ACOUSTICAL LETTER

Study on parameters of the variable threshold to detect local speech rate deceleration in Japanese spontaneous conversational speech

Keiichi Takamaru[1,*], Makoto Hiroshige[1], Kenji Araki[1] and Koji Tochinai[2]

[1]*Hokkaido University,*
N13-W8, Kita-ku, Sapporo, 060–8628 Japan
[2]*Hokkai-Gakuen University,*
4–1–10 Asahimachi, Toyohira-ku, Sapporo, 062–8605 Japan

(*Received 27 November 2002, Accepted for publication 17 February 2003*)

Keywords: Mora duration, Local speech rate, Spontaneous conversational speech, Variable threshold
PACS number: 43.71.Bp, 43.66.Mk, 43.72.Kb

1. Introduction

In human communication, speech conveys not only linguistic information but also emphasis, intention, attitude and so on. They are called paralinguistic information [1]. There are several researches on paralinguistic information [2,3]. Methods for modeling or detecting of paralinguistic information is useful for various application in man-machine communication such as speech synthesis with rich expressions and recognition of paralinguistic information in spontaneous speech. A speaker controls prosodic features such as fundamental frequency, power and temporal structures to express paralinguistic information. It is said that there are few speech rate variations in Japanese read speech. In spontaneous conversational speech, however, a speaker sometimes controls speech rate greatly to obtain a listener's attention. We previously found that speech rate of important words or portions of sentences is slowed to obtain the listener's attention [4]. In order to understand paralinguistic information using a computer, it is one of important issues to detect portions of sentences in which the speaker intentionally decelerates the speech rate. There are several studies on local speech rate variation [5–7]. However, there are few studies on detection of local speech rate variation.

We try to detect a local slower portion from a time series of mora duration [4,8]. When the speech rate of one portion is slower than that of other portions, the mora duration is longer than the durations of other morae. However, it is known that variation in time series of mora duration is caused not only by intentionally controlled speech rate variation but also by other factors such as difference of phonemes, length of a phrase or a sentence and a position of a mora in a phrase or sentence [9]. We have proposed the variable threshold (VT) [8] for detecting a local slower portion decelerated by a speaker from observed mora duration. The VT is applied to time series of mora duration. A mora whose duration exceeds the VT is detected as a local slower portion. The outline of the VT is described in section 2. In this paper, we examine the properties of parameters in the VT that are used for determining range and speed of variation of the VT. Three sets of parameters are prepared. We assume that these sets of parameters correspond to the levels of a listener's attention to utterances. We apply the VT to several sentences of Japanese spontaneous conversational speech. The detected portions are compared to the portions that human perceives slowness. Then we investigate the differences in detected portions using these three sets of parameters.

2. Threshold for detection of local speech rate deceleration

2.1. Detection process

A flow of the detection process is shown in Fig. 1. The first step is to calculate mora duration from speech signals. Calculation of the mora duration requires determination of the mora boundaries. The mora boundaries cannot be determined at several portions, e.g., at long vowels, diphthongs, double consonants and at portions with strong coarticulation. In such cases, plural morae are treated together with averaged mora duration.

At a next step, the mora duration adjusting factor (MDAF) [4,10] is applied to mora duration to obtain adjusted mora duration (AMD). Several kinds of morae have irregularly short durations. The MDAF is applied to mora duration to modify such fluctuation, which is mainly caused by phonemic nature. In this paper, the MDAF is applied to moraic nasal, long vowels, double consonants and diphthongs. By applying the MDAF to mora duration, adjusted mora duration (AMD) is obtained.

Finally, the VT is applied to the AMD to detect decelerated portions.

2.2. Outline of the variable threshold

When mora duration is greatly lengthened by a speaker's intentional control of speech rate, the listener would perceive the lengthening as a remarkable deceleration of the speech rate. Such large lengthening of mora duration seems to be "a caution signal" to obtain the listener's attention.

