Estimating Speaking Rate by Means of Rhythmicity Parameters

Christian Heinrich, Florian Schiel

Institute of Phonetics and Speech Processing, Bavarian Archive for Speech Signals, Ludwig-Maximilians-Universität, München, Germany

heinrich | schiel@bas.uni-muenchen.de

Abstract

In this paper we present a speech rate estimator based on so-called rhythmicity features derived from a modified version of the short-time energy envelope. To evaluate the new method, it is compared to a traditional speech rate estimator on the basis of semi-automatic segmentation. Speech material from the Alcohol Language Corpus (ALC) covering intoxicated and sober speech of different speech styles provides a statistically sound foundation to test upon. The proposed measure clearly correlates with the semi-automatically determined speech rate and seems to be robust across speech styles and speaker states.

Index Terms: speech rate, speech rhythm, Alcohol Language Corpus, BAS

1. Introduction

There exist several approaches to automatically estimate the rate of speech, mainly to enhance the performance of automatic speech recognition systems. Being able to classify speech material for example into slow, normal and fast speech allows for selecting adapted acoustic models for the recognition process which helps reducing the average word error rate by 32% as reported by Martinez et al. 1998 [4]. Speech rate also represents an important feature of the speech signal to be used for instance in empiric linguistics.

The straightforward way to obtain a speech rate measurement is to analyse a manually produced phonetic segmentation or the output of a speech recognition system. Both approaches have their disadvantages: manual segmentation is a costly and very time-consuming process, while automatic speech recognition still is too error-prone for many speech styles. Therefore it is desirable to assess the rate of speech automatically and directly from the speech signal. Another class of speech rate estimators hence builds on acoustic measurements only, even though calculating the speech rate with the help of a segmentation results in higher accuracy. However those estimators work independently of the recognition process and the obtained segmentation, they are no lexically-based measures, and the algorithms also work fast on large corpora.

Most of these rate of speech estimators offer a unit per time approximation, for example phones per second or syllables per second, where the number of phones or syllables is determined by evaluating acoustic properties. In the majority of cases prominent events such as peaks occurring in modified versions of the energy envelope are therefore counted by involving several peak counting algorithms, e.g. [3], [5] and [6]. Pfau and Ruske [10] presented a method where the smoothed modified loudness was used to detect vowels and therefore vowel clusters and syllable nuclei, which correspond to the number of syllables within an utterance. Verhasselt and Martens [17] proposed a rate of speech detector (phones per second), that was based on phone boundary probabilities provided by a Multi-Layer Perceptron. Narayanan and Wang ([17],[18]) introduced a method using temporal correlation and selected sub-band correlation (tcssbc), which performs a spectral and a temporal correlation and also involves a smoothing and a thresholding mechanism to improve the peak counting. A comparative study of eight different methods for speech rate estimation has been provided by Dekens et al. [1], who found the cssbc method to be the most reliable one. Two Years later Zhang and Glass [20] showed, that an envelope analysis on the input speech signal combined with an estimation of the global speech rhythm on the signal envelope brings further improvement compared to the tcssbc method. Other approaches work independently of any linguistic unit by exploiting the short-time stationarity of speech features (e.g. [14]).

The approach described here is based on the phonological concept of the syllable where a maximum in sonority is considered as the syllable nucleus and minima of sonority mark the syllable boundaries. We developed a fast and simple algorithm processing the short-time energy envelope. It provides so called rhythmicity features that are clearly correlated to segment-based speech rate. Furthermore it can be shown that it behaves consistently for different speech styles and both alcoholized and non-alcoholized speech.

The remaining paper is structured as follows: the next section describes the proposed rhythmicity method for speech rate estimation and the method to determine a reference for the evaluation. Section 3 gives details about the speech corpus which the presented approach was tested upon. The statistical framework to evaluate the new method and the obtained results are illustrated in Section 4.

2. Speech rate estimation

Below the method to obtain a reference speech rate for the Alcohol Language Corpus and the suggested speech rate estimator based on rhythmicity features are described.

