Detecting Japanese Local Speech Rate Deceleration in Spontaneous Conversational Speech Using a Variable Threshold

Keiichi Takamaru, Makoto Hiroshige, Kenji Araki and Koji Tochinai

Graduate School of Engineering,
Hokkaido University, Sapporo, Japan
takamaru@media.eng.hokudai.ac.jp

Abstract

The variable threshold(VT), which detects the speech rate deceleration, is proposed. The VT varies dynamically depending upon the duration of previous mora in the utterance. The VT should not change rapidly because listener cannot perceive small variations of mora duration. Thus, a set of functions with time constants which decide response speed of the VT is introduced. We apply the VT to six sentences of spontaneous conversational speech. The auditory test of detecting local speech rate deceleration is carried out for the evaluation. The possibility of detecting the local speech rate deceleration by the VT is indicated.

1. Introduction

Spontaneous conversational speech contains a lot of paralinguistic information[1], *i.e.*, speaker's attitude, emphasis, intention and so on. A speaker will control fundamental frequency, power and temporal structure to express paralinguistic information. Though we can observe many kinds of variations in prosodic feature by acoustic analysis, not all variations correspond to paralinguistic information. Sophisticated extraction of ingredients of speakers' intentional control is required to utilize the paralinguistic information for the better man-machine communication.

Speech rate variations which are occurred with prominence are fewer in Japanese than other prosodic features (fundamental frequency and power). Because of this rareness, however, the speech rate variation should be considered as a strong "caution signal" to call listener's attention, so that such strong "caution signal" should not be missed even in man-machine communication. Observing our usual conversation, we can easily confirm the existence of the local speech rate variations with speakers' intentional control.

From these points of view, we are aiming to detect the portions of intentional local speech rate variations in Japanese spontaneous conversational speech. When intentional local speech rate variation can be detected, the speech communication system can consider that there is possibility to contain paralinguistic information in the rate-varied portions. The possibility of the existence of paralinguistic information can be utilized to linguistic analysis of speech contents, and also utilized to establish warm communication between human and machine.

There are already several methods which express a speech rate. One method[4] using DTW needs reference speech, so that they are difficult to use for large amount of spontaneous conversational speech. Other methods[2][3] to decide precise rate variation for speech synthesis may not express speakers' intentional rate control. Different from these studies, our study aims to detect speech portions with intentional large variations from large amount of spontaneous conversations. Thus, expression of speech rate and its variation should be different from the other studies.

We have proposed a speech rate model[5], which have been designed to express a global tendency of speech rate(driving force) and a local transformation of speech duration(damming force) separately. This model had aimed to express speaker's control of local speech rate. However, when this model is applied into actual speech rate variation data, it is difficult to separate meaningfully the global tendency element and the local transformation element. Considering appropriate separation of these two elements, we introduce one kind of threshold operation hinted by listeners' perception of irregularity. In other words, there may be local transformation at the portion where the listener feels irregular rate variation.

In this paper, we propose the variable threshold(VT) to detect the local speech rate deceleration. The threshold is compared with variation of mora duration which is obtained from mora segmentation.

In section 2, the explanation of the VT and the process to detect a portion of local speech rate deceleration are described. In section 3, we carry out an auditory test on perception of local speech rate deceleration in Japanese spontaneous conversational speech. Then we apply the VT to those speech samples. A comparison between the human perception and the portion detected by the VT is discussed in section 4.

2. Variable threshold(VT)

2.1. Preparation of adjusted mora duration(AMD)

To obtain mora duration from speech signals, mora boundaries are needed. In cases that the mora boundary cannot be determined(long vowels, diphthongs, double consonants and strong coarticulation), plural morae are treated together with an averaged mora duration calculated within the morae.

As a pre-processing, MDAF(Mora Duration Adjusting Factor)[5] is applied to several kinds of mora, *i.e.*, moraic nasal, long vowels, double consonants and diphthongs., which have irregularly short duration. After applying the MDAF, we get a time series of adjusted mora duration(AMD).