In this study, we design a threshold for detection of local slower portions in Japanese spontaneous conversational speech. A hypothesis of a listener's perceptual threshold of speech rate deceleration is introduced into our detection process. The hypothesis is that a listener's threshold for perceiving slowness of speech rate depends on durations of preceding morae in time series of the utterance. When mora duration exceeds the threshold, a listener perceives local lengthening of the mora duration (i.e., slowdown of speech

[*] e-mail: takamaru@media.eng.hokudai.ac.jp

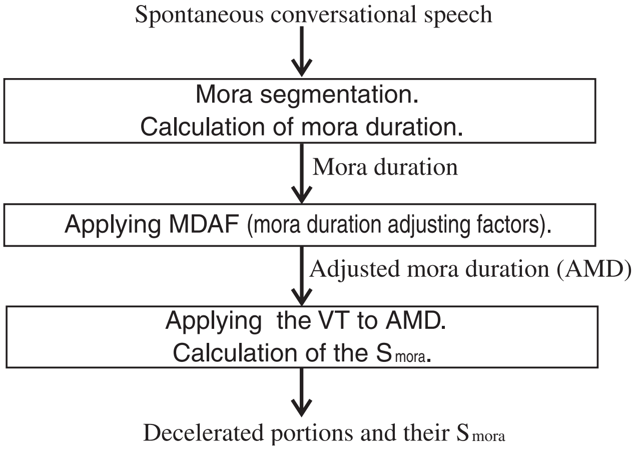

Spontaneous conversational speech
Mora segmentation.
Calculation of mora duration.
Mora duration
Applying MDAF (mora duration adjusting factors).
Adjusted mora duration (AMD)
Applying the VT to AMD.
Calculation of the S mora.
Decelerated portions and their S mora

Fig. 1 Flow of the detection process.

rate). Then the listener gradually increases the value of the threshold because of adaptation to the current lengthened duration (i.e., decelerated speech rate). When mora duration is shorter than the threshold, a listener slowly decreases the value of the threshold since the shorter duration should be employed as a new threshold for the next utterance. That is, the listener's threshold for perceiving slowness is adapted to the current speech rate.

According to the hypothesis, a threshold for detection of a local slower portion should dynamically vary depending on mora duration. The threshold should change slowly since a listener cannot rapidly change his/her perceptual threshold.

Thus, we have proposed the variable threshold (VT) [8] for detection of a local slower portion by the speaker's intentional control of speech rate. The VT is designed to be equipped with the features mentioned above. Examples of variation of the VT are shown in Fig. 2. The value of the VT increases up to the current mora duration when the mora duration exceeds the current value of the VT (Fig. 2(a)), and the VT decreases down to the current mora duration when the mora duration is shorter than the current value of the VT (Fig. 2(b)). A portion in which an AMD exceeds the VT is detected as a slower portion. An area of an AMD sequence exceeding the VT curve in the detected portion (hatched area in Fig. 2(a)) is considered to represent the degree of local slowness. We call the area value at a detected mora "S_{mora}."

