A Fundamental Study on a Method to Detect Slower Phrases in Japanese Dialog Speech

Keiichi Takamaru†, Makoto Hiroshige†, Kenji Araki† and Koji Tochinai‡

†Graduate School of Engineering, Hokkaido University, Japan
‡Graduate School of Business Administration, Hokkai-gakuen University, Japan
{takamaru;hiro;araki}@media.eng.hokudai.ac.jp, tochinai@econ.hokkai-s-u.ac.jp

Abstract

A slower phrase in spontaneous conversational speech is caused by emphasis, thinking during speaking and so on. To include such useful information with man-machine communication, we investigate a method to detect local slower phrase from time sequence of mora duration in Japanese dialog speech. At first we prepare speech samples, which contains phrases slowed considerably. Then the flow of the process to obtain phrase averaged mora duration is explained. In this method, speech period, mora boundaries and phrase boundaries are obtained from acoustical features. A threshold is applied to phrase averaged mora duration. An experiment to detect a local slower phrase is carried out. The slowed phrases are detected with high recall rate.

1. Introduction

In human communication, speech conveys not only linguistic information but also other useful information, which cannot be put down as characters. In spontaneous speech, people control acoustical features such as fundamental frequency, power and temporal structure to express such information. They sometimes slow down their speech rate locally. It is caused by emphasis, thinking during speaking and so on. There has been some research on synthesizing speech which contains rich expressions by controlling fundamental frequency, power and temporal structure[1]. However such research to detect these features is little. It is said that Japanese speech has few local variations of speech rates. However we can observe some local large variations of speech rates in Japanese spontaneous conversational speech[2]. People can perceive the local variation of speech rate[3]. In our daily experience, it seems that the contrast of the speech rate in adjacent phrases has a strong ability to draw the listener's attention. Thus, it is important to detect a slower phrase. In this paper, we investigate a method to detect local speech rate variation. We try to detect a local slower phrase from time sequence of mora duration since Japanese speech rates are conventionally expressed by mora duration.

It is known that mora duration is varied by several factors such as the number of mora in a phrase, a position of a mora in a phrase and a type of contained phoneme in a mora[4]. The variation of mora duration by such factors is observed throughout in an utterance. Thus, such changes don't draw people's attention. On the other hand, a variation of mora duration to draw a listener's attention should be much larger than the variation by the factors mentioned above. We are aiming to detect such a large variation of mora duration. To detect a slower phrase, a phrase averaged mora duration (Pave) has to be obtained from acoustical features. Then we try to apply a threshold to each Pave. This paper describes the preparation of speech samples which contained slowed phrase, the flow of the process to obtain Pave, and the fundamental experiment to detect slowed phrases in the speech samples.

2. Speech Samples

We need a speech sample which contains a phrase which is slowed considerably to draw a listener's attention. Read speech or narrated speech do not seem to contain many large slowed phrases. Although spontaneous conversational speech would contain local slower phrases, it is difficult to carry out auditory tests through large amounts of spontaneous speeches to distinguish slowed phrases from other phrases. Thus, we record dialog speeches which contain slowed phrases by instructing a speaker to slow several specific phrases.

We prepare a dialog script. The script has several underlined phrases which should be uttered slowly to draw a listener's attention. The script is based on a transcription of a dialog[5] between a customer (role A) and a car dealer (role B). Figure 1 shows a portion of the script.

The speakers in the recording are two graduate students (speaker 1 and 2). They are native Japanese and have some acting experience. We instruct them to utter the script naturally and to slow only the underlined phrases. The utterances are recorded to the DAT with 48kHz 16bit sampling via headset microphones. The recording is carried out in a soundproof room. The recorded speech is down-sampled to 16kHz for the analysis. We record 5 dialogs with the same script. The roles are alternately changed in each recording. The first recording is for the practice. The second to fifth recordings are used as speech samples.

