder if these prosodic changes occur regardless if the UFE is present or not. In this section, we analyze the prosodic differences between the utterances with/without the UFE, and examine whether or not the prosodic features project the following UFE.

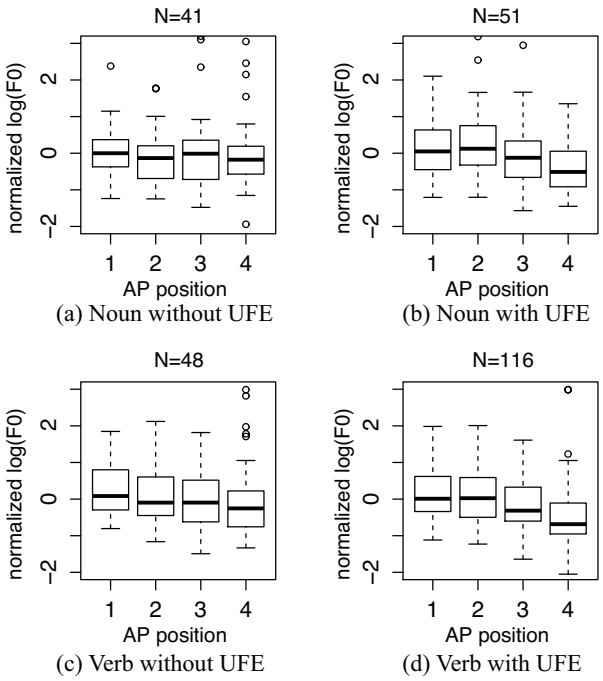
In Japanese, a noun can be located at the end of a sentence as a predicate except a verb, and the UFEs can follow the noun. The utterances from the Chiba three-party conversation corpus were classified into four groups as follows:

- whether the utterance has the UFE or not, and
- if the predicate of the utterance sentence is the verb or the noun.

Then, $F0_{mean}$ and P_{mean} were extracted for each AP. The changes in the $F0_{mean}$ and the P_{mean} between the AP positions were observed under the above classification.

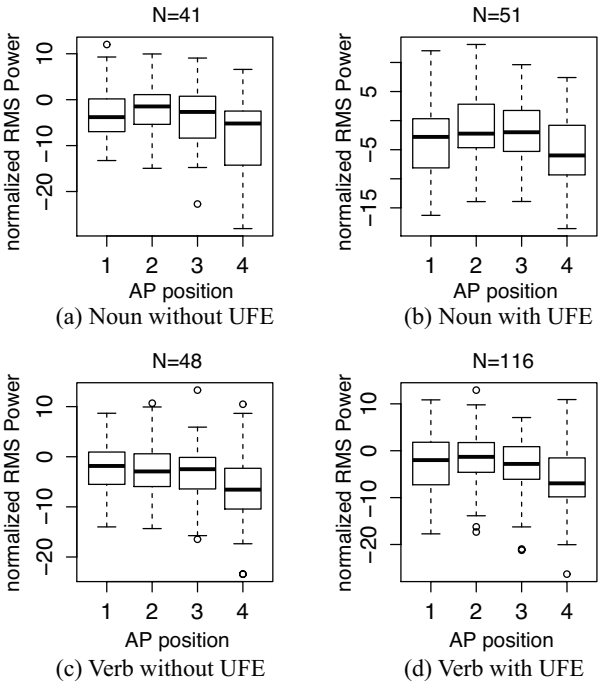
3.2.2. Results and discussion

Figure 6 shows the $F0_{mean}$ for the utterances with four APs, where N is number of utterances. The results for the number of APs excluding those of four APs also showed a similar tendency. As shown in Figure 6(b)(d), for the utterances with UFE, the F0 declines gradually toward the end-of-utterance, and the F0s rapidly drop at the final AP. These are the F0 declination and the final lowering as mentioned above. As shown in Figure 6(a), for the utterances without UFE when the predicate is



(a) Noun without UFE (b) Noun with UFE
(c) Verb without UFE (d) Verb with UFE

Figure 6: $F0_{mean}$ *for each AP position.*



(a) Noun without UFE (b) Noun with UFE
(c) Verb without UFE (d) Verb with UFE

Figure 7: P_{mean} *for each AP position. N: number of utterances.*

a noun, F0 declination is observed. However, the final lowering is not clearly seen. As shown in Figure 6(c), when the predicate is a verb, the decline in F0s is small compared to the utterance with UFE. That is, the utterance with the UFE has the final lowering informing the end-of-utterance, conversely, the utterance without the UFE shows special F0 changes. In other words, the final lowering principally occurs at the UFEs.

Figure 7 shows the P_{mean} for the utterances with four APs. The results for the other number of APs also showed a similar tendency. As shown in Figure 7(b)(d), for the utterances with UFE, there is a large decrease in power at the final AP, which is similar to the previous work as mentioned above. As shown in Figure 7(c), for the utterances without UFE when the predicate is a verb, the power also drops in the final AP. However, as shown in Figure 7(a), when the predicate is a noun, the power varies widely at the final AP, and the decline is small. That is, the decline in the power occurs just before the end of the utterance, and if the utterance does not have the UFE the power falls at the verb, and if the utterance has the UFE the power falls at the UFE.

Tanaka [4] classified the utterance that does not have the UFE and the predicate is the noun into the *iikiri* form and indicated that the last syllable of the utterance was emphasized and uttered with a clearly rising or falling final intonation. If many of the utterances in which the predicate is the noun are emphasized in that way, the results in this paper are consistent with the observations made by Tanaka.

In summary, it is possible that hearers can predict whether the utterance has the UFE by perceiving the power at the verb, and can feel the end-of-utterance by dropping of the F0s. The F0 declination also indicates the end of the utterance with the UFEs, and can be used for pre-announcing the end-of-utterances. The final lowering, which has been regarded as a marker of the end-of-utterance, mainly occurs at the UFEs, and it may be incompetent for projecting the end in the utterance without the UFE.

4. Conclusions

We conducted perceptual experiments to investigate the influence on cognition of the end-of-utterance for hearers by using utterances in spontaneous Japanese missing a part close to the end. The results showed that hearers can distinguish whether there is a following UFE at the point of the verb, and they feel that the end of the utterance arrives soon or has arrived already even if the last mora is missing when the utterance includes the UFEs. In addition, we analyzed the prosodic differences between the utterances with/without the UFE. The results showed that the F0 declines gradually toward the end-of-utterance, the final lowering of F0s remarkably occurs at the UFE, the power falls at the verb if the utterance does not have the UFE, and the power falls at the UFE if the utterance includes the UFE. We believe that hearers predict the syntactic completion point in an utterance by perceiving the F0 declination toward the end-of-utterance and the power falling at the verb.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24700109.

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

[3] 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.

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

[5] 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.

[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] 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.

[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.

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

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