Evaluation of the Method to Detect Japanese Local Speech Rate Deceleration Applying the Variable Threshold with a Constant Term

Keiichi Takamaru Makoto Hiroshige Kenji Araki and Koji Tochinai

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

ABSTRACT

We are aiming to detect local deceleration of Japanese spontaneous conversational speech. We have proposed the variable threshold (VT), which detects local speech rate deceleration from the sequence of time series of mora duration. In this paper, we add a constant term to the VT to detect local deceleration appropriately. The VT is applied to 167 samples of Japanese spontaneous speech taken from a spoken dialogue corpus. The results of the detection are compared with local decelerations which are perceived by a listener. The VT detects 64 phrases among whole 87 decelerated phrases. We confirm the reduction of the incorrect detection of non-decelerated portions by adding a constant term to the VT.

1. INTRODUCTION

In human communication, speech carries not only linguistic information but also paralinguistic information[1], namely, emphasis, intention, attitude and so on. A speaker controls prosodic features such as fundamental frequency, power and temporal structure to express the paralinguistic information. A listener perceives local changes of prosodic features and then understands the paralinguistic information. As the first step to catch paralinguistic information by a computer, local changes of prosodic features with speaker's intentional control have to be detected. A speaker sometimes controls speech rate locally. In Japanese spontaneous conversational speech, local deceleration of speech rate is observed at portions of speaker's thinking, emphasis, important words and so on. It is said that Japanese speech has few local rate variation. However, a listener should pay rather strong attention to the utterance because of rareness when the speaker decelerates speech rate intentionally.

The purpose of our study is to detect local deceleration of speech rate in Japanese spontaneous conversational speech. Speech rate in Japanese is conventionally measured by mora duration. Thus, we are aiming to detect local deceleration with speaker's intentional control from time series of mora duration. When intentional local speech rate variation can be detected, a speech communication system can consider that there is possibility to contain paralinguistic information in the rate-varied portions. To point out the portions which have 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 a human and a machine.

Several researches for representation of speech rate or segmental duration have been reported. Most of them are designed for speech synthesis (*e.g.,* [2][3]). Those methods can decide precise segmental duration. However, it seems that they are not suitable for detecting of intentional speech rate deceleration. The method using a DTW[4] can represent local variation of speech rate. However, it needs reference speech which is uttered neutrally. To detect local speech rate deceleration from large amount of speech, a simpler method should be needed.

We have proposed the variable threshold (VT)[7] to detect local speech rate deceleration. The VT is compared with variation of time series of mora duration which is obtained from mora segmentation. The VT is based on an assumption of listeners' perception of irregularity. In other words, a listener should consider that there may be speaker's intentional transformation of duration at the portion where the listener perceives large irregular rate variation.

In this paper, we try to improve the functions to express the VT. Based on the result of our previous study, a constant term is added to the functions.

We carry out the experiment to detect local speech rate deceleration from a spoken dialog database. We compare detected portions by the previous proposed VT and the new VT with a new constant term. Then, the results are discussed by classification based on the cause of the deceleration.

2. VARIABLE THRESHOLD (VT)

2.1. Flow of the Detection

A flow of the detecting process is shown in Figure 1. As pre-processing, we need to obtain mora duration from speech signals. Then, the mora duration adjusting factor (MDAF) is applied to mora duration to obtain adjusted mora duration (AMD).

