Temporal relationship between pause and utterance durations in speech of short sentence

Kazuto Kamoi, Tomohito Yamamoto, Yumiko Muto and Yoshihiro Miyake

Abstract— **In this paper we focused on a pause in speech, and analyzed the factors affecting pause duration. It has been considered that utterance duration just before the pause is the only factor affecting pause duration (preboundary effect), recently effect of utterance duration just after the pause has also been noticed (postboundary effect). However, the relation between two utterance durations and pause duration sandwiched by the durations (pre-post boundary effect) has not been analyzed. Therefore we analyzed these factors inclusively, by using a simple sentence (XY sentence) consisting of two words in speech experiment. Then we used two-way analysis of variance (ANOVA) for analyzing the contribution of factors, which were the utterance duration of these words. As a result, we found two factors affecting a pause. One is utterance duration just before the pause which was already observed, and the other is the ratio of prior and posterior utterance duration. These results suggest that not only a pre or postboundary effect but also a pre-postboundary effect exist in speech.**

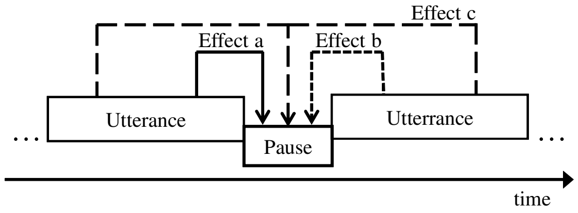


Fig. 1. Utterances and a pause in speech

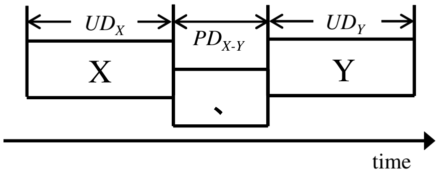


Fig. 2. XY sentence

I. INTRODUCTION

Human communication is composed of message exchanges by using various communication channels. These channels are divided into two channels. One is a verbal channel, and the other is non-verbal channels such as a utterance rhythm, a pause, an accent, a facial expression, a gesture and so on[1]. Recently some researchers have attempted to investigate this human communication mechanism and apply the results to design of robots and speech interfaces[2]. Especially, in the field of audio engineering, technology of speech processing has been significantly developed, and tone patterns and pause durations have been focused as important control factors in a speech synthesis [3].

In this study, we focus on a pause which is a non-verbal channel and an important component of speech. The previous researches have revealed the importance of pause in reading. For example, Sugitou et al.[4], [5] have investigated utterance duration, pause duration and a position of pause in reading a weather statement or folklore. The results showed that the position of pause was similar to that of punctuation which corresponded to the grammatical compartment, and also showed that there was positive correlation between utterance and pause duration. Based on these previous studies, effects from the speech duration to the pause duration can be classified the following three items Fig. 1).

a) Preboundary effect: Effect of the preceding utterance to the pause

b) Postboundary effect: Effect of the following utterance to the pause

c) Pre-Postboundary effect: Effect of the relationship between pre and postboundary utterances to the pause

Most of previous studies have focused the effect a) [6], [7]. On the other hand, recently the effect b) have been analyzed and attracted interest[8]. However, the effect c) has never been studied. Moreover, there is no research covering all of the effect a), b), c) because these effects in a long sentence are complicated. For example, there are a number of pauses in a long sentence, and some pre and postboundary effects are overlaid in an each pause. This situation makes it difficult to analyze each effect independently.

Therefore, in this study by using simple sentences, we analyze these three effects inclusively to clarify the production mechanism of a pause. In Chapter II, the experimental method using a simple sentences and the analysis of factors affecting a pause are described. In Chapter III, the experimental results are descried and in Chapter IV, the production mechanism of a pause is discussed based on the results.