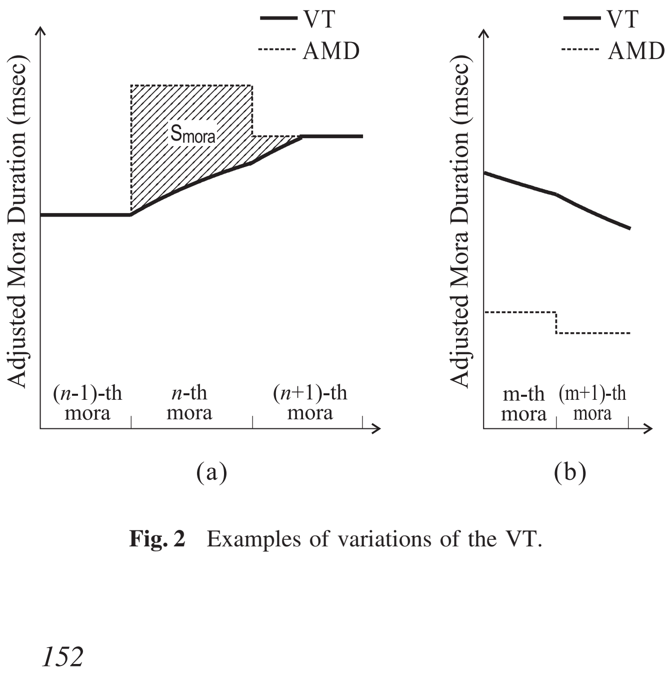


Fig. 2 Examples of variations of the VT.

2.3. Functions for the variable threshold

We introduce a set of functions into the VT. The VT is applied to the current mora, i.e., the N-th mora. AMD^n represents the AMD of the current N-th mora. V^n_{init} is the value of the VT at the beginning of the N-th mora. The V^n_{init} is equal to the value of the VT at the end of the $(n-1)$-th mora. $V_n(t)$ is a function of the VT within the N-th mora. The $V_n(t)$ is defined as follows:

1) When $\mathrm{AMD}^n > V^n_{\mathrm{init}}$:

$$V_{\mathrm{temp}}(t) = V^n_{\mathrm{init}} + A_{\mathrm{u}}\left\{1 - \exp\left(-\frac{t}{\tau}\right)\right\}$$

$$V^n(t) = \begin{cases} V_{\mathrm{temp}}(t) & V_{\mathrm{temp}}(t) < \mathrm{AMD}^n \\ \mathrm{AMD}^n & V_{\mathrm{temp}}(t) \geq \mathrm{AMD}^n. \end{cases} \tag{1}$$

2) When $\mathrm{AMD}^n \leq V^n_{\mathrm{init}}$:

$$V^n(t) = V^n_{\mathrm{init}} + A_{\mathrm{d}}\left(\mathrm{AMD}^n - V^n_{\mathrm{init}}\right)\left\{1 - \exp\left(\frac{t}{\tau}\right)\right\}, \tag{2}$$

where τ is a parameter that determines slowness of variation of the VT, and A_{u} and A_{d} are parameters that determine the degrees of increase and decrease in the VT.

2.4. Settings of the parameters

There are four parameters for a VT, i.e., A_{u}, A_{d}, τ, V^1_{init}. V^1_{init} is the initial value of the VT at the beginning of the first mora in a speech sample. V^1_{init} has an effect on detection at only the first several morae and does not affect detection at the following morae since the motions of the VTs for different values of V^1_{init} converge to a unique motion after the first several morae. In this paper, V^1_{init} is set to 100 [ms/mora] in consideration of average mora duration of recent Japanese speech. The τ is set as follows:

$$\tau = \begin{cases} 400 & \text{first time that } V^n_{\mathrm{init}} > \mathrm{AMD}^n \\ 200 & \text{otherwise.} \end{cases} \tag{3}$$

This is the same setting as our previous study [8]. A_{u} and A_{d} are parameters that determine the degrees of increase and decrease in the VT. We assume that the values of A_{u} and A_{d} can express the degree of the listener's attention to the utterance. Three kinds of settings shown in Table 1 are prepared. A VT with the setting of "param.1" is called "VT$_{\mathrm{param.1}}$." Similarly, "VT$_{\mathrm{param.2}}$" is a VT with "param.2" and "VT$_{\mathrm{param.3}}$" is a VT with "param.3." The setting "param.1" is the same as that in our previous study [8]. In the setting "param.1," the VT increases 26 [ms/mora] per typical one-word-length, i.e., about 400–500 [ms]. This is based on the results of a study on human perception of word-

Table 1 Three settings of parameters.