2.1. Syllable rate as reference

For the evaluation we need a reliable reference the proposed method can be compared to. We therefore estimate the syllable rate (SR) based on a semi-automatic phonetic segmentation provided by the Munich AUtomatic Segmentations MAUS [14]. This step requires considerable manual effort because MAUS needs an orthographic transcript as input. On the other hand this approach results in very precise estimates for the speech rate. In contrast to other automatic segmentation systems MAUS is able to detect deleted or inserted phones by comparing the speech signal to a potential selection of pronunciation variants pre-
dicted by a statistical language model. For this reason the segmentation does not necessarily contain the same number of syllables as the orthographic form of the utterance (in fact it rarely does).

From the MAUS segmentation we calculate the reference SR as the ratio of the number of vowel cluster nuclei (which corresponds to the number of syllables) to the total duration of the utterance in seconds. We did not calculate a local speech rate as Pfitzinger [11], since in this study we are only interested in the average speech rate of a complete recording. Considering that we also did not exclude silence intervals from the analysis as has been done in other studies.

2.2. Rhythmicity parameters

The short-time RMS of a speech signal shows the dynamic of the sound pressure energy which represents a sequence of alternating relatively loud and quiet parts. It can be used to describe rhythmicity features within the speech signal [15]. The basic idea for speech rate estimation is to automatically derive a sequence of alternating RMS maxima and minima which preferably resemble the syllable nuclei and syllable boundaries and then measure the time distances between peaks or valleys. The RMS analysis algorithm we use is the built-in kassp RMS algorithm of the Emu database system [2]. We use a Blackman window of 100 ms length and a window shift of 20 ms. These settings result in a moderate smoothing of the energy contour but still preserve short nuclei as are typical for unaccented or reduced syllables. Subsequently the obtained RMS contours are normalized for each utterance separately to ensure a comparable database. From this normalized RMS contour a sequence of local minima and maxima is determined. A simple thresholding mechanism filters all local maxima that are below mean of RMS within the utterance and all local minima that are above it. The resulting sequence of minima and maxima contains single minima and single maxima as well as clusters of minima and maxima. Since a cluster of maxima most likely represents a single syllable nucleus, we only keep the maximum with the highest RMS value and do likewise with minima clusters, where the minimum with the lowest RMS value is selected. This finally results in a consecutive sequence of alternating minima and maxima throughout the recording. Based on this min-max sequence we measure the time distances between successive minima and maxima. The described method does not require any spectral transformations and the computational effort is therefore negligible.

Figure 1 shows a part of the normalized RMS contour, the derived min-max sequence and the time distances \(d_1, \ldots, d_4\) between successive maxima for a single speech recording. The mean over all \(N\) time distances \(d_n\) between maxima is our proposed measure for speech rate (speech rate rhythmicity parameter, SRRP):

\[
SRRP = \frac{1}{N} \sum_{n=1}^{N} d_n
\]

Since the measured time distances are supposed to represent either the distance between syllable nuclei (distance between maxima in sonority) or the distance between syllable boundaries (distance between minima in sonority), we expect their average to be correlated to the inverse average syllable rate of the respective recording. For this study we only considered an SRRP concerning maxima, but it is also likely that working with the minima generates comparable results.

\[\text{Figure 1: Min-max contour (black) over normalized RMS contour (gray). Time distances } d_1, \text{ to } d_4 \text{ between successive maxima.}\]

3. Speech data

The speech data for this study is derived from the Alcohol Language Corpus (ALC) which comprises 37 hours of sober and intoxicated speech of 162 German speakers of both genders. ALC includes read, spontaneous and command & control speech. Read speech covers numbers, addresses and tongue twisters. Spontaneous speech comprises monologues and dialogues with the recording supervisor and the command & control speech consists of speech commands typically used for in-car communication with a vehicle computer. For a more detailed description of ALC see Schiel et al. 2011 [16].

