An improved pair-wise variability index for comparing the timing characteristics of speech

Hua-Li Jian

Department of Foreign Languages and Literature
National Cheng Kung University, Tainan, Taiwan

hljian@mail.ncku.edu.tw

Abstract

The pair-wise variability index has become a useful and widely used tool for comparing syllable timing of speech. In this paper we present an improved pair-wise variability index based on median instead of means that can more strongly amplify and reveal the differences in the timing characteristics of two datasets. Further, it places less stringent requirements on the pre-processing of the measurements, and it is therefore more robust to outliers in the dataset. The effectiveness of the new measure is demonstrated through an example where the measure is applied to data based on American English speech and Taiwanese English speech. The results obtained with the improved pair-wise variability index are compared to those of the standard pair-wise variability index and the rhythm ratio.

1. The pair-wise variability index applied

The pair-wise variability index, originally proposed by Low and Grabe [1] and later refined in [2], is a simple measure that allows the rhythm characteristics of different speech samples to be easily compared. In their original study, the authors compared the timing of British English and Singaporean English and showed that there is more variability in consecutive syllable timings for reduced vowels than full vowels in British English than in Singaporean English, which is more staccato sounding. British English reduced vowels therefore have a higher pair-wise variability index than those for Singaporean English, which has a smaller pair-wise variability index.

Similar patterns occur for other varieties of English and also in different languages, and the pair-wise variability index has recently been applied by a large number of studies in different domains within the broad field of phonetics. These studies include Whitworth’s study of rhythm in German English [3], Jian’s study of rhythm in Taiwanese English [4] and Keane’s studies into rhythm in Tamil [5], [6]. Further, Sandgren [7] used the pair-wise variability index to assess the rhythmic deviations among Swedish learners of French.

Variations on the theme include Patel’s study [8] where the pair-wise variability index is used to draw the parallels between the rhythm in language and music. Patel [9] also argues that the term pair-wise variability index is an inaccurate term and proposes to use the term pair-wise contrast index. We choose to use the term pair-wise variability index as this is the most widely known term in the prosodic research community.


Peter and Stoel-Gammon [13] have applied the pair-wise variability index to speech therapy research, and Kent and Kim [14] have applied it into clinical phonetics research.

Other measures of rhythmic variability have also been suggested in the literature, such as the rhythm ratio [15] proposed by Gibbon and Gut. The rhythm ratio is computed as the mean of consecutive syllable duration ratios. Further, Gut [16], [17] applied both the pair-wise variability index and the rhythm ratio to compare a wide range of languages including British English, Nigerian English, Anyi, Ega, Ibibio and German learners of English. The rhythm ratio has currently not been widely embraced in the research community.

2. Improving the pair-wise variability index

In acoustical phonetics speech is usually segmented and labeled such that boundaries are marked at instances of time $b_1, b_2, ... , b_n$ where $n$ is the number of syllable boundaries. The duration $d_i$ is therefore computed as follows:

$$d_i = b_{i+1} - b_i$$

(1)

Based on a sequence of durations the mean pair-wise variability index $v_i$ is defined as:

$$v_i = \frac{1}{n-1} \sum_{i=1}^{n-1} |d_{i+1} - d_i|$$

(2)

Further, the normalized pair-wise variability index $pv_i$ for two consecutive durations $d_i$ and $d_{i+1}$ is defined as:

$$pv_i = \frac{|d_{i+1} - d_i|}{d_{i+1} + d_i}$$

(3)

However, since 2 is a constant factor we redefine the index as follows:

$$pv_i = \frac{|d_{i+1} - d_i|}{d_{i+1} + d_i}$$

(4)

And the mean normalised pair-wise variability index can be expressed as

$$\overline{pv}i = \frac{1}{n-2} \sum_{i=1}^{n-2} pv_i$$

(5)

Note that the original pair-wise variability index proposed by Low and Grabe was scaled into percentage by multiplying it by 100. In this paper the term pair-wise variability index is used to refer to the normalized pair-wise variability index. For comparison, the rhythm ratio $rr_i$ proposed by Gibbon and...
Table 1: Comparison of pair-wise variability index and rhythm ratio based on mean and medians for Taiwan English and American English speech.