	A_{u}	A_{d}
"param.1"	26	1.0
"param.2"	75	0.3
"param.3"	13	1.0

based local speech rate [11]. According to [11], the differential limen on perception of lengthening of word-averaged mora duration is 26 [ms/mora]. In "param.2," A_u is larger than that in "param.1" and A_d is smaller than that in "param.1." Thus, the $VT_{param.2}$ tend to have a high value. We consider that the $VT_{param.2}$ expresses a situation in which the listener does not care much about speech rate variation. Only large variations should be detected by $VT_{param.2}$. In "param.3," A_u is smaller than that in "param.1." The value of $VT_{param.3}$ does not increase greatly even at a highly lengthened mora. We consider that $VT_{param.3}$ expresses a situation in which the listener nervously concentrate his/her attention on lengthening of mora duration, so that the listener keeps his/her threshold very low.

3. Experiments

3.1. Applying the VT to speech samples

We apply the VTs with three kinds of parameters to 10 sentences of spontaneous conversational speech. Figure 3 shows an example of the detection of a local slower portion by the VTs. The $VT_{param.2}$ detects only portions lengthening largely. The number of detected portions and their S_{mora} were increased in the order of $VT_{param.2}$, $VT_{param.1}$ and $VT_{param.3}$.

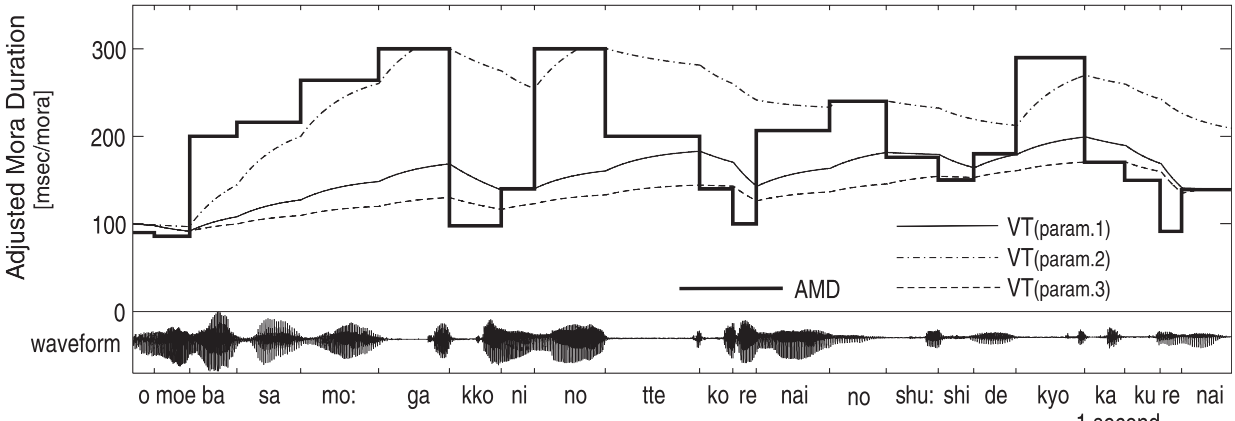
3.2. Auditory tests

We carry out several auditory tests to confirm portions in which human perceives slowness. The speech samples are 10 sentences of Japanese spontaneous conversational speech to which we applied the VT in the previous section. The subjects are 6 male university students. Three kinds of tests (test 1, 2 and 3) are carried out. In the test 1, the subjects listen to the speech samples once through a loudspeaker, and they point out portions where they feel the rate of the portion to be slower. The subjects are instructed to relax while listening to the speech samples. In tests 2 and 3, the subjects listen to the speech samples through headphones. They are instructed to listen attentively to the speech samples, and are allowed to listen to the samples repeatedly by their own operation. In test 2, the subjects are asked to listen to only phrase-final mora and to decide whether each phrase-final mora is lengthened or not. In test 3, the subjects are asked to listen to the speech samples except for the phrase-final morae and to point out portions where they feel the rate of the portions to be slower than rates of other portions.