This work was not supported by any organization

K. Kamoi is with the Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori, Yokohama, Japan kamoi@myk.dis.titech.ac.jp

T. Yamamoto is with the Department of Information and Computer Engineering, Kanazawa Institute of Technology, 7-1 Ohgigaoka, Nonoichi, Ishikawa, Japan tyama@infor.kanazawa-it.ac.jp

Y. Muto is with the Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori, Yokohama, Japan yumiko@myk.dis.titech.ac.jp

Y. Miyake is with the Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori, Yokohama, Japan miyake@dis.titech.ac.jp

TABLE I
CLASSIFICATION OF XY SENTENCE

		Postboundary(UD_Y)	
		Group S	Group L
Preboundary	Group S	SS sentence	SL sentence
(UD_X)	Group L	LS sentence	LL sentence

Fig. 3. A scene of speech experiment

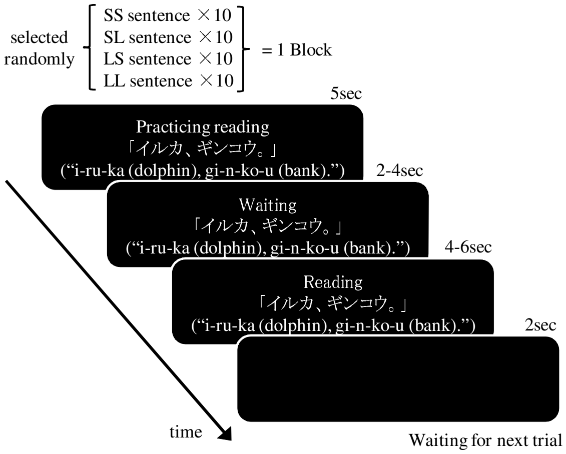


Fig. 4. Experimental procedure

TABLE II
TWO FACTORS EFFECTING THE PAUSE OF XY SENTENCE

Factor	UD_X (A)	UD_Y (B)
Level 1	short (a1)	short (b1)
Level 2	long (a2)	long (b2)

II. SPEECH EXPERIMENT

A. Task and Condition

In this experiment, "XY sentence" which included only one pause was used. This sentence was special comparing to a natural sentence, however it made it possible to analyze three effects independently. XY sentence was composed of two words (see Fig. 2). In the Figure, UD_X, UD_Y and PD_{X-Y} indicate X utterance duration, Y utterance duration and a pause duration respectively. By controlling the length of X and Y, the temporal relationship between these durations were analyzed.

In the experiment, four types of XY sentences (SS, SL, LS, LL sentence, see TABLE I) were prepared. The word X and Y were divided into two groups. The group S was composed of a short utterance duration from 3 to 4 letters (2 to 4 moras). The group L was composed of a long utterance duration from 8 to 9 letters (6 to 9 moras). For example, the SS, SL, LS, and LL sentences would be "i-ru-ka (dolphin), gi-n-ko-u (bank).", "i-ru-ka (dolphin), se-i-sa-n-ka-ku-ke-i (equilateral triangle).", "so-re-ni-mo-ka-ka-wa-ra-zu (nevertheless), gi-n-ko-u (bank).", and "so-re-ni-mo-ka-ka-wa-ra-zu (nevertheless), se-i-sa-n-ka-ku-ke-i (equilateral triangle).". All words of the group S and L applied in this experiment were selected randomly in the basic word database (for group S from 5730 and L from 243 words)[10], and had a high value of familiarity by eliminating the illegibility of word combination. All of the experiments were performed by using Japanese.

B. Participants and Equipment

The participants were healthy 13 students (12 male and 1 female). They were native Japanese speaker and have no disabilities in hearing, sight and speech. The mean age of them was 23 years old. During the experiment, we asked the participants to sit on a chair (see Fig. 3) and read the XY sentences displayed by LCD monitor of the computer (LATITUDE E5400, DELL). The distance between the participant and the monitor was 50cm. XY sentences were presented automatically on the monitor by MAT-LAB (version7.8.0.347, Psychtoolbox-3).