2.2. Basic concept of VT

Mora duration should be lengthened when the local speech rate becomes slower. In some cases of local speech rate deceleration, duration of the whole phrase which carries a meaning become longer uniformly. But in most cases, a

portion within the phrase are lengthened largely by the speaker's control of decelerating. If an average value is calculated within a phrase, the rate deceleration with such a partial lengthening may disappear (*e.g.*, *A* in Fig. 1). Usually, the convex portions in the utterance can be considered as slower portion, but it is not always slower portion. For instance, *B* in figure 1 is a convex portion with a duration over the averaged rate, but it should not be slower portion.

To distinguish a portion of speech whose rate becomes slower locally, several kinds of threshold operations are required. Since it seems that the listener's perceptual standard of local speech rate deceleration depend on the duration of past morae, the threshold value should vary dynamically depending upon the duration of past morae. However, the threshold should not change rapidly by each mora because listener cannot perceive small variations of mora duration. Thus, we propose the variable threshold (VT) which has the features mentioned above to detect the local speech rate deceleration. The threshold value increases when the mora duration exceeds the current value of threshold up to the current mora duration. While, the threshold decreases when mora duration becomes shorter.

2.3. Design

To express the VT, we introduce a set of functions with time constants which decide response speed of the VT. In the following equations, AMD^n is the AMD of the current *n*-th mora, $\mathrm{V}^n_{\mathrm{init}}$ is a VT value at the beginning of the *n*-th mora. The $\mathrm{V}^n_{\mathrm{init}}$ is equal to the VT value at the end of the (*n*-1)th mora. $\mathrm{V}^1_{\mathrm{init}}$ is set to be a constant. τ is a time constant. The VT function within the *n*-th mora $\mathrm{V}^n(t)$ is defined as follows(Fig.2):

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

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

$$\mathrm{V}^n(t) = \begin{cases} \mathrm{V}_{\mathrm{temp}}(t) & (\mathrm{V}_{\mathrm{temp}}(t) < \mathrm{AMD}^n) \quad (1a) \\ \mathrm{AMD}^n & (\mathrm{V}_{\mathrm{temp}}(t) \geq \mathrm{AMD}^n) \quad (1b) \end{cases}$$

where *A* is a sensitivity constant described below.

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

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

In this Paper, $\mathrm{V}^1_{\mathrm{init}}$ is set as 100[ms] considering global speech rate of our sample data. The *A* is set as 26[ms]. And τ is decided as follows:

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

Since averaged word duration can be roughly considered as about 400~500msec, the VT increases 26msec per one word longer mora by these settings, and the VT decreases down to the AMD for shorter mora. The value 26msec is selected considering our previous study, which declares that the differential limen for word-based local speech rate deceleration is 26msec[6].

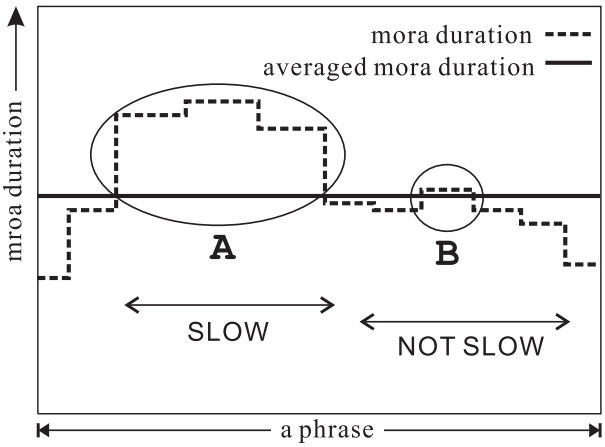

Figure 1: *An example of variation of mora duration*

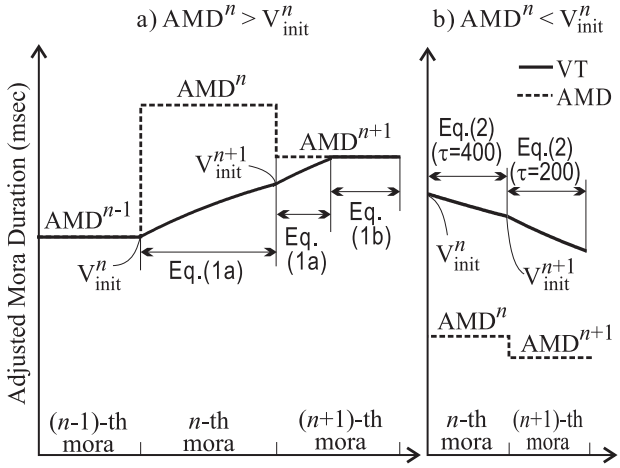

Figure 2: *Variation of the VT*

The VT is calculated based on the AMD. When the AMD exceeds the VT, we consider the speech become slower locally.

