2013 IEEE RO-MAN: The 22nd IEEE International Symposium on
Robot and Human Interactive Communication
Gyeongju, Korea, August 26-29, 2013

A generative model of pause duration considering the relation between utterances before and after a pause

Tomohito Yamamoto[1], Kazuto Kamoi[2], Yoshihiro Miyake[2]

Abstract— **Utterance durations just before and after a pause have been considered to be the only factors affecting pause duration (Preboundary and Postboundary Effects). Recently, using an "*XY* utterance phrase" composed of two words, we discovered that the ratio of the two utterance durations before and after a pause affects pause duration (Pre-postboundary effect). However, it is not obvious whether such effects are useful for speech processing applications. In this research, we developed a generative model of pause duration based on multiple regression analysis from our experimental data (Primal Model), and derived two additional models with different parameters. Furthermore, we evaluated them, comparing them to a model whose pause duration is constant (Constant Model). The result was that the subjects' impressions, such as "natural," "like," and "familiar," of the Primal Model were more positive than those of the Constant Model. Moreover, when compared with the two additional pause duration models, the Primal Model gave the best results. From these results, we discuss the validity of the Primal Model and the relationship between the parameters and the subjective evaluation.**

I. INTRODUCTION

Human communication is composed of message exchanges through various communication channels. These channels are divided into two types [1]. One is the verbal channel and the others consist of non-verbal channels such as utterance rhythms, pauses, accents, and gestures. Recently, some researchers investigated the mechanisms of this type of human communication and applied the results to design communication robots and speech processing systems [2].

In this research, we focus on the non-verbal channel consisting of pauses, a basic component of human speech. Previous research has already revealed the importance of pauses in reading. For example, Sugitou et al. [3] investigated the relations between utterance duration, pause duration, and the position of a pause when reading a weather report. The results showed that the position of a pause was similar to that of punctuation.

In our previous research, we analyzed the relations between a pause and the utterances before and after the pause [4]. We classified the effects of the utterance duration on the pause duration into the following three categories (Fig. 1).

[1]Tomohito Yamamoto is with Department of Information and Computer Engineering, Kanazawa Institute of Technology, 7-1 Ohgigaoka, Nonoichi, Ishikawa tyama at neptune.kanazawa-it.ac.jp

[2]Kazuto Kamoi, Yoshihiro Miyake are with Department of Computational Intelligence and Systems Science, Tokyo Institute of Technology, 4259 Nagatsuta, Midori, Yokohama kamoi at myk.dis.titech.ac.jp, miyake at dis.titech.ac.jp

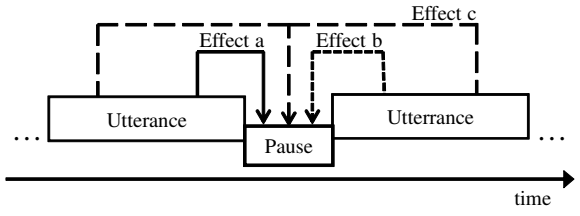


Fig. 1. Effects of the utterance on the pause in speech

(a) Preboundary Effect: the effect of the preceding utterance
(b) Postboundary Effect: the effect of the following utterance
(c) Pre-postboundary Effect: the effect of the relationship between the preceding and following utterances

Among these, the Preboundary and Postboundary Effects have been analyzed in [5], [6], [7]; however, the Pre-postboundary Effect and the relationships among all three effects had not been analyzed. Therefore, in our previous research, we inclusively analyzed these effects. In our experiment, we proposed a simple phrase (an *XY* utterance phrase, Fig. 2) that was composed of two words without restriction on context or breath, and analyzed the contribution of various factors. As a result, we found two factors that affected the pause. One was the utterance duration just before the pause, which had been observed by other researchers, and the other was the ratio between the preceding and following utterance duration [4].