B: hai. kyo:wa do:mo iraqshaimase. (May I help you, today?)
A: kurumano kotowo oshiete itadakitaiNdesukeredomo.
(I would like you to tell me about a car.)
B: hai. (Sure.)
A: chiisakutemo fo:doa tokaqte aruNdesuyone.
(There is a car which has four doors in spite of a small sized car, isn't it?)
B: dono teidono chiisasaka kokusande iimasuto,
ta:serutoka arekuraino saizuni narimasuka.
(Figuratively speaking in a Japanese car, the size is same as *taaseru* or the like?)
A: so:desune. (That's right.)

Figure 1: *A portion of the dialog script (A: customer, B: car dealer)*

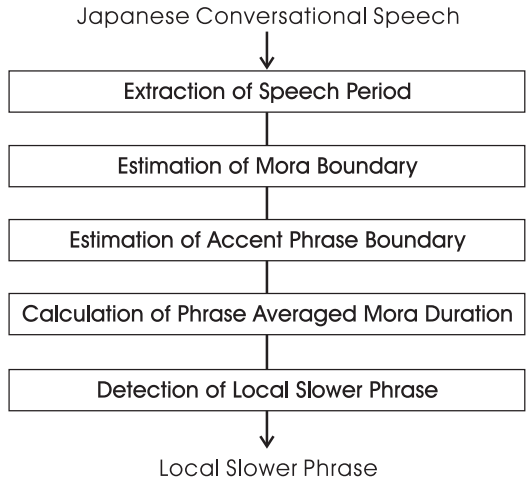

Figure 2: *Flow of the detection process*

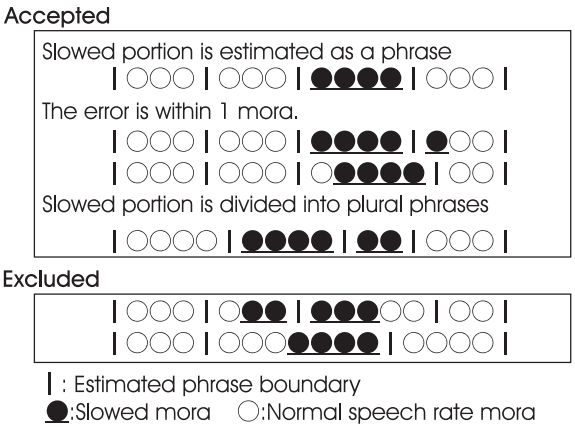

Figure 3: *Rule to exclude a speech period*

3. Flow of the detection process

Figure 2 shows the flow of the detection process. Speech periods, mora boundaries and phrase boundaries are estimated from acoustical features. Then phrase averaged mora duration (Pave) is calculated. Finally, a threshold to detect slowed phrases is applied to Pave.

3.1. Extraction of speech period

A speech period is estimated from log RMS power. Log RMS power is calculated as follows:

$$\text{Log RMS} = \log_{10} \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} x_i^2} \qquad (1)$$

where x_i is a windowed value of a speech signal, and N is a frame length. The analysis parameters are shown in Table 1.

Table 1: *Analysis parameter of log RMS power*

Frame length	1,000 [ms]
Frame shift	100 [ms]
Window function	Hamming

When log RMS power is over a specific value, the period seems to be a speech period. The value depends on the recording environment and loudness of an utterance. In this study, the threshold of log RMS power is set to 5.3. When an interval between an estimated speech period and the next one is within 1,000[ms], these periods are treated as one speech period since a short interval may be a silent period of a double consonant, a short pause in a sentence or the like.

The samples of speaker 1 has 110 speech periods and the samples of speaker 2 has 92 periods. Total durations are 457[sec] and 430[sec] respectively.

3.2. Estimation of mora boundary

To calculate duration of each mora, phoneme boundaries are needed. Ideally, we would like to obtain phoneme boundaries by using a speech recognition system. However it is difficult to recognize conversational speech with high accuracy so far. We cannot obtain enough results by using a speech recognition system. Thus, in this study, we use a forced alignment of phoneme sequences by using a HMM based phoneme segmentation tool[6] instead of a speech recognition system. In this case, a sequence of phoneme to align must be given to the segmentation tool. We give the transcription for each speech period. A position of a short pause is estimated automatically to make an acoustical likelihood highest.

