Evaluation of a threshold for detecting local slower phrases in Japanese spontaneous conversational speech

Keiichi Takamaru

Department of Urban Economics, Nasu University
takamaru@nasu-u.ac.jp

Abstract

I have proposed a method for detecting local slower phrases in Japanese spontaneous conversational speech. A threshold is applied to phrase-averaged mora duration in this method. It is considered that relative variation of time sequence of phrase-averaged mora duration should be taken into account for detecting slower phrases correctly. In this paper, preliminary experiments are carried out to obtain optimal threshold. Then the optimal threshold is applied to Japanese conversational speech. I have confirmed that the threshold in consideration of relative variation of phrase-averaged mora duration has better result than the simple straight threshold.

1. Introduction

Speech conveys not only linguistic information but also some useful pieces of information such as speaker’s intention, emphasis in human communication. They are not usually expressed in written words. Though speech recognition system cannot obtain such useful information so far, they must be also detected and used in speech understanding by computer. In spontaneous speech, human controls prosodic features such as power, fundamental frequency and temporal structure to express them. Recently, there are some researches to recognize or identify paralinguistic information[1]. This paper focuses on local variation of speech rates. People sometimes slow down their speech rates locally for emphasizing a specific portion, thinking during speaking and so on. Though it is said that Japanese speech has few local variation, the contrast of the speech rate in adjacent phrases has a strong ability to draw the listener’s attention in daily conversation. It is important to detect local changes of speech rate. Especially, local slower portions seem to contain more important information compared with local faster portions. Therefore I am trying to detect local slower portions in Japanese spontaneous conversational speech. I have proposed a method for detecting local slower portions in Japanese spontaneous conversations[2]. A threshold is applied to phrase-averaged mora duration in this method. However, in the previous study[2], only a simple straight line has been used as a threshold. To detect the contrast of speech rates, a relative variation of phrase-averaged mora duration must be taken into account for designing the threshold.

In this paper, the threshold for detecting local slower portions is proposed. At first, basic method for detecting slowed phrases is explained. Then I explain about a threshold and appropriate parameters for the threshold are decided in preliminary experiments. Finally, the threshold is compared with a simple straight threshold.

2. Method

Figure 1 shows the flow of the detecting process that I have proposed in [2]. Speech periods, mora boundaries and phrase boundaries are estimated from acoustical features. Then phrase-averaged mora duration is calculated from phrase duration and the number of morae in a phrase. Finally, a threshold is applied to phrase-averaged mora duration. Each process up to obtaining phrase-averaged mora duration is explained briefly in following subsection. The threshold operation is explained in next section.

2.1. Extraction of speech period

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

\[
\text{Log RMS} = \log_{10} \left( \frac{1}{N} \sum_{i=0}^{N-1} x_i^2 \right)
\]

where \(x_i\) is a windowed value of a speech signal, and \(N\) is a frame length. In this study, frame length is 1000[ms], frame shift is 100[ms] and a type of window function is Hanning.

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 1000[ms], these two periods are treated as one speech period since a short interval seems to be a silent period of a double consonant, a short pause in a sentence or the like.

![Flow of the detection process](image-url)
2.2. Estimation of mora boundary

To calculate duration of each mora, phoneme boundaries are needed. Ideally, phoneme boundaries should be obtained by using a speech recognition system. However, it is difficult to recognize conversational speech with high accuracy so far. Especially, in the case of containing large durational variation, recognition rates seem to be lower. Then, in this study, a forced alignment of phoneme sequences by using a HMM based phoneme segmentation tool[3] is used instead of a speech recognition system. A sequence of phoneme to align must be given to the segmentation tool. It simulates a case where a result of speech recognition is perfect.

2.3. 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[4]. Local minimal points of the line are estimated as phrase boundaries. In the F0 analysis of the dialog speech, there are a lot of extraction errors. Therefore we introduce a reliability rate of each F0 for calculating mean squared error. F0 extraction is based on the auto-correlation method[5]. The peak value of the auto-correlation coefficient is used as a reliability rate. The mean squared error is calculated as follows:

\[
\text{Mean squared error} = \frac{1}{N} \sum_{i=1}^{N} (F0_i - L_i)^2 \times \text{ac_peak}_i^n
\]

where, \(N\) is the number of F0 points within a period to calculate mean squared error. F0 is a value of F0 at the \(i\)th

![Figure 2: Calculation of phrase-averaged mora duration](image2)

![Figure 3: Example of obtaining phrase-averaged mora duration](image3)
slower phrase. The parameters, phrase is longer than this value, the phrase is detected as a mora duration. The threshold should be decided on the basis of a threshold is applied to the time sequence of phrase-averaged mora duration within previous and next phrases of a target phrase. I designed a threshold as follows:

$$\text{Threshold} = \text{Ave}[N_{\text{pre}}: N_{\text{post}}] + B$$  \hspace{1cm} (3)

where $N_{\text{pre}}$, $N_{\text{post}}$, and $B$ are decided in the preliminary experiment.