3. Applying the VT to spontaneous conversational speech

On detection of local speech rate deceleration, the portion which is detected by the VT should correspond to the portion which is perceptually detected by human. An auditory test on perception of local speech rate deceleration is carried out for the comparison with the portion which is detected by the VT.

3.1. Speech samples

Spontaneous conversational speech uttered by two male university students have been recorded for our research. Free conversations are recorded in their familiar place, *i.e.*, in their laboratory. Typical six sentences are selected to be used as speech samples in this study.

3.2. Auditory test

Five subjects participated in the test. They are requested to

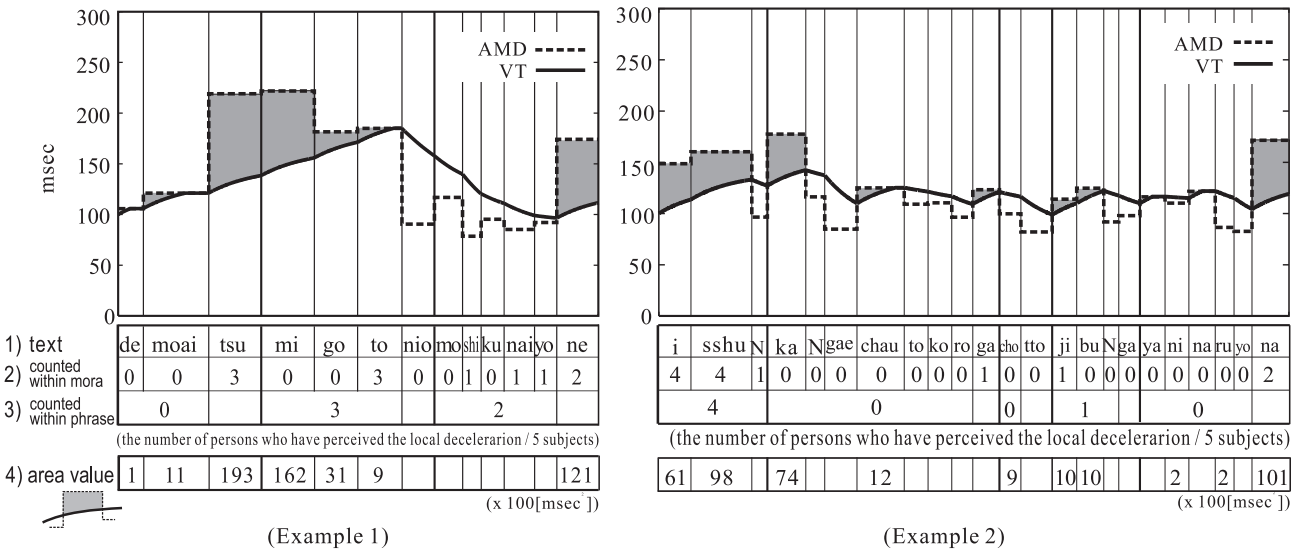

Figure 3: Examples of the VT applied to spontaneous conversational speech and the result of auditory test
The meaning of the utterance of the example 1 is "He is not amusing at all."
The meaning of the utterance of the example 2 is "I become to dislike myself since I let me think about it in a moment."

distinguish a slower portion in each speech sample by listening to the sample as often as they needed. Following three kinds of tests are carried out:

Test 1. Subjects detect the local rate deceleration and mark them into the transcribed text of the speech samples. The transcribed texts has no additional information such as a phrase boundary.

Test 2. Subjects check the lengthening of only phrase final mora and mark them into the transcribed text if they feel lengthening. Phrase boundary information is given to the transcribed texts so that the subjects easily find the phrase final mora.