In that research, some aspects of the effects from the utterance to the pause have been analyzed. However, it is not obvious whether such results are useful for speech processing applications. Much of the previous research on speech processing has focused only on utterances, introducing the pause as a constant value [8]. Recently, a generative model of pause duration is being gradually developed. For example, there is a Markov model-based speech processing system that considers the effect of the preceding pause and adjusts the present pause [9]. However, these models have only focused on information before the pause, and have not considered the effects from the utterances before and after the pause.

Therefore, in this research, we first develop a generative model of pause duration based on multiple regression analysis from our experimental data, and then derive two additional models that have different parameters. Next, we evaluate them, along with a model whose pause duration

978-1-4799-0509-6/13/$31.00 ©2013 IEEE

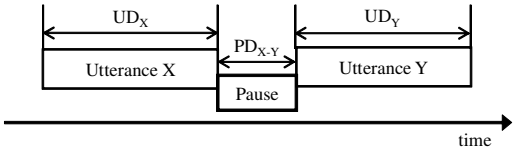


Fig. 2. *XY* utterance phrase

is constant, using a general sentence that is useful for applications.

II. A GENERATIVE MODEL OF PAUSE DURATION AND ITS PARAMETERS

A. A generative model of pause duration

In this subsection, a generative model of pause duration is developed based on the results of previous research [4]. In the following derivation, we assume a sentence with $n+1$ utterances and n pauses, and denote the kth utterance and pause as PD_k and UD_k, respectively.

At first, we consider the Preboundary Effect of the preceding utterance on the pause. In this case, the precedent utterance is directly proportional to the following pause. Therefore, it is necessary to judge the length of utterance durations in a sentence. However, the same utterance duration may be judged long or short depending on the sentence. For that reason, it is necessary to normalize the utterance duration with respect to its sentence. In this research, we define the kth normalized utterance duration UD'_k as

$$UD'_k = \frac{nUD_k}{\sum_{i=1}^{n} UD_i} \tag{1}$$

Note that UD'_k becomes greater than 1 when it is bigger than the mean of the utterance durations.

Next, we consider the Pre-postboundary Effect, the relationship between the preceding and following utterances to the pause. In [4], we described the ratio between the preceding and following utterance durations by variable σ calculated by

$$\sigma = \frac{Max(UD_X, UD_Y)}{Min(UD_X, UD_Y)} \tag{2}$$

The ratio σ is directly proportional to the pause duration. In this research, we normalize this ratio similarly to Equation (1). Here, σ_k is the ratio of utterance durations before and after the kth pause, and σ'_k is its normalized value, as defined below:

$$\sigma_k = \frac{Max(UD_k, UD_{k+1})}{Min(UD_k, UD_{k+1})} \tag{3}$$

$$\sigma'_k = \frac{n\sigma_k}{\sum_{i=1}^{n} \sigma_i} \tag{4}$$

To develop the generative model of pause duration, we assume that the Preboundary and Pre-postboundary Effects are independent, and the multiple regression equation of PD'_k, derived from the normalized utterance duration UD'_k and the normalized ratio between utterance durations σ'_k, can be written as

$$PD'_k = \alpha UD'_k + \beta \sigma'_k + \gamma \tag{5}$$

where the parameters α, β, and γ are explained in the next subsection.

The generative model of pause duration is constructed by multiplying the standard pause duration PD_S by the factor PD'_k. The kth pause duration PD_k may then be described by

$$PD_k = PD'_k \cdot PD_S \tag{6}$$

B. Parameters for the generative model

In this subsection, the parameters of a generative model of pause duration are discussed. The parameters of the multiple regression in Equation (5) are calculated from the results reported in [4]. In that research, 15 subjects spoke four types of *XY* utterance phrases (where the *X* and *Y* words were long or short) repeated 30 times. The forms of UD'_k and σ'_k were derived from these data. The parameters α, β, and γ, calculated from multiple regression analysis based on Equation (5), were found to be:

$$\begin{aligned} \alpha &= 0.1673\,(p < .001) \\ \beta &= 0.0858\,(p < .01) \\ \gamma &= 0.7473\,(p < .001) \end{aligned} \tag{7}$$

All of the regression coefficients were found to be significant or marginally significant.