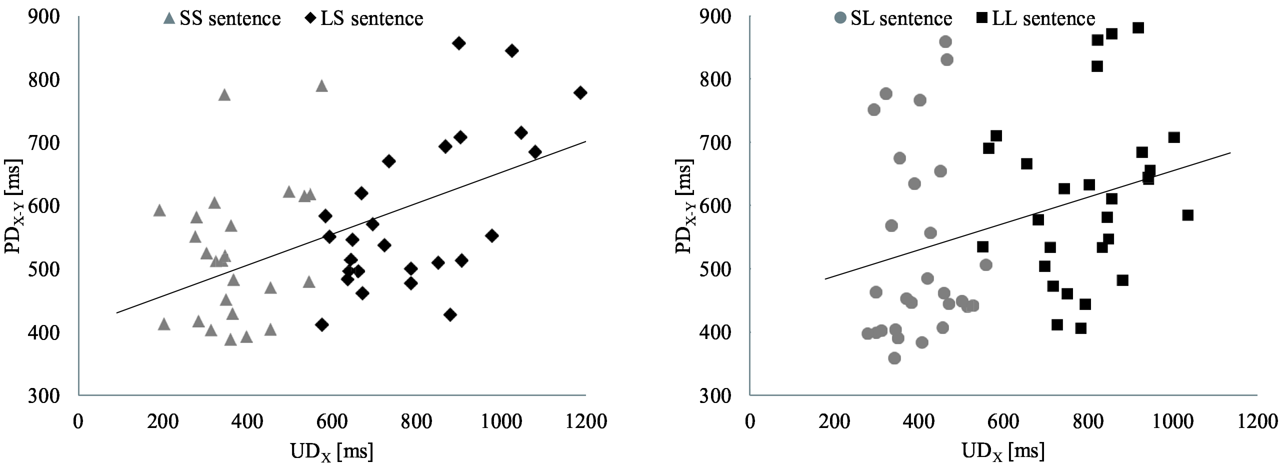
After the experiment, audio data was saved as a wav format, and the average value of sound pressure was calculated every millisecond to analyze utterance and pause durations. The experiment was conducted in the soundproof room (SILENT DESIGN, 2.1m length, 2.6m width, 1.7m height, Fig. 3) with comfortable temperature and brightness.

C. Experimental Procedure

Fig. 4 shows the experimental procedure. First, the participant practiced reading XY sentence for 5 seconds. After finishing the practice, the participant asked to wait 2-4 seconds (determined randomly), and then they started to read this sentence again. This is a series of the steps and was repeated for four types reading (SS, SL, LS, LL sentence) in the experiment. Each sentence was applied 10 times, which is named a "Block". The orders of displaying each sentence were decided randomly. One experiment was composed of three blocks, and at the end of each block, participants had enough rests. In addition, participants were instructed to take a pause at a comma and speak at natural speed. Also, they were asked not to take breaths in the reading of XY sentence, and to ignore the context of sentences.

D. Data Analysis

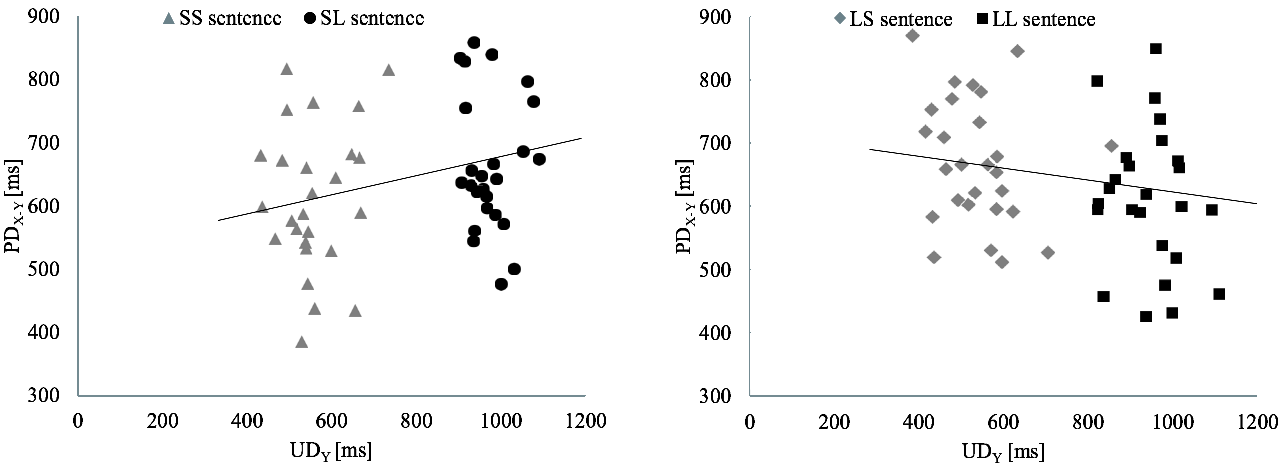
In this study, two effects for the pause: utterance duration before the pause: UD_X (A), and utterance duration after pause: UD_Y (B) were analyzed by using two-way ANOVA



