INFLUENCE OF PROSODY, CONTEXT, AND WORD ORDER IN THE IDENTIFICATION OF FOCUS IN JAPANESE DIALOGUE

Tatsuya Kitamura Kayo Itoh Toshihiko Itoh Shigeyoshi Kitazawa

Faculty of Information, Shizuoka University
3-5-1 Johoku, Hamamatsu, Shizuoka 432-8011, Japan
{kitamura, cs8005, t-itoh, kitazawa} @cs.inf.shizuoka.ac.jp

ABSTRACT

This paper studies the influence of prosodic features, context, and word order on the identification of focused clauses in Japanese dialogue, using a psychoacoustic experiment. In the experiment, question and answer speech was used as stimuli. The questions were to create two different contexts in the stimuli, and the answers had focal prominence at different clauses and had different word orders. The experimental results indicate that (1) prosodic characteristics are more significant for focus identification, (2) context has some effect on identification, and (3) it is probable that the word order has some effect on identification.

1. INTRODUCTION

The aim of this paper is to investigate the relation of prosodic, semantic, and syntactic clues in the identification of focal prominence in Japanese dialogue.

Focal prominence appears when a speaker focuses on a part of a sentence, and various prosodic changes are observed on and around the focused part. The speaker uses the focal prominence to make his or her aim known clearly; on the other hand, the listener uses it to understand the speaker's intention correctly.

The addition of focal prominence would make synthesized speech more natural and easier to understand. Many studies have thus analyzed the prosodic features of focal prominence and have proposed rules to emulate focal prominence in speech synthesis[2, 3, 4, 5, 6].

While focal prominence has been studied mainly from the viewpoint of prosody in the speech science area as mentioned above, the influence of semantic and syntactic clues in focal identification have not been reported. However, these could assist in focal identification. For example, word order in a sentence can be changed freely in Japanese in contrast to English, and in spontaneous speech the important word is frequently placed at the beginning of the sentence. It is therefore important to clarify whether these factors should influence the synthesis of answers in a spoken dialogue system.

The study described here used question and answer speech as stimuli. The questions were to create two different contexts in the stimuli, and the answers had focal prominence in different clauses and had different word orders. The influences of context and word order were investigated by a psychoacoustic experiment.

2. SPEECH DATA

The speech data consisted of Japanese questions and answers about schedules, and had different contexts and word orders. The following sentences were spoken by one male and one female native Japanese speaker. Spaces in the sentences below signify the boundaries of clauses (*bunsetsu* in Japanese).

Q1. *asjita dochirani ikarerundesuka ?*
(Where are you going tomorrow ?)
Q2. *asjita nanininotte ikarerundesuka ?*
(How are you travelling tomorrow ?)

A1. *oHsakae jidoHshade ikimasu.*
(I am going to Osaka by car.)
A2. *jidoHshade oHsakae ikimasu.*
(I am going to Osaka by car.)

These sentences were recorded in the following three contexts.

Context 1. **A1** and **A2** were spoken independently without any focus. (These speech data are denoted in italics below as $A1_{no}$ and $A2_{no}$ respectively.)

Context 2. **A1** and **A2** were spoken as the answers to **Q1**. The clause "*oHsakae*" was focused on, thus there is a focal prominence at this clause (These speech data are denoted as $A1_{1st}$ and $A2_{2nd}$ respectively. The cardinal numbers show the order of the focused clauses in the sentence.)

Context 3. **A1** and **A2** were spoken as the answers to **Q2**. Thus, the clause "*jidoHshade*" was focused on. (These speech data are denoted as $A1_{2nd}$ and $A2_{1st}$ respectively.)

The speech data of the sentence **Q1** and **Q2** are also denoted in italics below as *Q1* and *Q2* respectively.

Contexts 2 and 3 were recorded in the face-to-face conversation style by the two speakers, whilst Context 1 was recorded as monologues. The speakers played the roles of questioner and answerer in rotation. They were not instructed on which clause to focus on or how to pronounce focal prominence in the recording.

The speech data were recorded using two headsets (SENNHEISER HMD-410) and a DAT recorder (SONY DTC-ZA5 ES). The the data were then saved in a personal computer at a sampling rate 16 kHz with 16-bit resolution.

3. COMPARISONS OF PHYSICAL CHARACTERISTICS

To clarify the prosodic features of focal prominence, the physical characteristics of the speech data of answers spoken in the above three conditions were compared.