Test 3. Subjects detect the deceleration within each phrase except for phrase final mora and mark them into the text which has phrase boundary information.

The test 1 is designed to avoid influence of given phrase boundaries in the transcribed text. Table 1 shows the percentage of disagreement between the position of marks of the three tests. Marks in test2 and test3 are mixed together, then compared with the marks in test1. The result of test 1 agrees with test 2 and 3 on approximately 90% of morae in the samples. We confirmed that that phrase boundary information does not hardly affect the result of the auditory test. Thus in the following section, we only discuss about the result of test 2 and 3.

Phrase final lengthening(PFL) is often observed remarkably in spontaneous speech. However, in many cases, PFL is considered to be a filled pause which contains no intentional control of speech rate. For the subjects, who are not experts of speech rate analysis, it is difficult to avoid the PFL(which is remarkable and less meaningful) and to concentrate rate variation in other portions(which may be slight but meaningful). Thus, we prepare two different kinds of auditory tests(*i.e.* test 2 and 3) to concentrate to PFL and the other portions separately.

3.3. Appling the VT to speech samples

We applied the VT to the six speech samples. Examples are shown in figure 3. The thick line in the graph is the VT and the dotted line is the AMD. The tables below the graph contain 1) the transcribed text of the utterance, 2) the number of persons who have perceived the deceleration of speech rate counted within each mora-based segment and 3) within each phrase (final mora is excluded from the phrase). In case the AMD exceeds the VT, an area value of a region enclosed by the VT and step-wise curve of AMD sequence is calculated(Fig.3(4)).

At the portion around "*mi-go-to*" in the example 1, three subjects perceived the local deceleration, and the AMD exceeds the VT. At the small convex portions "*mo*" and "*ku*", following the "*mi-go-to*", no subjects perceived the

Table 1: Disagreement between the result of test1 and that of Test 2, 3

Subject	MD	FK	HD	NS	NG	Average
Disagreement [%]	3.5	10.6	17.1	13.5	7.1	10.4

Table 2: The number of morae exceeded areas

α	The number of persons who perceive the deceleration	0	1	2	3	4	5
β	The numbers of all of segments	79	20	18	7	7	0
γ	The number of segments which exceed the VT	15	8	9	5	7	0
δ	Hit ratio $(\gamma/\beta \times 100)$ [%]	19.0	40.0	50.0	71.4	100.0	N/A