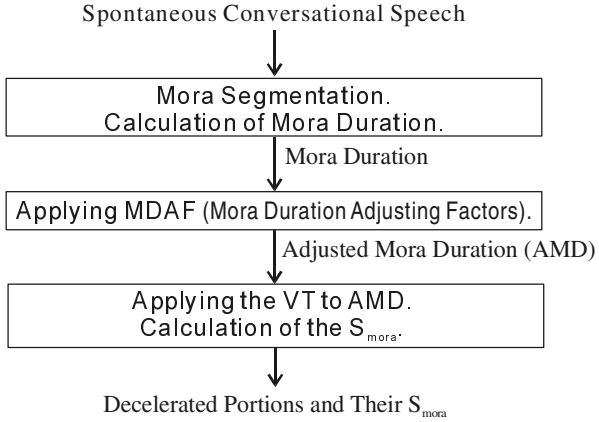

Fig. 1: The flow of the detecting process

2.1.1. Mora Duration

To obtain mora duration from speech signals, we have to know 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. We do not have to know all mora boundaries if only average mora durations are calculated in such portions.

2.1.2. Adjusted Mora Duration (AMD)

Several kinds of morae have irregularly short duration. The mora duration adjusting factor (MDAF)[5] is applied to mora duration values to modify such fluctuation, which is mainly caused by phonemic nature. In this paper, 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.

2.2. Basic Concept of the VT

Duration of mora should be lengthened when the local speech rate becomes slower. In several cases of local deceleration, durations of the whole phrase become longer uniformly. But in most cases, a particular portion within the phrase is largely lengthened by the speaker's control of deceleration. In Figure 2, there are 2 phrases, *i.e.*, phrase A and B. The phrase A has a mora which is lengthened largely. This is local a decelerated portion. While the phrase B has no particularly decelerated portion. However, average mora durations of those 2 phrases are not largely different. By averaging mora durations in a phrase, it becomes difficult to find partial deceleration within the phrase. Therefore, we cannot compare average mora durations in each phrase though it is one of the conventional methods.

To distinguish a portion of speech whose rate becomes slower locally, a threshold operation is required. Since it seems that a listener's perceptual standard of local speech rate deceleration depends on the durations of past morae, the threshold should vary dynamically depending on them. Since a listener cannot perceive small variations of mora duration, the threshold should not change rapidly according to rapid variation of each mora duration. Thus, we have proposed the variable threshold (VT) which has the features mentioned above to detect the local speech rate deceleration. A value of a threshold increases when the mora duration exceeds the current value of the threshold up to the current mora duration. While, the threshold decreases when mora duration becomes shorter.

Portions that the AMD sequence exceeds the VT are local decelerated portions detected by the threshold operation. In this paper, "S_{mora}" represents the area surrounded by the AMD sequence and the VT curve where the AMD exceeds the VT (shown by the shaded area in Figure 3.) We consider the "S_{mora}" area represents degree of the local deceleration created by the speaker. Figure 3 shows several examples of the VT.

2.3. VT Functions

2.3.1. Functions

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, V^n_{init} is a value of the VT at the beginning of the n-th mora. The V^n_{init} is equal to a value of the VT at the end of the (n-1)th mora. V^1_{init} is set to be a constant. τ is a time constant.