3.1. F_0 Contours and Durations

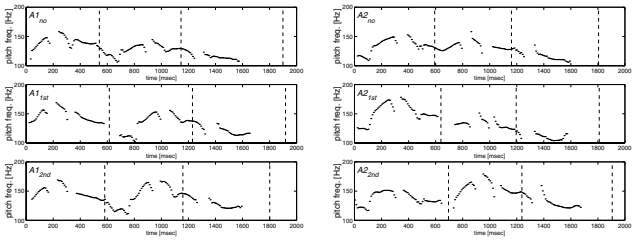
The fundamental frequency (F_0) contours of the speech data of the male speaker are shown in Figure 1 and those of the female speaker are shown in Figure 2. The package Praat[1] was adopted to extract F_0 and the extraction error was corrected manually. Vertical dashed lines in the figures represent the boundaries of clause.

On and around the focused clause the following phenomena are observed. First, the maximum F_0 of the clause is higher and the F_0 fluctuation is larger. Next, F_0 of the clause following the focused one is lower. Finally, in the contours of the speech data with the focal prominence in the second clause ($A1_{2nd}$ and $A2_{2nd}$) of the female speaker, a rapid rise of the fundamental frequency at the post positional particle, corresponding to a preposition in English, is observed.

There is no significant feature of the focused clause with respect to duration.

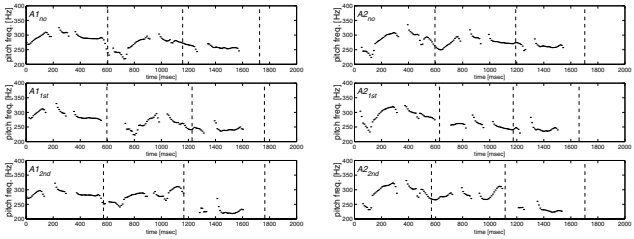
3.2. Power Contours

The power contours of the speech data of the male speaker are shown in Figure 3 and those of the female speaker are shown in Figure 4. When the first clause is focused on ($A1_{1st}$ and $A2_{1st}$), there is a significant fall in power after the focused clause.



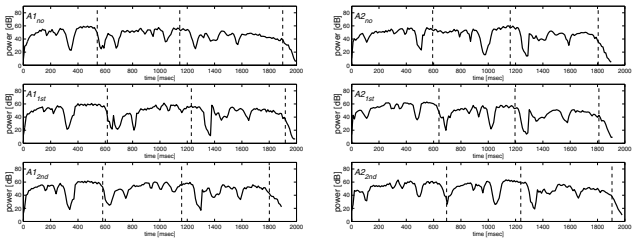
(top) $A1_{no}$, (middle) $A1_{1st}$, (bottom) $A1_{2nd}$ (top) $A2_{no}$, (middle) $A2_{1st}$, (bottom) $A2_{2nd}$

Fig. 1. F_0 contours of the speech data of the male speaker.



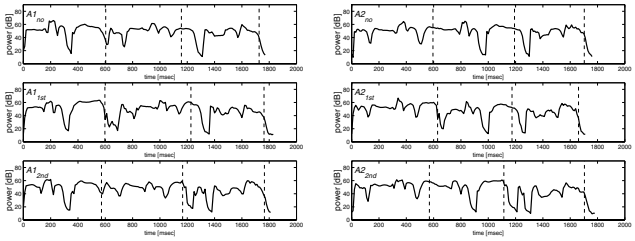
(top) $A1_{no}$, (middle) $A1_{1st}$, (bottom) $A1_{2nd}$ (top) $A2_{no}$, (middle) $A2_{1st}$, (bottom) $A2_{2nd}$

Fig. 2. F_0 contours of the speech data of the female speaker.



(top) $A1_{no}$, (middle) $A1_{1st}$, (bottom) $A1_{2nd}$ (top) $A2_{no}$, (middle) $A2_{1st}$, (bottom) $A2_{2nd}$

Fig. 3. Power contours of the speech data of the male speaker.



(top) $A1_{no}$, (middle) $A1_{1st}$, (bottom) $A1_{2nd}$ (top) $A2_{no}$, (middle) $A2_{1st}$, (bottom) $A2_{2nd}$

Fig. 4. Power contours of the speech data of the female speaker.

4. EXPERIMENT

An experiment was carried out to clarify the effects of prosodic features, context, and word order in identifying the focused clause.