4. Results and discussion

We calculated both SR and SRRP for each of the 162 speakers, every speech style and alcoholized versus non-alcoholized. Therefore our data matrix consists of 162 \(\times 3 \times 2 = 972\) values for SR and SRRP. We also calculated SR and SRRP without differentiating between the speech styles (162 \(\times 2 = 324\) data values) to allow for establishing correlations for the complete speech material of the speakers.

4.1. Global means

Table 1 shows the means of SR and SRRP for all speakers, separately for the three speech styles read, spontaneous and command speech and both intoxicated and sober speech. SR is specified in syllables per second and SRRP in milliseconds.
Table 1: Means of SR and SRRP for all speakers, intoxicated and sober speech (a and na) as well as the three speech styles read (read), spontaneous (spont) and command speech (comm).

<table>
<thead>
<tr>
<th>style</th>
<th>intox.</th>
<th>read a</th>
<th>read na</th>
<th>spont a</th>
<th>spont na</th>
<th>comm a</th>
<th>comm na</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>2.94</td>
<td>3.17</td>
<td>3.01</td>
<td>3.2</td>
<td>3.66</td>
<td>3.72</td>
<td></td>
</tr>
<tr>
<td>SRRP</td>
<td>349</td>
<td>317</td>
<td>494</td>
<td>468</td>
<td>318</td>
<td>323</td>
<td></td>
</tr>
</tbody>
</table>

As you can see both measures indicate slower rates for intoxicated speech than for sober speech, except for command speech, where the SRRP for non-alcoholized speech is slightly higher than for alcoholized speech. Considering that silence intervals are included in both SR and SRRP, the relatively high mean values for SRRP in spontaneous speech suggest, that there are quite a number of rather large max to max time distances comprising these silence intervals. This can also be seen in Figures 4 and 5 where the SRRP values of spontaneous speech almost reach 900 ms for intoxicated and nearly 750 ms for sober speech. In contrast to spontaneous speech, read speech and command speech do not contain as many and mainly large silence intervals because speakers can prepare themselves before speaking. Furthermore the utterances that are part of the read material are relatively short compared to the spontaneous utterances. Therefore it is to be expected that the correlation coefficient for SR and SRRP regarding spontaneous speech is the poorest.

4.2. Correlation between SRRP and 1/SR

The scatter plots of 1/SR and SRRP can be seen in Figure 2 for intoxicated speech and Figure 3 for sober speech. 1/SR is given in milliseconds and corresponds to the average syllable length. As expected SRRP clearly correlates with 1/SR. The correlation coefficient for the intoxicated material of all speakers is r=0.74 (Fig. 2) and for the sober material r=0.72 (Fig. 3).

A Repeated Measures ANOVA exposes significant differences (p < 0.001) between the three speech styles and also between intoxication levels for SRRP.

Figure 2: Correlation between SRRP and 1/SR (both in ms) of alcoholized speech (a).

Figure 3: Correlation between SRRP and 1/SR (both in ms) of non-alcoholized speech (na).

Figure 4: Correlation between SRRP and 1/SR (both in ms) of alcoholized speech shown separately for the three speech styles read (read), spontaneous (spont) and command speech (comm).

Figure 4 presents the correlations between SRRP and 1/SR for the three speech styles and alcoholized speech. Figure 5 for non-alcoholized speech. In both cases read speech as well as command speech feature relatively high correlation coefficients whereas for spontaneous speech they are marginally lower. The separated scatter plots show that the proposed measure based on the short-time energy envelope behaves consistently and thus can be applied to different speech styles.
We would like to thank the ALC team for providing the speech data and the orthographic transcription.

5. Conclusion

In this study we presented a new method to estimate the rate of speech by finding peaks in the short-time energy envelope of an acoustic speech signal. The method does not require a high computational load and seems to be quite robust across different speech styles and different speaker states such as intoxicated and sober speech. Clearly correlating with the segment-based SR the measure based on the rhythmicity parameters respectively the energy envelope (SRRP) reliably reflects the rate of speech and for this purpose can be used to classify speech material adequately.

6. Acknowledgements

This work was partly supported by the DFG, contract number SCHI1117/1-1. We would like to thank the ALC team for providing the speech data and the orthographic transcription.

7. References