<table>
<thead>
<tr>
<th>measure</th>
<th>category</th>
<th>Taiwan English</th>
<th></th>
<th>American English</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>full</td>
<td>reduced</td>
<td>Diff</td>
<td>full</td>
</tr>
<tr>
<td>duration</td>
<td>mean</td>
<td>0.15</td>
<td>0.14</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.14</td>
<td>0.13</td>
<td>0.01</td>
<td>0.13</td>
</tr>
<tr>
<td>Pair-wise variability index</td>
<td>mean</td>
<td>0.19</td>
<td>0.21</td>
<td>0.02</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.17</td>
<td>0.19</td>
<td>0.02</td>
<td>0.19</td>
</tr>
<tr>
<td>rhythm ratio</td>
<td>mean</td>
<td>0.30</td>
<td>0.32</td>
<td>0.02</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>median</td>
<td>0.29</td>
<td>0.32</td>
<td>0.03</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Gut, of two consecutive syllable durations $d_i$ and $d_{i+1}$ are defined as:

$$rr_i = 1 - \frac{\min(d_i, d_{i+1})}{\max(d_i, d_{i+1})}$$  \hspace{1cm} (6)

Note that the original rhythm ratio proposed by Gibbon and Gut was expressed in terms of percentages. Further, we have made it comparable to the pair-wise variability index such that a small rhythm ratio signals more equal durations than a larger ratio – the original formulation was interpreted the other way around.

Finally, expressing the rhythm ratio in terms of the maximum and minimum function is more concise than the original formulation. The mean rhythm ratio $rr$ is therefore expressed as

$$rr = \frac{1}{n-1} \sum_{i=1}^{n-1} rr_i$$  \hspace{1cm} (7)

The problem with the mean is that it is sensitive to outliers, and outliers also frequently occur in speech data. We therefore propose to base the pair-wise variability index and the rhythm ratio on the median of the sequences of pair-wise variability indices $pvi_1, pvi_2, \ldots, pvi_{n-2}$ and rhythm ratios $rr_1, rr_2, \ldots, rr_{n-1}$ respectively. The next section demonstrates the difference between using the median and using the mean on acoustic speech data taken from American English and Taiwan English speakers, as well as the difference between using the pair-wise variability index and the rhythm ratio.

3. Method

As an illustrative example we have taken data from a study by Jian [4], which investigated the timing of Taiwan English compared to that of American English. A vast body of literature exist addressing the rhythm and timing in British English (BE) and American English (AE), which are both known to be stressed-timed. However, very little has been written about the rhythm and timing of Taiwan variety of English (TE), although some studies have addressed other South Asian varieties of English – especially Singapore (SE) and Malaysian English (ME) since English is one of the official languages in Singapore. Most studies describe Singaporean English as being syllable-timed and also staccato sounding. In this study we investigate Taiwan English from the same viewpoint. English is not an official language in Taiwan though everyone is now taught English from an early age. Taiwan English is a distinct variety of English, although American pronunciation is generally taught throughout the island. Factors influencing the Taiwan variety of English include Japanese-style English brought over during the second world war occupation, mainland Chinese English mainly brought over just before 1945, British textbooks used before the introduction of the Kenyon and Knott general American pronunciation system in 1969, persistent defective pronunciation of English in classrooms, and effects from local languages – especially Mandarin and Southern Min (see Chung [18]). The assumption of Jian’s study and the current study are that the Taiwan English variety shares some characteristics with the Singaporean English due to similar influences such as the Chinese language. Jian’s original data based on the pair-wise variability index suggests that Taiwan English is syllable-timed. This study is based on the same data, but applied to both the pair-wise variability index and rhythm ratio based on both means and medians. The objective is to assess the respective suitability of the different measures for capturing key timing characteristics of speech rhythm.

3.1. Subjects

Five Taiwanese English (TE) and five American English (AE) speakers took part in the recording task. All subjects live in Taiwan. The native Taiwanese TE speakers were students at the National Cheng Kung University in Tainan. The native American AE speakers were mostly teachers in the Tainan and Kaohsiung areas in the south of Taiwan. The TE speakers were mostly in their early twenties, while the age of the AE ranged from mid twenties to early forties. None of the subjects reported having any articulation disability.