4.1. Method

4.1.1. Stimulus Set 1

The stimuli were dialogue-type speech data that were full combinations of the speech data of the questions ($Q1$ and $Q2$), and the answers ($A1_{no}$, $A2_{no}$, $A1_{1st}$, $A1_{2nd}$, $A2_{1st}$, and $A2_{2nd}$). The speakers of the question and answer were different in each stimulus. The number of stimulus types is 24 (2 speakers $\times$ 2 questions $\times$ 6 answers = 24).

4.1.2. Stimulus Set 2

The speech data $A1_{no}$, $A2_{no}$, $A1_{1st}$, $A1_{2nd}$, $A2_{1st}$, and $A2_{2nd}$ were used independently as stimuli. The number of stimulus type is 12 (2 speakers $\times$ 6 answers = 12).

4.1.3. Subjects

Fifteen subjects (14 males and one female) participated in this experiment. All were native speakers of Japanese and none had a known hearing impairment.

4.1.4. Procedure

The experiments with stimulus sets 1 and 2 were carried out separately. Stimulus set 1 was used first.

The stimuli were low-pass filtered with a cutoff frequency of 8 kHz, and presented through binaural earphones (STAX λ Nova Signature) at a comfortable loudness level. Each stimulus was presented to the subjects three times randomly. The task was to identify focused clauses, and when the subjects judged a stimuli had no focused clause they responded with "no focus". The subjects were allowed to listen to each stimulus twice.

4.2. Results and Discussions

Focus identification for $A1_{1st}$ in stimulus set 1 was dispersed for several subjects. Detailed analysis showed that $A1_{1st}$ with the male speaker had indistinct focus for the subjects. Thus, we will analyze only results with respect to the sentence **A2**.

Figure 5 shows the focus identification rates of stimulus set 1 pairing *Q1* with $A2_{no}$, $A2_{1st}$, and $A2_{2nd}$. Figure 6 shows focus identification rates of stimulus set 1 pairing *Q2* with $A2_{no}$, $A2_{1st}$, and $A2_{2nd}$. Figure 7 shows focus identification rates of stimulus set 2 ($A2_{no}$, $A2_{1st}$, and $A2_{2nd}$). These results are averaged across the speakers and the subjects.

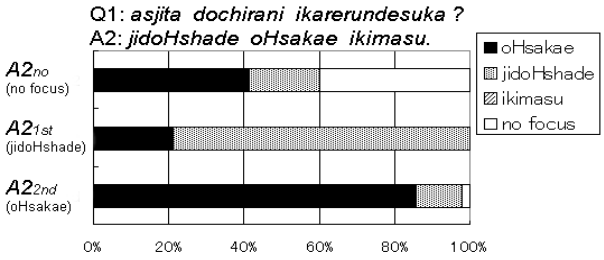


Fig. 5. Focus identification rates for stimulus set 1 pairing *Q1* with $A2_{no}$, $A2_{1st}$, and $A2_{2nd}$. The clause in parentheses shows the focused clause according to the prosodic characteristics.

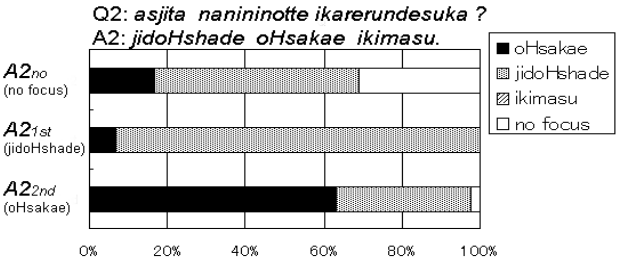


Fig. 6. Focus identification rates for stimulus set 1 pairing *Q2* with $A2_{no}$, $A2_{1st}$, and $A2_{2nd}$. The clause in parentheses shows the focused clause according to the prosodic characteristics.

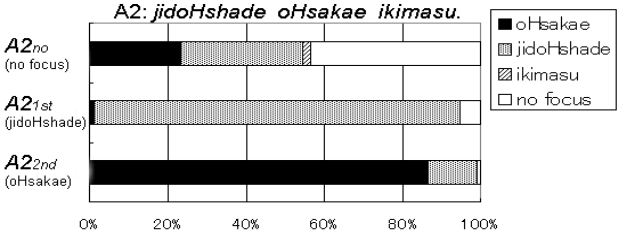


Fig. 7. Focus identification rates for stimulus set 2 ($A2_{no}$, $A2_{1st}$, and $A2_{2nd}$). The clause in parentheses shows the focused clause according to the prosodic characteristics.

4.2.1. Effects of Prosodic Feature and Context

The differences between the stimuli that have a conflict between the focal prominence and the context focus were tested by ANOVA. The critical F value is $F(1, 28) = 4.16, p < .05$.