We can then approximate PD'_k by the following equation, considering that the normalized value should change around the value of 1, and $\alpha+\beta+\gamma = 1.0004 \fallingdotseq 1$.

$$\begin{aligned} PD'_k &= \alpha\left(UD'_k - 1\right) + \beta\left(\sigma'_k - 1\right) + (\alpha+\beta+\gamma) \\ &\fallingdotseq \alpha\left(UD'_k - 1\right) + \beta\left(\sigma'_k - 1\right) + 1 \end{aligned} \tag{8}$$

In this equation, α implements the effect of the preceding utterance on the pause, and β implements the effect of the relationship between the preceding and following utterances on the pause.

In this research, to evaluate the generative model of pause duration based on our experiment data, the Primal Model (P) using the original α and β is given by the following, where K_α and K_β are the values obtained from the multiple regression analysis.

$$\begin{aligned} PD'_k &= 0.1673\left(UD'_k - 1\right) + 0.0858\left(\sigma'_k - 1\right) + 1 \\ &= K_\alpha\left(UD'_k - 1\right) + K_\beta\left(\sigma'_k - 1\right) + 1 \end{aligned} \tag{9}$$

Next, an Emphasized Model (E) that is more affected by the utterance durations than the Primal Model is considered.

In this model, K_α and K_β are twice as big as those of the Primal Model as given by the following.

$$\begin{aligned} PD'_k &= 0.3346\left(UD'_k-1\right)+0.1706\left(\sigma'_k-1\right)+1 \\ &= 2K_\alpha\left(UD'_k-1\right)+2K_\beta\left(\sigma'_k-1\right)+1 \end{aligned} \quad (10)$$

Further, we also consider an Opposite Model (O) such that the Preboundary and Pre-postboundary Effects have the opposite effect of the Primal Model. In this model, K_α and K_β are multiplied by -1 and the model is given by

$$\begin{aligned} PD'_k &= -0.1673\left(UD'_k-1\right)-0.0858\left(\sigma'_k-1\right)+1 \\ &= -K_\alpha\left(UD'_k-1\right)-K_\beta\left(\sigma'_k-1\right)+1 \end{aligned} \quad (11)$$

Finally, a Constant Model (C) with a constant pause duration is given by the following equation, where, α and β are 0 and the model is not affected by the Preboundary and Pre-postboundary Effects.

$$PD'_k = 1 \quad (12)$$

In this research, these four models are evaluated by a paired comparison method.

III. METHOD

A. Evaluation method

To evaluate the proposed models, subjects listened to an avatar's speech synthesized by the models, and were asked for their impressions. An avatar with no facial expressions (Fig.3) [10] was used and the speech was synthesized by VoiceText (SAYAKA, HOYA). Standard pause duration and speech speed were set based on preliminary experiments. Concretely, the pause duration was 550 ms (the default value of VoiceText), and the speech speed was 7.414 mola/s.

Two sentences from a weather report used in previous research [3] were used in this experiment. In these sentences, the punctuation (,) and (.) are inserted according to grammatical and utterance structure from previous research [3].

(A) Nishinihon wo ootteiru, idousei koukiatsu ha, shidaini higashi he idou shi, itsuka ha, kiatsu no tani ga tsuuka suru mikomi desu. (High pressure covering the west side of Japan is going to move east, and low pressure is coming on 5th day of the month.)

(B) Konotame, asa kara, ame no furutokoro ga ooku, nicchu ha, kakuchi tomo, tokidoki ame ni naru deshou. (Therefore, it is partly likely to rain from morning, and by day time, it is going to rain around Japan.)

Pause durations for these sentences were calculated based on the four models. To compare the models, five adjectives were prepared based on previous research [11], [12].

(1) natural
(2) like
(3) polite

Fig. 3. The avatar presented on a PC

Fig. 4. A picture of an evaluation experiment

TABLE I
FOUR-LEVEL LIST OF QUESTION ITEMS

Value	Item
2	The **latter** represents a question item **much** better than the former.
1	The **latter** represents a question item **little** better than the former.
−1	The **former** represents a question item **little** better than the latter.
−2	The **former** represents a question item **much** better than the latter.