(a) SS and LS condition

(b) SL and LL condition

Fig. 5. A relationship between preboundary utterance duration and pause duration in XY sentence (A typical example of one subject. Fig. 5(a) are ploted 26 reading data (removed misreading 4 data from 30) of SS sentence and 26 of LS. Fig. 5(b) are ploted 28 reading data of SL sentence and 28 of LL.)



(a) SS and SL condition

(b) LS and LL condition

Fig. 6. A relationship between postboundary utterance duration and pause duration in XY sentence (A typical example of one subject. Fig. 6(a) are ploted 26 reading data of SS sentence and 26 of SL. Fig. 6(b) are ploted 27 reading data of LS sentence and 27 of LL.)

(TABLE II). When a significant effect was observed, we analyzed main effects and an interaction of two factors. Misreading data(approximately 6% of all) were removed. It was also confirmed for the normality of all data on pause duration.

III. RESULT

A. Relationship Between Pause Duration and Utterance Duration Before or After The Pause

Fig. 5a shows an example of the relationship between UD_X and PD_{X-Y} in SS and LS sentences. Fig. 5b shows in SL and LL sentences. In addition, TABLE IIIa shows the mean of pause duration corresponding to the condition of each sentence in the Fig. 5. From these results, pause duration of LS and LL sentences tends to be longer than that of SS and SL respectively. This result means that the longer the UD_X becomes, the longer the PD_{X-Y} becomes. Fig. 6a shows an example of the relationship between UD_Y and PD_{X-Y} in SS and SL sentences. Fig. 6b shows in LS and LL sentences. In addition, TABLE IIIb shows the mean of pause duration corresponding to the condition of each sentence in the Fig. 6. From these results, pause duration of

TABLE III

MEAN PAUSE DURATIONS OF EACH XY SENTENCE SHOWN IN FIG. 5 AND 6

(a) Preboundary

XY sentence	SS	LS	SL	LL
PD_{X-Y}[ms]	515.82	585.54	529.46	613.57

(b) Postboundary

XY sentence	SS	SL	LS	LL
PD_{X-Y}[ms]	612.051	670.95	667.27	641.077

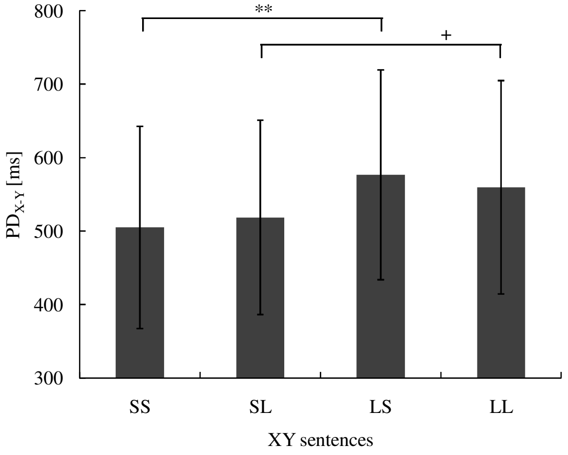


Fig. 7. Mean pause durations of each XY sentence (two-way ANOVA,**:$p < .01$,+:$p < .10$)

SL sentences tends to be longer than that of SS, however pause duration of LL sentences tends to be shorter than that of LS. This result means that if the UD_Y becomes longer, the PD_{X-Y} does not always becomes longer and vice versa.