The results show that there are significant differences

between the answers selecting "*jidoHshade*" as the focused clause and the answers selecting "*oHsakae*" as the focused clause in $A2_{1st}$ of Figure5 ($F = 30.8$). There are also significant differences between the answers selecting "*jidoHshade*" as the focused clause and the answers selecting "*oHsakae*" as the focused clause in $A2_{2nd}$ of Figure 6 ($F = 4.59$). These results indicate that prosodic features are more significant than context for focus identification.

4.2.2. Effect of the Presence of a Question

ANOVA was adopted to clarify the difference of the existence of the question.

First, we tested whether there was a significant difference in the identification of the focused clause ("*oHsakae*") between the results in Figure 6 and 7. The results show that there are significant differences in $A2_{1st}$ ($F = 7.22$), but not in $A2_{no}$ ($F = 2.04$) or $A2_{2nd}$ ($F = 0.02$).

Second, we tested whether there was a significant difference in the identification of the focused clause ("*jidoHshade*") between the results in Figure 6 and 7. The results show that there are no significant difference in $A2_{no}$ ($F = 1.35$), $A2_{1st}$ ($F = 0.00$), or $A2_{2nd}$ ($F = 3.74$).

These results indicate that the effect of the existence of a question is minimal. The result shown above in 4.2.1, that prosodic features are more significant than the context for focus identification, is supported by this outcome.

4.2.3. Effect of Different Questions

ANOVA was adopted to clarify whether there is an effect of the different question between the results of Figure 5 and 6.

First, the results with $A2_{no}$, which have no focal prominences, were tested. There were significant differences in the results with respect to the clauses "*jidoHshade*" ($F = 9.83$) and "*oHsakae*" ($F = 5.96$).

Next, the results of $A2_{1st}$, which have the focal prominence in the clause "*oHsakae*" were tested. There are significant differences in the results with respect to the clauses "*jidoHshade*" ($F = 4.92$) and "*oHsakae*" ($F = 4.78$).

However, in the results of $A2_{2nd}$, which have the focal prominence in the clause "*jidoHshade*", there were no significant differences with respect to the clauses "*jidoHshade*" ($F = 2.79$) or "*oHsakae*" ($F = 2.79$).

These results indicate that if focal prominence does not exist in the sentence or the focal prominence is in the second clause in the sentence, focus identification is affected by a difference in the question, in other words a difference of context. However, if the focal prominence is in the first clause in the sentence, it is not affected by contextual difference.

In other words, if there is no focal prominence in the sentence, a focus is suggested by the context. Further, if the focal prominence is in the second clause in the sentence, the perceived focus is moved to the focus suggested by the context.

This results show that context affects focus identification, and it is probable that word order has some effect for the identification.

5. SUMMARY AND CONCLUSIONS

In this paper, we have shown that focus identification in Japanese dialogue speech is affected by not only prosodic clues but also by semantic and syntactic clues. Therefore, when generating answers in a a spoken dialogue system, we should consider the context and the word order of the answer sentence. Several subjects reported that they had a feeling of wrongness with some dialogues in the experiment, so our next task is to clarify the conditions that cause the feeling of wrongness.

Acknowledgement

This work was supported by Grant-in-Aid for Scientific Research on Priority Areas (B) "Prosody and Speech Processing", Ministry of Education and Science.

6. REFERENCES

[1] P. Boersma, and D. Weenink, "Praat: a system for doing phonetics by computer," `http://www.fon.hum.uva.nl/praat/`, 2000.

[2] H. Fujisaki, K. Hirose, N. Takahashi, and M. Yokoo, "Realization of accent components in connected speech," Trans. Committee Speech Res., Acoust. Soc. Jpn, 279–286, 1984.

[3] K. Shirai, and K. Iwata, "Prosodic rules for speech synthesis representaing word emphasis," IEICE trans. J70-A(5) 816–821, 1984.

[4] S. Takeda, and A. Ichikawa, "Analysis of prosodic features of prominence in spoken Japanese sentences," J. Acoust. Soc. Jpn. 47(6) 386–396, 1991.

[5] S. Takeda, and A. Ichikawa, "Production and evaluation of prominence rules for spoken Japanese sentences," J. Acoust. Soc. Jpn. 47(6) 397–404, 1991.

[6] M. D. Pell, "Influence of emotion and focus location on prosody in matched statements and questions," J. Acoust. Soc. Am. 109(4), 1668–1680, 2001.