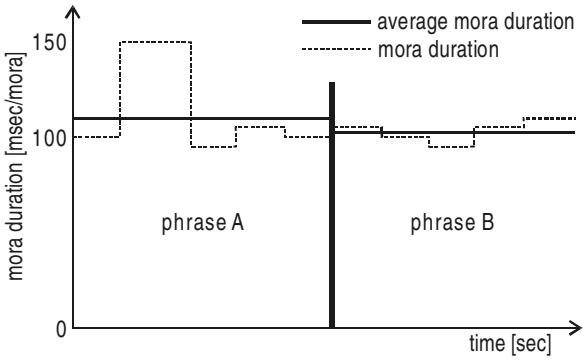

Figure 2: An example of average mora duration

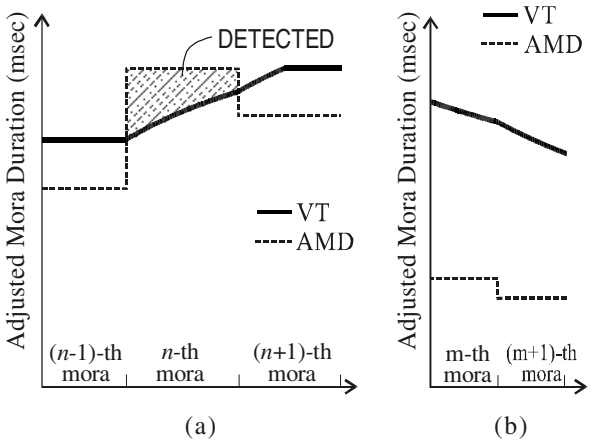

Figure 3: Variation of the VT

k is a constant term that is newly added to the functions in this report. The VT function within the n-th mora $V^n(t)$ is defined as follows:

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

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

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

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

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

where A_u and A_d are sensitivity constants of ascending and descending of the VT.

2.3.2. Adding a constant term

In our previous report[7], the VT without a constant term k has been proposed. (It is equivalent to k=0. Thus, we call it $VT_{k=0}$.) The $VT_{k=0}$ have applied to 10 sentences of spontaneous conversational speech. We have often observed incorrect detection of the non-decelerated portions. The S_{mora} of such portions are not so large. Therefore, we consider that the VT should not detect such non-decelerated portions by adding a constant term to the VT functions. We try to detect local deceleration appropriately.

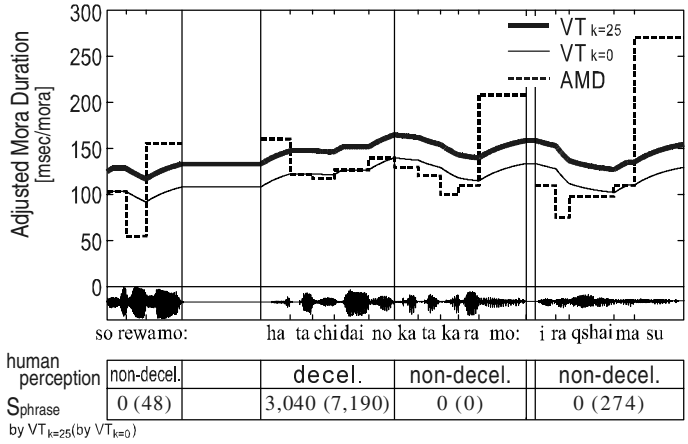

Figure 4: An example of detecting of local deceleration in spontaneous speech.

Table 1: results of the informal auditory test and detected phrases by the VTs

	Human's perception	Detected phrases	
		$VT_{k=0}$	$VT_{k=25}$
Decelerated phrases	87	81	64
Non-decelerated phrases	467	324	156

2.3.3. Parameters

In this Paper, V^1_{init} is set as 100[msec] considering global speech rate of our sample data. A_u is set as 26[msec]. A_d is set as 1.0. τ is decided as follows:

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

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

In our previous experiment, the majority of subjects of the auditory tests have perceived local deceleration at the phrases which have more than 10,000[msec^2/mora] in the area. The average duration of the phrases is about 400[msec]. Therefore, a value of a constant term *k* which is set as 25[msec/mora]

3. EXPERIMENTS

3.1. Applying the VT to spoken dialogue corpus

Speech samples are taken from RWCP spoken dialogue corpus. 167 samples are selected from the corpus. Each sample has more than 1 second and more than 2 phrases. To determine phoneme boundaries, HMM based forced alignment is executed. Then we manually revise the mora boundaries to get more accurate boundaries.

Two kinds of VTs are applied to the speech samples. One is the VT whose constant term *k* is set to 25 ($VT_{k=25}$) and another is the $VT_{k=0}$. We compare detected portions by two sets of VTs. Then detected portions by the $VT_{k=25}$ are evaluated by a comparison with the results of the informal auditory test mentioned in the next section.