### 3.2. Speech samples

At first a dialog script is prepared. The script has several underlined phrases in which a speaker is instructed to utter slowly. The script is based on a transcription of a dialog[6] between a customer (role A) and a car dealer (role B).

The speakers in the recording are two graduate students (speaker 1 and 2). They are native Japanese and have some acting experience. They are instructed to utter the script naturally and to slow only the underlined phrases. The utterances are recorded to 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 experiments. 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.

As shown in Table 1, 148 speech periods uttered by speaker 1 and 151 speech periods uttered by speaker 2 are obtained from these speech samples. 27 and 30 slowed phrases are contained respectively. Average phrase durations are 721.4[ms] and 732.1[ms]. In the following experiments, the slowed phrases are the targets to detect.

### 3.3. Preliminary experiment for parameters

Parameters for the threshold, $N_{\text{pre}}$, $N_{\text{post}}$ and $B$, are decided in the preliminary experiment. Speech samples within first 30 seconds of utterance are used. The rest of samples are used for evaluation.

$N_{\text{pre}}$ and $N_{\text{post}}$ are changed from 2.5 to 30[s] every 2.5[s]. $B$ is changed from 0 to 50[ms/mora] every 5[ms/mora]. Precision rates, recall rates and F-measures are calculated in all these combinations. They are calculated with following equations:

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

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

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

Parameter sets are evaluated by F-measure. In those all combinations, the highest F-measure is obtained when $N_{\text{pre}}$ equals 5, $N_{\text{post}}$ equals 5 and $B$ equals 35. In this situation, F-measure was 0.875 for speaker 1 and 0.933 for speaker 2. $N_{\text{pre}}$ and $N_{\text{post}}$ are 5[s]. 5[s] contains about 7-8 phrases. It seems that it should be effective for a threshold to consider the contrast of adjacent phrases. It has been reported the differential limen of perception of local slower phrase is about 30[ms/mora] in phrase-averaged mora duration[6]. This almost corresponds to result of parameter $B$.

### 4. Evaluation

#### 4.1. Results

According to the preliminary experiments, the threshold for detecting slower phrase is Ave[5.5]+35, that is, 35 plus averaged mora duration of the phrases within previous and next 5[s] of the detection target.

This threshold is applied to speech samples without phrases used in preliminary experiments. Figure 4 shows the results for speaker 1 and 2. Table 2 shows precision rates, recall rates and F-measures. F-measures are 0.89 and 0.74 respectively. Recall rates are both 0.94. Almost all slowed phrases are correctly detected by the proposed threshold. Precision rates are 0.85 and 0.60. There are 15-40[%] of misdetections.

#### 4.2. Considerations

##### 4.2.1. Phrase averaged mora duration

The phrase-averaged mora duration in slowed phrases is 402.5[ms/mora] averagely. That in other phrases is 127.0[ms/mora]. In the figure 4, slowed phrases (shaded

<table>
<thead>
<tr>
<th>Table 1: Speech samples</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>the number of speech periods</td>
</tr>
<tr>
<td>speaker 1</td>
</tr>
<tr>
<td>148</td>
</tr>
</tbody>
</table>
bars) are obviously longer than normal rate phrases (white bars). It is known mora duration is varied by various factors. However, when a speaker slows the speech rate to draw a listener’s attention, mora durations are lengthened much larger than other factors. It suggests that variation of phrase-averaged mora duration can reflects variation of local speech rate.

4.2.2. Misdetections

Precision rates are 0.85 and 0.61 respectively. 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. If such phrases are perceptually slow, they are not misdetection. So we have to carry out an auditory test to check whether the phrases are slow.

Most misdetected phrases are slightly over the threshold. It seems that the difference of adjacent phrases should be more strongly taken into account for the threshold. Further investigation will be needed to design appropriate threshold.

4.2.3. Comparison

I compared proposed threshold with fixed threshold which is used in previous study[2]. The fixed threshold is averaged mora duration of speech samples used in preliminary experiments.

Table 3 shows recall rates, precision rates and F-measures of the detection results with fixed threshold. Recall rates are 0.94. It is same as the proposed threshold. However, the proposed threshold has slightly higher precision rate than the fixed threshold. Therefore, the threshold in consideration of the contrast of adjacent phrases has slight higher effectiveness.

5. Conclusions

In this study, I examined the parameters of the threshold for detecting local slower phrases. The threshold in consideration of the contrast between adjacent phrases has had slightly better result than simple straight threshold.

Further investigations will be needed to avoid misdetections. Pre-processes must be brushed up for applying this method to natural conversations.

6. References