(4) familiar
(5) fast

Items (1)-(4) were prepared for evaluating the impression of an avatar's speech, and item (5) was prepared to evaluate the subjective time of speech length. The item "natural" means that the sentence is natural compared to human speech. In this experiment, two of the four models were chosen, and speech modified by each of these models was presented to a subject. Subjects were asked which speech was more suitable with respect to each item by four degrees, and a pair-wise comparison was conducted on the results (Table I). In this experiment, the Nakaya variation that considers all combinations but does not consider comparison order was used.

B. Subjects and experiment system

The subjects consisted of twelve healthy male students. They were all native Japanese speakers and had no disabilities with respect to hearing, sight, or speech. Their mean age was 23 years old.

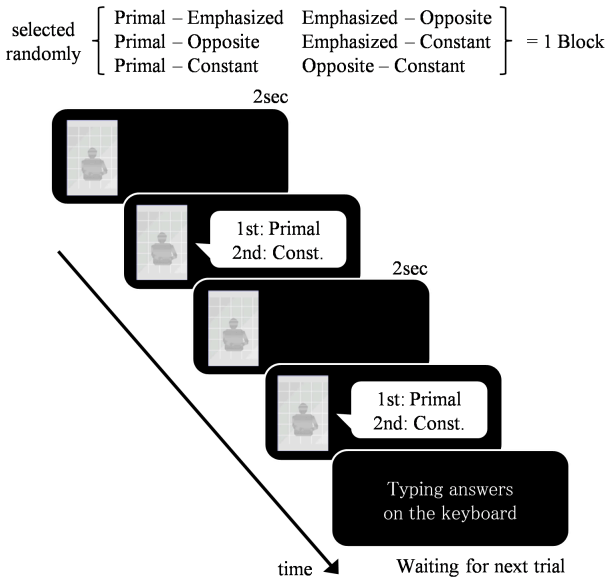


Fig. 5. Experimental evaluation procedure. In this trial, a subject listens to the Primal Model first, and then the Constant Model.

The avatar was displayed on the LCD monitor of a computer (LATITUDE E5400, DELL). Speech was presented automatically by MATLAB (Version 7.8, Psychtoolbox-3) from a speaker (MS-105USV, ELE-COM).

The experiment was conducted in a soundproof room (produced by SILENT DESIGN, prefabricated soundproof rooms, 2.1*2.6*1.7 m (l*w*h)) with a comfortable temperature and level of illumination. During the experiment, the participants were asked to sit on a chair (Fig. 4) and listen to the synthesized speech of the avatar. The distance between the participant and the monitor was 50 cm and the gaze of the avatar was coordinated with the subject's gaze. The size of the avatar was 108*72 mm (h*w).

C. Experimental procedure

An overview of the experimental procedure is shown in Fig. 5. First, the avatar was presented to a subject on the monitor and after two seconds, speech was presented. For each trial, two models were randomly chosen from the four models and speech from each model was presented to the subject. After the presentation, the subject answered questions about their impressions using the computer.

An experimental block was composed of six trials. Before the experiment, the subject was given two practice trials. An experiment was composed of two blocks (one block for Sentence (A), and the other for Sentence (B)). After each block, subjects were allowed to rest. In this experiment, the subjects were divided into two groups, and the order of the sentences was switched for each group.

IV. RESULTS

Table II shows the result of ANOVA results to for each question item. The results show that there is a significant or marginally significant difference between among the models in for Ssentences (A) or and (B). In this section, the results of the comparison between the Constant model and Primal Model is are described at first, and then next, the results of the comparison between of all models is are described.