In Fig. 3, the table under the graph shows the number of subjects who perceive lengthening at each mora in the example. The upper row shows the results by listening via a loudspeaker (test 1). The lower row shows the results by listening via headphones (mixed results of tests 2 for phrase-final mora and test 3 for non phrase-final mora). We can find that "**ba sa**" (the latter half of the first phrase "**o mo e ba sa**"), "**no tte**" and "**na i no**" in the example are perceived as slower portions by most subjects. Although there is a slight tendency for more portions to be perceived in the tests via a headphone than the test via a loudspeaker, there is no obvious difference between the results of the test via a loudspeaker and the results of the tests via headphones. Even in the case of the tests via a loudspeaker, the subjects have strongly concentrated on "local slower portions."

4. Comparison of detected portions and listener's perception

We consider that a speaker controls speech rate phrase by phrase since a phrase carries a specific meaning. At the portion that majority of listeners perceive lengthening or slowness of speech rate, values of S_{mora} should be large. For evaluation, we calculate the summation of S_{mora} in each phrase without phrase-final mora. We call this summation S_{phrase}. Then we compare the value of S_{phrase} with the total number of subjects who perceive slowness within the phrase (N_{phrase}.) We have classified the phrases according to their values of S_{phrase}. In each class, the average N_{phrase} is calculated by dividing the summation of N_{phrase} of the phrases by the number of phrases in the class. Figure 4 shows the results in the case of the $VT_{param.1}$. Few subjects perceive slowness in a phrase in which S_{phrase} is small. The average N_{phrase} increases as S_{phrase} increases. A similar tendency are found between N_{phrase} and S_{phrase} in the cases of the $VT_{param.2}$ and the $VT_{param.3}$. Thus, it seems that a phrase in which the majority



Number of subjects who perceive slowness or lengthening at the mora

via a loudspeaker	1	2	6	4	0	0	0	0	5	4	0	0	1	1	1	1	2	0	0	0	0	0
via headphones	0	0	3	6	3	0	0	0	6	6	1	1	3	3	1	0	3	0	2	0	0	0

Fig. 3 An example of the detection of local deceleration.

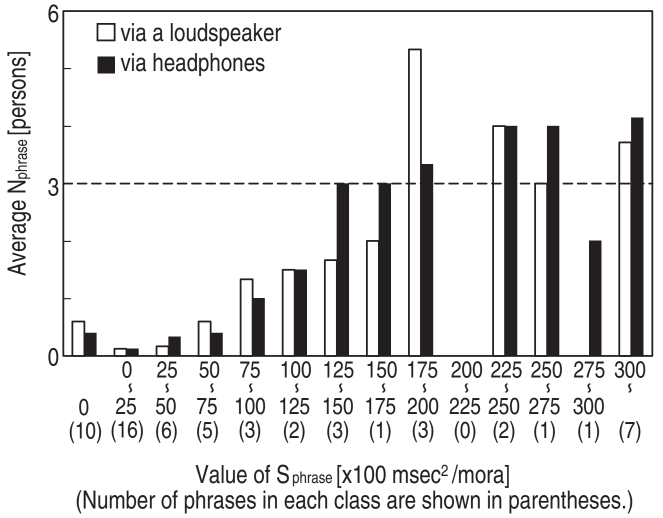


Fig. 4 S_{phrase} by $\mathrm{VT}_{\mathrm{param.1}}$ and the number of subjects who perceive a local slower portion in the phrase.

Table 2 S_{T} that $N_{\mathrm{average}}(S_{\mathrm{T}})$ exceeds 3.

	via a loudspeaker	via headphones
"param.1"	8,000	7,000
"param.2"	3,000	1,000
"param.3"	11,000	11,000

of listeners perceive local slowness has large S_{phrase}.