3.2. Informal Auditory Test

The informal auditory test on local deceleration is carried out. The subject is one of the authors. He has judged the portions of local deceleration. He also checks causes of the deceleration. The causes are roughly classified into three kinds of categories; 1) the speaker's thinking or hesitation during speaking, 2) lengthening of phrase final mora, 3) expressions to call listener's attention, to emphasize and so on.

3.3. Results

Two kinds of VTs have been applied to 167 samples of speech. There are 554 phrases in the samples.

The detection of local deceleration by the VT is carried out with a resolution of a mora. However, the result of the detection and that of the informal auditory test have to be compared by a resolution of a phrase which carries a meaning. We consider that a listener should not realize deceleration of a particular mora precisely. The listener should perceive deceleration of a particular mora to be deceleration of a phrase which include the decelerated mora. So we sum up the results of detection by the VT within each phrase, and the summed-up results are compared with the results of the auditory test. In the auditory test, subjects are asked to detect a deceleration within each phrase. The S_{phrase} that is summation of S_{mora} in a phrase except for the phrase final mora is calculate to represent the degree of the local deceleration in each phrase.

An example of the results is shown in Figure 4. In the table below the graph, the first row shows the result of informal auditory test. The "decel." means that the subject perceives deceleration at the phrase. The "non-decel." means that the subject does not perceive deceleration at the phrase. The second row shows the values of S_{phrase} detected by the $VT_{k=25}$ in each phrase. The values in the parentheses are S_{phrase} detected by the $VT_{k=0}$. The $VT_{k=25}$ detects only the second phrase in the example. It corresponds to the result of informal auditory test. While, the $VT_{k=0}$ detects first, second and fourth phrases. The $VT_{k=0}$ can detect decelerated phrase correctly, however, non-decelerated phrases are often detected by the $VT_{k=0}$. The lengthening of phrase final mora is detected at three phrases by both $VT_{k=0}$ and $VT_{k=25}$.

In the informal auditory test, the subject perceives local deceleration at 87 phrases among whole 554 phrases. Table 1 shows the number of the detected phrases by the VTs.

4. DISCUSSIONS

4.1. Comparison between two settings of the VT

As in Table 1, the $VT_{k=0}$ detects 81 phrases from 87 decelerated phrases. However, 324 non-decelerated phrases are detected incorrectly. Thus, in our previous report[7], we have calculated S_{phrase} for screening the results. Table 2 shows the number of the phrases whose S_{phrase} of detection by the $VT_{k=0}$ are more than 5,000[msec^2/mora] or 10,000[msec^2/mora].

The $VT_{k=25}$ detects 64 decelerated phrases and 156 non-decelerated phrases. Compared with $VT_{k=0}$, Incorrect detection decreases to half. The result of the detection by the $VT_{k=25}$ is equivalent to the result of $VT_{k=0}$ whose threshold S_{phrase} are more than 5,000[msec^2/mora]. Therefore, by adding a constant term to the VT, the decelerated phrases can be detected more appropriately without giving the phrase boundaries. However, The number of the decelerated phrases

Table 2: The number of phrases whose $VT_{k=0}$'s S_{phrase} are more than 5,000 and 10,000

	$S_{phrase} > 5{,}000$	$S_{phrase} > 10{,}000$
decelerated phrases	68	51
non-decelerated phrases	166	73

Table 3: Causes of deceleration at detected phrases by the $VT_{k=25}$

cause	frequency	average S_{phrase}
to call listener's attention (emphasis etc)	42	18,815
thinking or hesitation	22	190,044

Table 4: Phrase final mora

	frequency	average S_{mora}
perceived lengthening	106	57,140

detected by the $VT_{k=25}$ decreases slightly compared with $VT_{k=0}$. Further adjustments of parameters of the VT and a value of the constant term are needed to detect decelerated portions appropriately.