A. Comparison between the Primal Model and the Constant Model

Fig. 6 shows the values of the evaluation questions for the Primal and Constant Models. For the question item "natural," the evaluation score of the Primal Model is significantly higher than that of the Constant Model for Sentence (A) ($p < .05$). For Sentence (B), the difference between the models is marginally significant ($p < .10$). Moreover, for the question item "familiar," the evaluation score of the Primal Model is significantly higher than that of the Constant Model for Sentence (A) ($p < .05$). These results indicate that both the Preboundary and Pre-postboundary Effects of the Primary Model improve these impressions.

For the question item "like," the evaluation score of the Primal Model tends to be higher than that of the Constant Model, and the scores of Sentence (A) tend to be lower than those of Sentence (B). For the question item "polite," there is no significant difference between the Primal and Constant Models. These results indicate that the choice of sentence affects the evaluation of the models, and the question item "polite" has a different connotation to the other question items.

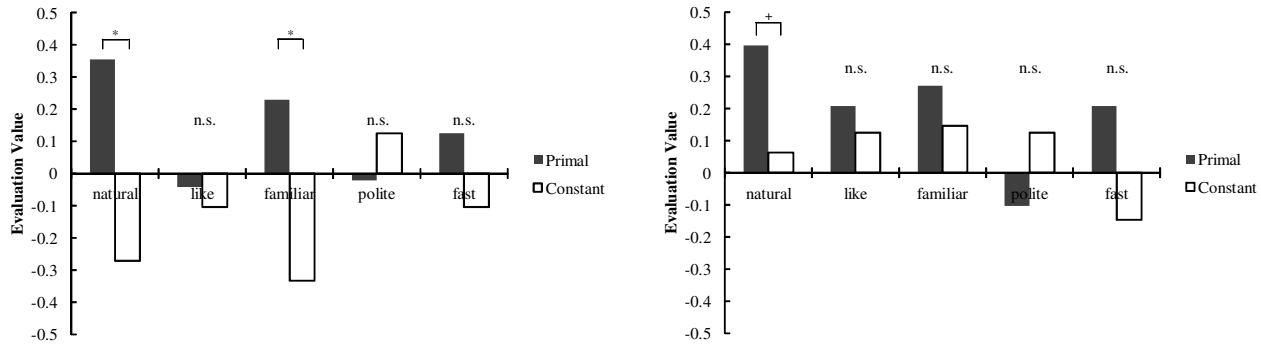
For the question item "fast," there is no significant difference between the Primal and Constant Models. In this experiment, generative models calculated the pause duration based on the corrected value. Therefore, the total length of the pause duration for both models was the same. Moreover, the speech speed of both models was also equal, and as a result, there was no difference between the total speech duration of the models. This result means that the subjects evaluated the speed of speech properly.

B. Parameters of genearative models and its their evaluation

Fig. 7 shows the relationship between the parameters of the models and their evaluation. For the question items "natural," "like," and "familiar," the Primal Model ($\alpha = K_\alpha$, $\beta = K_\beta$) gets mostly higher scores than the Emphasized Model ($\alpha = 2K_\alpha$, $\beta = 2K_\beta$). In particular, for Sentence (B), the Primal Model gets significantly higher evaluation scores than the Emphasized Model ($p < .05$). On the other

TABLE II
THE RESLUTS OF ANOVA IN EACH QUESTION ITEM ($***$:$p < .001$, $**$:$p < .01$, $+$:$p < .10$, $n.s.$:NON SIGNIFICANT)