3.2. Materials

It was decided to adopt material with a similar structure to that described in Low, Grabe and Noland [1]. The TE and AE subjects were asked to read 10 sentences consisting of five full and potentially reduced vowels and five sentences comprising only full vowels. The sentences were shuffled into pseudorandom order and presented to the speakers without context and fillers.
3.3. Recording procedure

The subjects were recorded in an office in the National Cheng Kung University using a Shure KSM32 studio microphone and a portable mixer with a built-in microphone amplifier. The microphone was placed approximately 20 cm in front of the speakers. The subjects were given time to read the material prior to the recording and the subjects were manually informed to repeat utterances when mistakes were made. The material was digitally recorded onto a Minidisk using a stationary Sony deck. The digitally recorded material was subsequently transferred to audio files stored on a personal computer.

3.4. Post-processing and analysis

The speech material comprising 100 utterances of which 50 utterances were produced by the TE speakers, and the remaining 50 were produced by the AE speakers. The digital recording was partitioned so that each audio-file consisted of one sentence. The open source analysis package PRAAT by Boersma and Weenink [19], [20] was used to segment and label the utterances. The spectrogram feature and formant analysis tool in PRAAT was used to select the vowel boundary, which is not a trivial task. See Peterson and Lehiste [21] for a general discussion on segmentation of vowels by the means of spectrograms.

PRAAT stores segmentation information in separate text files. A small script was written to extract the relevant vowel information from the segmentation file so that it could be exported into a Microsoft Excel spreadsheet. Subsequent calculations were performed in Excel and statistical significance tests were conducted by the means of the data analysis plug-in for Excel.

4. Results and discussion

The results of the experiment are summarised in Table 1, Figures 1 and 2. Table 1 enumerates the pair-wise variability indices and rhythm ratio computed as means and medians in addition to mean and median syllable durations for full and reduced vowels from both Taiwan English and American English speech.

The table conforms to the expectations, i.e. that there is only a very small difference in pair-wise variability indices and rhythm ratios for full and reduced vowels in Taiwan English and that there is a larger difference between the indices for full and reduced vowels in American English, which is a stress-timed language. The table shows that the differences between full and reduced vowel duration for both the mean and median duration are larger in American English than in Taiwan English, while the differences between the mean and median values for pair-wise variability indices and for rhythm ratios are much larger comparatively.

These differences are more apparent in Figures 1 and 2. Figure 1 shows the differences (in percentages) between the median-based and mean-based pair-wise variability index for the full and reduced vowels. For Taiwan English the difference in percentage relative to the full vowel is 10.5 % and 11.8 % (see Figure 1 left) for the mean and the median respectively. Clearly, these differences are quite similar. However, for American English the differences using the mean and the median are 17.4 % and 31.6 % (see Figure 1 right), i.e. the difference for the median based pair-wise variability is nearly twice that of the mean-based index. This evidence indicates that the median pair-wise variability is better at amplifying and revealing the difference between the full and the reduced vowels in American English.

Figure 2 shows similar details for the rhythm ratio measure. For Taiwan English the difference between full and reduced vowels using mean- and median-based rhythm ratios are 6.6 % and 10.3 % (see Figure 2 left) respectively, and for American English the corresponding measures are 17.6 % and 25 % (see Figure 2 right). These differences are smaller than the differences for the pair-wise variability index. These observations suggest that the rhythm ratio is not as good as revealing the difference between full and reduced vowels as the pair-wise variability index. Further, the data suggests that the rhythm ratio is not as sensitive to the use of mean or median measures as the pair-wise variability index is. Overall, the results for this data set suggest that a median- based pair-wise variability index results in the most visible contrast in rhythm characteristics for the two different classes of speech samples.
5. Summary

In this study four variability measures were investigated, namely the mean pair-wise variability index, mean rhythm ratio and newly proposed median-based pair-wise variability index and median rhythm ratio. The results show that both the pair-wise variability index and rhythm ratios are able to express the rhythmic variations in the investigated data set. However, the data shows that median-based measures result in larger differences than the means for different rhythmic categories. Median-based variability measures are therefore probably a better means of amplifying and identifying rhythmic contrasts in acoustic measurement data than mean-based measures. Further, the median pair-wise variability index produced results that are less ambiguous and easier to interpret than the two rhythm ratio based quantities.

Acknowledgements

The author would like to thank the assistants Hoi Wee-Wee, Tan Te-Kong, and Chen Chia-Sian for assisting me with the overwhelming job of corpus labeling and segmentation. I also am grateful for the subjects that took the time to conduct the recording. This work is supported by the National Science Council of Taiwan grant NSC91-2411-H006-020.

6. References