4.2. Classification by the area value

In this section, the phrases detected by the $VT_{k=25}$ are classified by the S_{phrase}. Figure 5 shows the relative frequencies of the decelerated and the non-decelerated phrases classified by the S_{phrase} of the $VT_{k=25}$. The S_{phrase} are 0[msec^2/mora] for 70[%] of non-decelerated phrases. However, the small amounts of non-decelerated phrases are also distributed to classes less than 10,000[msec^2/mora]. The perception of deceleration of the phrases whose S_{phrase} are more than 0[msec^2/mora] and less than 10,000[msec^2/mora] in the detection should depends upon listeners and degree of listener's concentration to the utterance. Therefore, we have to examine the human perception of local deceleration by carrying out more auditory tests.

4.3. Cause of Deceleration

The subject of the informal auditory test roughly classified causes of local deceleration into three categories, *i.e.*, the expression to call listener's attention (emphasis etc.), the thinking or hesitation during speaking and lengthening of phrase final mora. Table 3 and Table 4 show the results of classification. In the 42 phrases, the decelerations are classified into calling listener's attention. In the 22 phrases, the decelerations are from thinking and hesitation. In Table 3, the average S_{phrase} of the latter phrases (thinking and hesitation) are about ten times as large as that of former phrases (calling attention). Table 4 shows the number of the perception of the lengthening of phrase final mora. In 106 phrases, the lengthenings of phrase final mora are perceived. The average S_{mora} of them are larger than average S_{phrase} of the phrases decelerated to call listener's attention.

The local decelerations to call listener's attention are detected by the $VT_{k=25}$ appropriately. However, their S_{phrase} are quite smaller than the portions of thinking or hesitation. Thinking or hesitation mainly appears as variations of durations. A speaker should also vary various prosodic features to call listener's attention or to emphasize a phrase. Therefore, we have to examine variations of other prosodic features.

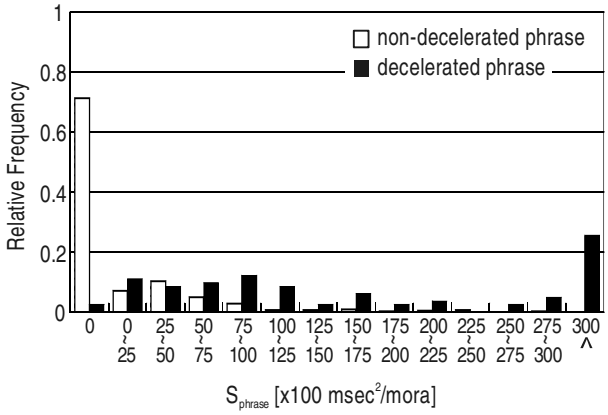

Figure 5: Relative Frequency of decelerated and non-decelerated phrases detected by the $VT_{k=25}$

5. CONCLUSIONS

We have proposed the variable threshold (VT) to detect Japanese local speech rate deceleration. Based on our previous study, a constant term has been added to the VT.
The VT has been applied to 167 sentences from spoken dialog corpus. We have compared the 2 sets of VTs, *i.e.*, the VT which has no constant term ($VT_{k=0}$) and the VT which has a constant term ($VT_{k=25}$.) Incorrect detection of non-decelerated phrases decreases to half (from 324 to 156.) Therefore we can conclude that adding a constant term to the VT is effective for appropriate detection. By the VT added a constant term, local decelerations are detected without giving the information of phrase boundaries beforehand.

Carrying out more auditory test to confirm perception by human and refining the parameters of the VT to achieve more accurate detection are future issues.

Acknowledgements This work was supported by a Grant-in-Aid for Encouragement of Young Scientists (13750373) from Japan Society for the Promotion of Science.

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: "Temporal rate change of dialogue speech in prosodic units as compared to read speech," Speech Communication, Vol.36, pp.97-111 (2002).
[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).
[7] K.Takamaru, M.Hiroshige, K.Araki, K.Tochinai: "Detecting Japanese local speech rate deceleration in spontaneous conversational speech using a variable threshold," Proc. of Eurospeech2001, pp.935-938 (2001).