Fig. 7 shows the mean of pause durations in each condition from all experimental data. From result of ANOVA for these values, there is a significant effect of UD_X: the factor A($F(1,12) = 6.466,\ p < .05$), and there is no significant effect of UD_Y: the factor B($F(1,12) = 0.036,\ p = .854$). Moreover, there is a marginal significance in the interaction of utterance duration before and after the pause ($F(2,24) = 3.863,\ p < .10$)

In addition to the result of ANOVA, simple main effects were analyzed. As a result, there is a significant or marginally significant effect of UD_X in all combinations: A[b1]($F(1,24) = 9.298,\ p < .01$) and A[b2]($F(1,24) = 3.066,\ p < .10$). On the other hand, there is no significant effect of UD_Y in all combinations: B[a1]($F(1,24) = 1.262$, $p = .272$) and B[a2]($F(1,24) = 1.993,\ p = .171$). These results mean that the effect a): preboundary effect exists, but the effect b): postboundary effect does not exist.

B. Relationship Between The Ratio of Utterance Duration Before and After The pause and Pause Duration

It is suggested that there is the interaction of utterance duration around the pause from the result of the ANOVA. This result suggests that the effect c): pre-postboundary effect exists. In previous studies, a quantitative method to analyze this effect has not been proposed. Therefore, we introduce a measure to analyze this effect, based on the relationship between utterance and pause in XY sentence.

In this experiment, the pause duration of SL and LS sentences tends to be longer than that of SS and LL. This result implicates that the ratio between utterance duration before and after the pause affects the pause duration. Therefore, we focus on this ratio quantitatively. First, the ratio between utterance durations: σ is defined the following in equation (1).

$$\sigma = \frac{Max(UD_X, UD_Y)}{Min(UD_X, UD_Y)} \qquad (1)$$

From this definition, the magnitude of the σ value becomes larger when the change of utterance duration becomes larger.

Fig. 8 shows an example of the relationship between σ and PD_{X-Y}, and TABLE IV shows the mean of pause duration corresponding to each sentence. From this result, pause duration of large σ sentences: SL, LS tends to be longer than that of small σ: SS and LL. This result indicates that the larger the σ becomes, the longer the PD_{X-Y} becomes.

Fig. 9 shows the mean of pause duration for each condition from all experimental data. As a result of paired t-test, there is a significant difference between the large and small σ sentences ($t(12) = 2.657$,$p < .05$) . These results mean that the effect c): pre-postboundary effect exists.

IV. DISCUSSION

In this study, we analyzed the effect from the utterance before and after the pause to the pause, by using XY sentences. The results showed that the effect a) and the effect c) existed. These results suggest that the pause length should be based on two mechanisms. One is the “causal” mechanism: the mechanism by the relationship, which is composed of the past utterance. The other is the “synchronic” mechanism: the mechanism by the relationship, which is composed of the past and anticipated future utterances. The causal mechanism has been featured in previous studies. On the other hand, the synchronic mechanism is mentioned for the first time in this study.

The causal mechanism has been discussed in the field of cognitive science so far. For example, on the research on time perception, there is a hypothesis that humans are equipped with an internal clock. In this hypothesis, the brain accumulates durations, and then compares it to memorized durations and produce time estimations[11]. Moreover, this internal clock has been observed in multiple sensory modalities independently[12]. If it is assumed that utterance and pause duration are associated with memorized and accumulated durations in internal clock model, effect a): preboundary effect could be explained. There are another explanation which are based on the HMM[9] and the multi-space probability[13] to determine the pause. Effect a) could be modeled by using these approaches. Thus, the causal mechanism has been discussed in previous studies.