Next, we examine the average N_{phrase} of the phrases in which the values of S_{phrase} are greater than a specific value S_T. Namely, the summation of N_{phrase} of the phrases in which the S_{phrase} are over S_{T} is divided by the number of the phrases. We call the average value $N_{\mathrm{average}}(S_{\mathrm{T}})$. Table 2 shows the values of S_{T} when $N_{\mathrm{average}}(S_{\mathrm{T}})$ is over 3, i.e., over half of the total number of the subjects (6 subjects). In the cases of the $\mathrm{VT}_{\mathrm{param.1}}$ and the $\mathrm{VT}_{\mathrm{param.3}}$, the S_{T} are about 7,000–11,000 [ms^2/mora]. These results agree with those obtained in our previous study [8]. The results of the auditory tests, however, differ depending on the subjects. In the detected portions in which the N_{phrase} are less than 3, the judgement of slowness by listeners should depend on the degree of their attention to the utterance and other unknown factors. In the case of $\mathrm{VT}_{\mathrm{param.2}}$, the S_{T} require to maintain N_{phrase} over 3 are about 1,000–3,000 [ms^2/mora]. These values are smaller than those of other settings. This means the $\mathrm{VT}_{\mathrm{param.2}}$ detects only large speech rate variations, which most listeners can perceive.

We have confirmed that the number of detected portions and their S_{phrase} can be controlled by setting the parameters for the VT.

5. Conclusions

In this paper, we have studied the parameters of the VT for detecting local slowness of speech. Three kinds of parameter for the VT have been prepared. The VTs using these parameters have been applied to 10 sentences of Japanese spontaneous conversational speech. The portions detected by the VTs and the slower portions perceived by the subjects have been compared. It have been shown that the VT can detect a local slower portion in Japanese spontaneous conversational speech appropriately by adjusting the parameters of the VT.

We need to apply the VT to more Japanese spontaneous conversational speech. Then we have to adjust the parameter for more appropriate detection of local slower portions.

References

[1] H. Fujisaki, "Prosody, models, and spontaneous speech," in *Computing Prosody*, Y. Sagisaka *et al.*, Eds. (Springer-Verlag, New York, 1997), Chap. 3, pp. 27–42.

[2] K. Hirose, "Para- and non-linguistic information in speech information processing," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 243–246 (2002).

[3] K. Maekawa, "Issues in the study of paralinguistic information," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 247–250 (2002).

[4] K. Takamaru, M. Hiroshige, K. Araki and K. Tochinai, "A proposal of the model to extract Japanese voluntary speech rate control," *Proc. ICSLP 2000*, Vol. III, pp. 654–657 (2000).

[5] H. R. Pfitzinger, "Local speech rate perception in German speech," *Proc. ICPhS 1999*, Vol. 2, pp. 893–896 (1999).

[6] K. Hirose and H. Kawanami, "Temporal rate change of dialogue speech in prosodic units as compared to read speech," *Speech Commun.*, **36**, 97–111 (2002).

[7] S. Ohno and H. Fujisaki, "Quantitative analysis of the local speech rate and its application to speech synthesis," *Proc. ICSLP '96*, Vol. 3, pp. 2254–2257 (1996).

[8] K. Takamaru, M. Hiroshige, K. Araki and K. Tochinai, "Detecting Japanese local speech rate deceleration in spontaneous conversational speech using a variable threshold," *Proc. Eurospeech 2001*, pp. 935–938 (2001).

[9] Y. Sagisaka and Y. Tohkura, "Phoneme duration control for speech synthesis by rule," *Trans. Inst. Electron. Commun. Eng. Jpn.*, **J67-A**, 629–636 (1984).

[10] K. Takamaru, K. Suzuki, M. Hiroshige and K. Tochinai, "A study on the expression of speech rate for local speech rate analysis of spontaneous conversational speech," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, pp. 285–286 (1999).

[11] M. Hiroshige, K. Suzuki, K. Araki and K. Tochinai, "On perception of word-based local speech rate in Japanese without focusing attention," *Proc. ICSLP 2000*, Vol. 3, pp. 255–258 (2000).