3.3. Estimation of phrase boundary

A phrase boundary is estimated from fundamental frequency (F0). F0 is approximated by a broken line to have the lowest mean squared error[7]. A local minimal point of the line is estimated as a phrase boundary. In the F0 analysis of the dialog speech, there are a lot of extraction error. Therefore we introduce a reliability rate of each F0 for calculating mean squared error. F0 extraction is based on the auto-correlation method[8]. The peak value of the auto-correlation coefficient is used as a reliability rate. The mean squared error is calculated as follows:

$$\text{Mean squared error} = \frac{1}{N} \sqrt{\sum_{i=1}^{N} (\text{F0}_i \times \text{ac\_peak}_i^{\,n} - \text{L}_i)^2} \qquad (2)$$

where, F0_i is a value of F0 at the ith frame, ac_peak_i is an auto-correlation coefficient of ith F0, and L_i is a value of approximated line at the ith F0.

A position of a phrase boundary does not absolutely correspond to a position of a mora boundary. To obtain the number of morae in a phrase, a position of a phrase boundary should correspond to a position of mora boundary. Thus, a phrase boundary is adjusted to the nearest mora boundary.

We are aiming to detect slowed phrases. To do so, the slowed portion has to be separated as a phrase. When a slowed portion is not separated from other portions, the speech period is excluded in this study. An error within 1 mora is permitted (Figure 3). We exclude 22 and 43 speech periods from the samples of speaker 1 and 2 respectively. The

number of the estimated phrases used in the later detection experiment are shown in Table 2.

3.4. Calculation of phrase averaged mora duration

A phrase-final mora is excluded when phrase averaged mora duration (Pave) is calculated (Figure 4). A final mora in a phrase has higher flexibility than other morae. It is called filled pause or phrase-final lengthening. Lengthening of a phrase-final mora is observed throughout an utterance. Thus, it is usual for a listener and does not draw his/her attention. We have to treat a final mora in a phrase separately.

An example of the process of obtaining Pave is shown in Figure 5.

3.5. Detection of local slower phrase

A threshold is applied to the time sequence of Pave. The threshold in this study is the average of Pave in the first 60 seconds of the samples dialog.

Table 2: *The number of estimated phrases*

	speaker 1	speaker 2
The number of estimated phrase	148	151
The number of slowed phrase	27	30
Phrase duration [ms]	721.4	732.1