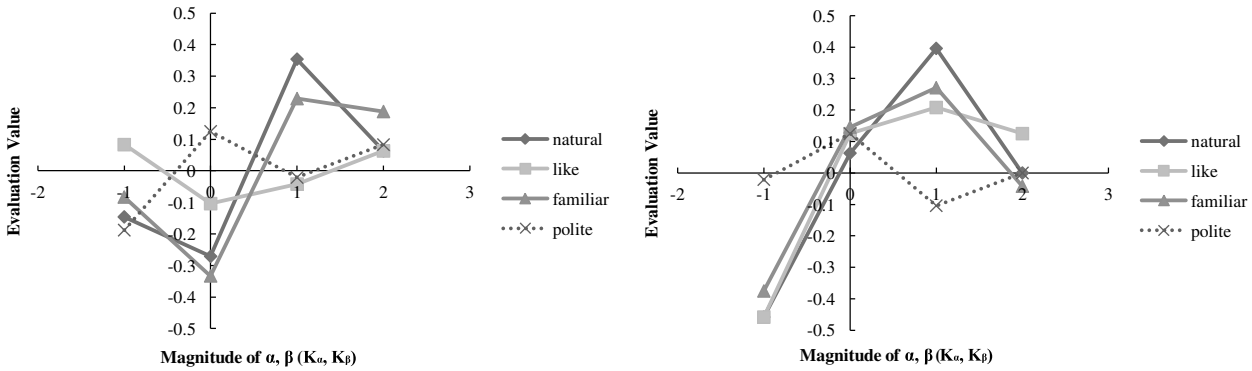
	Sentence A		Sentence B	
natural	7.17	$F(3,22)=5.45^{***}$	11.86	$F(3,22)=15.34^{***}$
like	0.75	$F(3,22)=0.52^{n.s.}$	9.11	$F(3,22)=9.18^{***}$
familiar	6.58	$F(3,22)=6.49^{***}$	7.58	$F(3,22)=7.51^{***}$
polite	1.86	$F(3,22)=2.22^{+}$	0.86	$F(3,22)=1.08^{n.s.}$
fast	1.03	$F(3,22)=1.06^{n.s.}$	4.14	$F(3,22)=5.03^{**}$



(a) Sentence (A)

(b) Sentence (B)

Fig. 6. Evaluation value of question items (natural, like, familiar, polite and fast) between Primal and Constant Model in (a) Sentence (A), and (b) Sentence (B) (pairwise comparison, $*$:$p < .05$, $+$:$p < .10$, *n.s.*:non significant)



(a) Sentence (A)

(b) Sentence (B)

Fig. 7. Relationship between the evaluation scores of the question items and the magnitude of α, β in (a) Sentence (A), and (b) Sentence (B)

hand, the Opposite Model ($\alpha = -K_\alpha$, $\beta = -K_\beta$) tends to get lower values than the Primal or Emphasized Models. There is a particularly significant difference between the Primal and Opposite Models for the question item "like" of Sentence (A), and for all the question items of Sentence (B). These results shows that adjusting the parameters of model to the K_α and K_β obtained from the results of the original experiments is effective for improving a subject's impressions.

V. DISCUSSION

In this research, we developed a generative model of pause duration based on multiple regression analysis from our experimental data (Primal Model), and derived two additional models with different parameters. Further, they were compared to a model with a constant pause duration (Constant Model). The results show that for the question items "natural," "like," and "familiar," the evaluation scores of the proposed Primal Model tended to be higher than those of the Constant Model. Further, the Emphasized Model that was more affected by the utterance durations and tended to get lower scores than the Primal Model. The Opposite Model that was affected oppositely by the utterance durations tended to get lower scores than all other models. On the other hand, the question items "polite" and "fast," did not score significantly differently among the models.

The fact that the Primal Model tends to get higher scores than the Constant Model means that the generative model of pause duration based on two effects (the Preboundary and Pre-postboundary Effects) effectively improves a subject's impressions. The result also means that the model based on the experimental results of an XY utterance phrase is applicable for long sentences that contain several pauses.

Further, in this experiment, the total length of pause duration in each model was the same, and there was no difference between speech speeds. Nevertheless, the impression of each model was different. This result suggests that it is possible to change a subject's impressions without an apparent change of duration, and the relation between each utterance and pause

duration is an important factor in a subject's impressions.

In the results for the various parameters of the models and their impressions, the Primal Model gets higher scores than the Emphasized and Opposite Models. In particular, the Opposite Model gets the lowest scores. These results means that the values for K_α and K_β obtained from the original experimental data is the most preferable for improving a subject's impressions. Muto et al.[10] reported that the proper switching pause duration between a speaker and an avatar is 600 ms for the question item "natural" and 900 ms for "polite," and that other durations get lower scores than these. The results of our experiment also suggest that there are optimized parameters for the generative model.