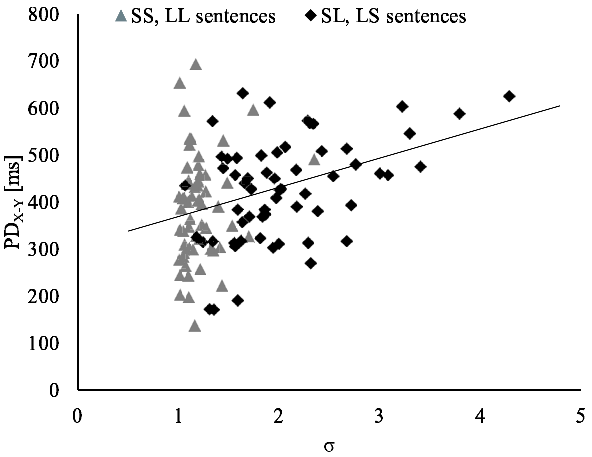


Fig. 8. A relationship between pause duration and the ratio between before and after utterance duration in XY sentence (A typical example of one subject. Fig. 8 are ploted 60 reading data of SS, LL sentences and 60 of SL, LS.)

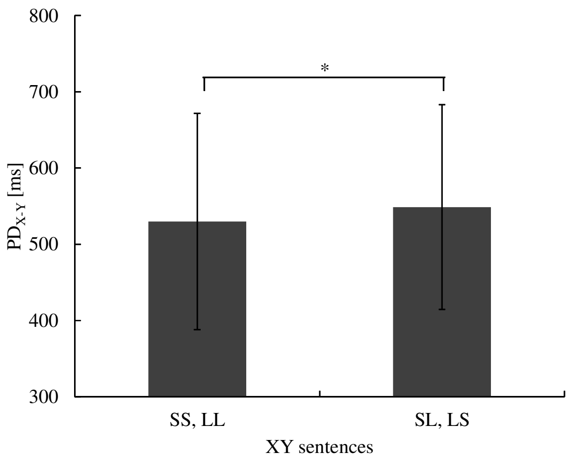


Fig. 9. Mean pause durations of SS, LL sentences and SL, LS sentences (paired t-test, *:$p < .05$)

TABLE IV
MEAN PAUSE DURATIONS OF XY SENTENCES SHOWN IN FIG.8

XY sentences	SS, LL	SL, LS
PD_{X-Y} [ms]	388.26	428.18

On the other hand, these discussions cannot explain effect c) which contains future utterance and not sufficient to handle the synchronic mechanism. One explanation for the synchronic mechanism is perceptual grouping phenomena. When people perceive the successive stimuli, they are affected by the relationship between before and after the current stimulus. Kurosawa et al.[14] investigated the effect of grouping phenomena from the relationship between before and after stimulation in experiments using tone-burst series. As a result, the perception of sensory stimulus followed in the preceding stimuli, not only in the visual system but also in the auditory system. In addition, Yonezawa and Akagi[15] showed two effects from phonetic stimuli close to the current stimulus and modeled them. The first effect is an assimilation effect contributing to reduce the variation in patterns of perception, when humans perceive the same stimuli. The second effect is a contrast effect contributing to emphasize the difference of perception and separating patterns of perception, when humans perceive the dissimilar stimuli. In this study, experiments were performed using XY sentence. In the practice procedure, it is considered that participants regard this sentence as one coherent speech behavior consisting of three elements: word X, pause and word Y. This process causes the grouping between the pause and utterances before and after it, and produces the effect c).

In human speech or dialogue, we considered these mechanisms play an important role for realizing its smoothness. For example the causal mechanism, based on the past speech, would contribute to reducing the temporal deviation of the pause and stabilizing the rhythmic structure of speech. The synchronic mechanism, based on the past and anticipated future speech, would contribute to predicatively stabilize the rhythmic structure of speech with a range of time. It is speculated that human speech achieves its smoothness by these two mechanisms working simultaneously.

On the other hand, the results of our experiment are contrary to those of Krivokapić's[8], which supported effect b). There are two reasons which are able to explain the difference. The first reason is that there is the difference between "period" and "comma"[5]. Although Krivokapić used a period, we focused on a comma. This difference possibly affected the results. The second reason is the effect of a breather. Sugitou[5] has been reported that the existence of a breather affects the importance for the semantic punctuation. Sentences of Krivokapić's study may be too long to breathe and its affected the pause duration. In XY sentence, it was not necessary to breath in speech, and it made a difference of the results between our and previous study.