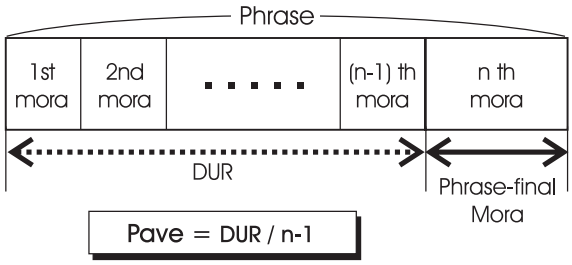

Figure 4: *Phrase averaged mora duration*

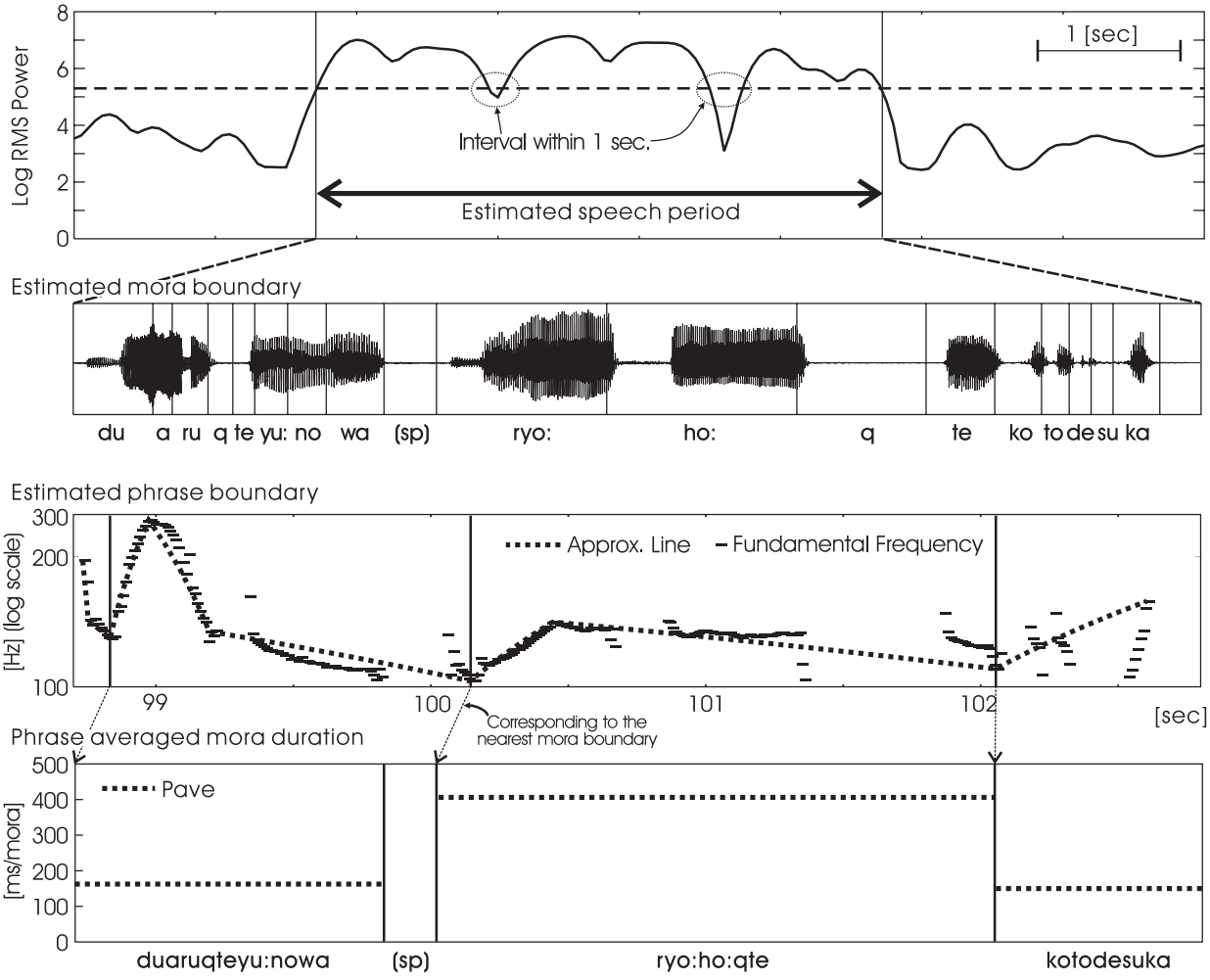

Figure 5: *An example to obtain phrase averaged mora duration*

4. Detection Experiment

4.1. Results

The threshold is applied to the time sequence of Pave uttered by speaker 1 and 2. Precision, recall and f-measure are calculated with the following equations:

$$\text{Precision} = \frac{\text{the number of detected slowed phrases}}{\text{the number of detected phrases}},$$

$$\text{Recall} = \frac{\text{the number of the detected slowed phrases}}{\text{the number of slowed phrases}},$$

$$\text{F-measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}}.$$

Table 3 shows the precision, the recall and the f-measure. The recall rate is very high rate in both speakers. The threshold has not failed to detect slowed phrases. However, the precisions are about 60[%]. That is to say, 40[%] of the detected phrases were not required to be uttered slowly.

Figure 6 shows all the result of the detection in the case of speaker 1. The white bars are normal phrases which were not instructed to be uttered slowly. The dark bars are the slowed phrases. All the slowed phrases are over the threshold. However we can observe that several normal rate phrases are also over the threshold. Thus, the optimum threshold seems to be higher than the threshold in this study.

4.2. Discussion

In figure 6, some normal phrases have longer Pave than the Pave of the shortest slowed phrase. Most of the misdetected phrases are isolated utterances, *e.g.*, short answers and back channels. It seems that a speaker has slowed such phrases although there are no instructions to slow the phrases. So we have to carry out an auditory test to check whether the phrases are slow. If such phrases are perceptually slow, they are not misdetection. Otherwise it may be impossible to distinguish the slowed phrases from normal phrases by using only a fixed threshold to Pave. Further study will be needed to decide an appropriate threshold. The other misdetection is caused by incorrectly estimating mora boundaries and phrase boundaries. To estimate these boundaries with high accuracy is an important future issue.