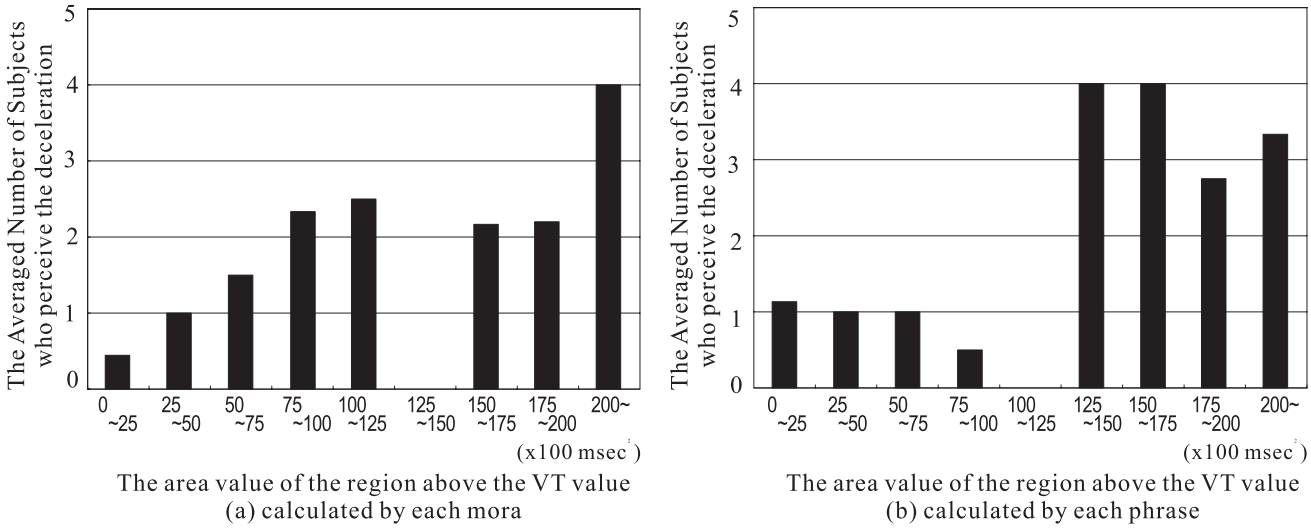

Figure 4: *The Averaged Number of Subjects who perceive the deceleration and the areas which exceeded the VT 12500~15000[msec^2] in (a) and 10000~12500[msec^2] in (b) have no data.*

deceleration, and the AMD are below the VT. The portion around "*i-sshu*" at the beginning of the example 2, four subjects perceived the local deceleration, and the AMD exceeds the VT. In these cases, the results of the detection by VT agree with the results of auditory tests.

However, no subject perceived the deceleration of local speech rate at the mora "*ka*" following the "*i-sshu-N*" although the AMD remarkably exceeds the VT.

In such cases, other prosodic features such as large variations of f0 or power are considered to affect the perception of speech rate.

At the phrase final mora "*ne*" in the example 1 and "*na*" in the example 2, the AMD exceed the VT. The decelerations of both morae are perceived by two subjects.

4. Analysis and discussions

Table 2 shows the number of segments which are perceived to contain deceleration(β) classified by the number of subjects who declare the existence of the deceleration(α). The table 2 also shows the number of segments at which the AMD exceeds the VT(γ). The hit ratio(δ) calculated by γ / β also shown in Table. 2. There are no portions which are consistently perceived to be slower portions by all five subjects. This means that detection of speech rate variation by mora is difficult for listeners. The hit ratio δ increases as the number of persons who declare the deceleration (α) increases. This result reveals a positive correlation between the perception of deceleration and the portion detected by the VT.

The averaged number of persons who perceive the deceleration classified by the area value of the region above the VT value are shown in Figure 4. Figure 4(a) is calculated by each mora and figure 4(b) is calculated by each phrase (phrase final mora is excluded). In figure 4(a), the number of persons who perceive the deceleration increases as the area value increases. On the other hand, in figure 4 (b), the number of persons who perceive the deceleration also increases as the area value increases, and the number of persons who perceive the deceleration are rapidly increases around 10000[msec^2]. In our speech samples, it seems that the detection threshold for local deceleration is around 10000[msec^2] of area value.

5. Conclusions

We have proposed a variable threshold(VT) to detect the local speech rate deceleration. This threshold takes into account the perception of local speech rate deceleration by listeners.

The VT has been applied to the six samples of spontaneous conversational speech. The areas of the AMD above the VT have been compared with the result of the auditory test. A positive correlation between the perception of deceleration and the portion detected by the VT have been confirmed, so that the VT have possibility to detect the local decelerated portion of speech rate.

Carrying out more auditory test and apply the VT to more speech samples, and refinement of the parameter in the VT to achieve more accurate detection are future issues.

6. References

[1] H.Fujisaki: "Prosody, Models, and Spontaneous Speech", In Y.Sagisaka *et al.*(ed.) Computing Prosody, Springer, pp.27-42 (1997)

[2] H.R.Pfitzinger: "Local Speech Rate Perception in German Speech", Proc. of ICPhS 1999, vol.2, pp.893-896 (1999)

[3] K.Hirose, H.Kawanami: "On the relationship of speech rates with prosodic units in dialogue speech", Proc. of ICSLP '98 (1998)

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

[5] K.Takamaru, M.Hiroshige, K.Araki and K.Tochinai: "A Proposal of the Model to Extract Japanese Voluntary Speech Rate Control", Proc. of ICSLP2000, Vol.III pp.654-657 (2000)

[6] M.Hiroshige, K.Suzuki, K.Araki, K.Tochinai: "On Perception of Word-based Local Speech Rate in Japanese without Focusing Attention", Proc. of ICSLP2000, Vol.III pp.255-258 (2000)