Two mechanisms discussed here may contribute to realize smooth human-human or human-robot communications. If the pause is affected by an immediately preceding or following utterance, it might be possible to apply the effect to general sentences including multiple pauses. In future works, we will develop the pause decision model based on our results and evaluate its effectiveness.

In this study, we analyzed the effect of utterance that is temporally closest to pause. On the other hand, Ozeki et al.[13] have proposed a model that determines the natural pause duration and position from a preceding pause and segment intensity. Moreover, pause is affected by linguistic attributes[6], speech rate[7] and intension and so on. Therefore, in future works, we are going to take these factors into consideration and expand the model.

V. CONCLUSION

In this paper we focused on a pause in the speech, and analyzed the factors affecting pause duration. As a result, we found two factors affecting a pause. One is utterance duration just before the pause which was aleardy observed, and the other is the ratio of prior and posterior utterance duration. These results mean that not only a pre or postboundary effect but also a pre-postboundary effect exist in speech, and we discussed its mechanism. In future works, we are going to take another factor such as linguistic attributes into consideration and expand the model.

REFERENCES

[1] V. P. Richmond and J. C. McCroskey, *Nonverbal Behavior in Interpersonal Relations*, Allyn & Bacon, 2007

[2] T. Hayashi, S. Kato and H. Itoh, "A Mental Rhythm Synchronous Model Using Paralanguage for Communication Robot(in Japanese)," The 23rd Annual Conference of JSAI, 1H2-4, 2009, pp.17-19

[3] M. Yamada, K. Iwano and S. Furui, "A Study on F_0 Contour Generation Factors Using Categorical Multiple Regression(in Japanese)," The Special Interest Group Technical Reports of IPSJ, **38**-3, 2001, pp.15-20

[4] M. Sugitou, "The Relation between Punctuation and Prosodic Features of Utterances in Weather Forecast Sentences(in Japanese)," Osaka Shoin Women's College collected essays, **22**, 1985, pp.1-7

[5] M. Sugitou and G. Ohyama, "Studies in Phonetics and Speech Communication(in Japanese)," Kinki Society of Phonetics, 1990, pp.199-211

[6] N. Kaiki and Y. Sagisaka, "Study of Pause Insertion Rules Based on Local Phrase Dependency Structure(in Japanese)," The transactions of IEICE, **J79-D-II**, 9, 1996, pp.1455-1463

[7] N. Minematsu, Y. Kataoka, S. Nakagawa, "Analysis of Spoken Language in Lecture Style(in Japanese)," IPSJ SIG Notes, **95**-100, 1995, pp.39-46

[8] J. Krivokapić, "Prosodic planning: Effects of phrasal length and complexity on pause duration," Journal of Phonetics, **35**, 2007, pp.162-179

[9] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, "Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis(in Japanese)," The transactions of IEICE, **J83-D-II**, 11, 2000, pp.2099-2107

[10] S. Amano, T. Kobayashi and NTT Communication Science Laboratories, *Basic word database of Japanese with semantic word familiarity*, Gakken, 2008

[11] T. Michel, "Temporal discrimination and the indifference interval, implications for a model of the 'internal clock'," Psychology Monographs, **77**, 1963, pp.1-31

[12] W. H. Meck and R. M. Church, "Simultaneous temporal processing," Journal of Experimental Psychology, **100**, 1984, pp.1-29

[13] H. Ozeki, T. Masuko and T. Kobayashi, "A pause modeling technique based on multi-space probability distribution(in Japanese)," IEICE technical report. Speech, **104**-29, 2004, pp.41-46

[14] T. Kurosawa, R. Nishimura and Y. Suzuki, "Context effect on perceptual grouping of toneburst sequences(in Japanese)," IEICE technical report. HIP, **101**-512, 2001, pp.13-18

[15] Y. Yonezawa and M. Akagi, "Modeling of Contextual Effects and Its Application to Word Spotting(in Japanese)," The transactions of IEICE, **J80-D-II**, 1, 1997, pp.36-43