On the other hand, there is no significant difference between models for the question item "polite." These results may be due to the literal meaning of "polite." For example, Yamamoto et al. have reported that the time difference between utterance and body motion affected a subject's impressions, however, the result of the question item "polite" was different from the result of the other impressions [11]. Nagaoka et al. also have reported that in human dialogue, the pause duration of a speaker or switching pause duration affected the question items "natural" and "like," however, these durations did not affect the question item "polite" [12]. Our results are consistent with these results, and the word "polite" is different from the other words used to evaluate human communication.

In this research, a generative model was developed using a parameter α that represents the Preboundary Effect, and a parameter β that represents the Pre-postboundary Effect. However, it is unclear which parameter is more effective at improving a subject's impressions. For example, the value of K_α is double that of K_β, and in this experiment, the ratio was fixed. Therefore, the relationship between the values is not clear. In future work, it will be necessary to investigate this relationship in more detail.

Further, there is the difference between some of the results of Sentence (A) and Sentence (B). One reason could be the length of the words before and after the commas. In Sentence (A), there are long words before and after the comma, but in Sentence (B), there are short words before and after the comma. The generative model calculates pause duration based on the total length of pauses, and the difference between the word length of sentences affected the impressions. In addition to this effect, pause duration in general sentences is affected by the preceding pause, language attributes, a speaker's intention, and so on. In future work, it will be necessary to take other factors that affect pause duration into consideration, and to expand the generative model of pause duration.

VI. CONCLUSIONS

In this research, we developed a generative model of pause duration based on multiple regression analysis from our experimental data (Primal Model), and derived two additional models with different parameters. Furthermore, they were compared to a model with constant pause duration (Constant Model). The results show that for the question items "natural," "like," and "familiar," the evaluation scores of the proposed Primal Model tended to be higher than those of the Constant Model, and compared to the two additional models, the Primal Model gave the best results.

In future work, it will be necessary to investigate the relationship between the parameters of the generative model more thoroughly. Further, it will be necessary to consider other factors that affect pause duration to develop a more general model of pause duration.

REFERENCES

[1] R. Virginia and M. James, Nonverbal behavior in interpersonal relations, Ally & 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, pp.17-19, 2009

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

[4] K. Kamoi, T. Yamamoto, Y. Muto, and Y. Miyake, "Temporal Relationship between Pause and Utterance Durations in Speech of Short Sentence," Proc. of the SI International 2011, pp.100-105, 2011

[5] 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, pp.1455-1463, 1996

[6] J. Krivokapić, "Prosodic Planning: Effects of Phrasal Length and Complexity on Pause Duration," Journal of Phonetics, Vol.35, pp.162-179, 2007

[7] J. Krivokapić, "Speech Planning and Prosodic Phrase Length," Speech Prosody, 100311, pp.1-4, 2010

[8] 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, No.11, pp.2099-2107, 2000

[9] H. Ozeki, T. Masuko, and T. Kobayashi, "A Pause Modeling Technique Based on Multi-Space Probability Distribution (in Japanese)," IEICE Technical Report, Speech, Vol.104, No.29, p.41-46, 2004

[10] Y. Muto, K. Takano, H. Ora, Y. Kobayashi, T. Yamamoto, and Y. Miyake, "Utterance Timing Control on Speech Dialog Interface and its Evaluation (in Japanese)," Proc. of Human Interface Symposium 2007, pp.639-642, 2007

[11] M. Yamamoto and T. Watanabe, "Timing Control Effects of Utterance to Communicative Actions on Embodied Interaction with a Robot and CG Character," International Journal of Human-Computer Interaction, Vol. 24, No.1, pp. 87-107, 2008

[12] C. Nagaoka, M. Draguna, M. Komori, and T. Nakamura, "The Influence of Switching Pauses on Interpersonal Perception in Dialogues (in Japanese)," Proc. of Human Interface Symposium, pp.171-174, 2002