Table 3: *The results of the detection*

	Precision	Recall	F-measure
speaker 1	56.3 %	100.0 %	72
speaker 2	58.0 %	96.7 %	72.5

5. Conclusions

In this paper, we have studied a method to detect local slower phrases in Japanese dialog speech. At first we have prepared speech samples which contained considerably slowed phrases. Then the flow of the process to obtain phrase averaged mora duration (Pave) has been explained. Finally, the experiment to detect local slower phrases has been carried out. The average of Pave in the first 60 seconds is used as the threshold. The slowed phrases are detected correctly. However the precision has not been so high. We need further studies about the threshold and high accuracy estimation of mora boundaries and phrase boundaries.

6. References

[1] Iida, A.; Higuchi, F.; Campbell; N., Yasumura, M., 2002. A corpus-based speech synthesis system with emotion. *Speech Communication* 40(1-2), 161-187.

[2] K.Takamaru; K.Suzuki; M.Hiroshige; K.Tochinai, 1999. A study on several features of local speech rate variations in spontaneous conversational speech. *Proc. ITC-CSCC 99,* 390-393.

[3] M.Hiroshige; K.Suzuki; K.Araki; K.Tochinai, 2000. On perception of word-based local speech rate in Japanese without focusing attention. *Proc. ICSLP2000 Vol. III,* 255-258.

[4] Y.Sagisaka; Y.Tohkura, 1984. Phoneme Duration Control for Speech Synthesis by Rule. *Trans. the Institute of Electronics and Communication Engineers of Japan* .J67-A, 629-636.

[5] http://www.rwcp.or.jp/wswg/rwcdb/speech/

[6] http://julius.sourceforge.jp/ouyoukit.htm

[7] A.Komatsu; E.Ohira; A.Ichikawa, 1988. Conversational speech understanding based on sentence structure inference using prosodics and word spotting. *IEICE Trans. inf. & Syst.* J71-D(7), 1218-1228.

[8] D. Talkin, 1995. A Robust Algorithm for pitch tracking (RAPT). in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal (eds.). New York: Elsevier, 495-518.

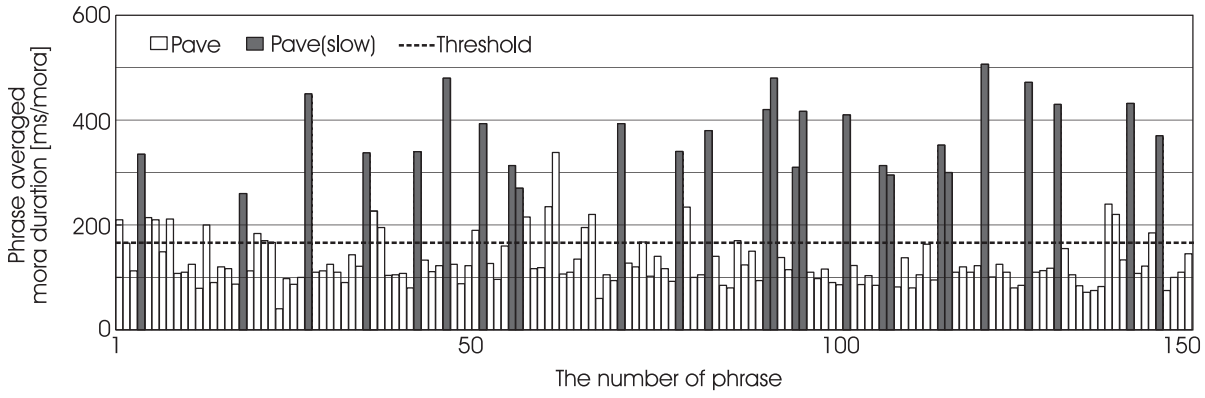

Figure 6:*The result of the detection (Speaker: 1, Threshold: the average of Pave in first 60 seconds)*