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. Using speech rate dependent models together with a speech rate classifier reduces the average word error rate by 32% as reported by Martinez et al. [4]. Speech rate also represents a more general feature of the speech signal to be used for instance in empiric linguistics.

The easiest 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 this is not as reliable as calculating the speech rate with the help of a segmentation. However those estimators work independently of the recognition process and the obtained segmentation, 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 and not by counting the number of units per time directly, i.e. by accessing a segmentation. 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 ([7],[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 *tcssbc* 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 does not presume to compete against the already established rate of speech estimators. But it shows that a fast and simple algorithm processing the short-time energy envelope provides so called rhythmicity features which are clearly correlated to segment-based speech rate. Furthermore we can show 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

Following 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

To evaluate it, 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 values for the speech rate. In contrast to other automatic segmentations MAUS is able to detect deleted or inserted phones by comparing the speech signal to a potential selection of pronunciation variants predicted by a sta-

tistical 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. Following this approach 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 *tkassp* RMS algorithm of the *Emu* database system ([2])[1]. 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.[2] 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 maxima and minima. 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 curve, the derived *min-max* sequence and the time distances $d_1 ... d_4$ 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) or the distance between syllable boundaries (distance between minima), we expect their average to be inversely correlated to the true average syllable rate of the respective recording.

[1]For further information visit *http://emu.sourceforge.net/*

[2]We also considered using an adaptive mean calculated successively from a larger time span (2 sec), since in rare cases there might be a slow change of loudness within a recording. But this approach leads to meaningless *min-max* sequences for longer silence intervals and was therefore not pursued.

For this study we only considered an SRRP concerning maxima, but it is also likely that working with the minima generates comparable results. [3]

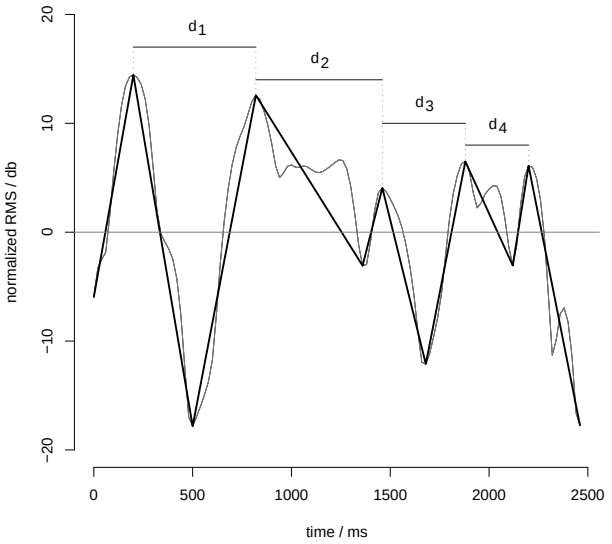

Figure 1: *Min-max curve (black) over normalized RMS curve (gray). Time distances d_1 to d_4 between successive maxima.*

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 "Alcohol Language Corpus" [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 $162 * 3 * 2 = 972$ values for SR and SRRP as well. We also calculated SR and SRRP without differentiating between the speech styles ($162 * 2 = 324$ 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 the segment-based syllable rate (SR) and the speech rate based on rhythmicity parameters (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. As you can see both speech rates indicate slower rates for intoxicated speech than for sober

[3]Other rhythmicity features which represent the intrinsic rhythmical syllable structure can be derived from the *min-max* sequence [15].

Table 1: *Means of the syllable rates (SR) and speech rates derived from the rhythmicity parameters (SRRP) for all speakers, intoxicated and sober speech (a and na) as well as the three speech styles read, spontaneous and command speech.*

style intox.	read a	read na	spont a	spont na	comm a	comm na
SR	2.94	3.17	3.01	3.2	3.66	3.72
SRRP	349	317	494	468	318	323

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 obvious that the correlation coefficient for SR and SRRP regarding spontaneous speech is the poorest.

4.2. Correlation between SRRP and SR

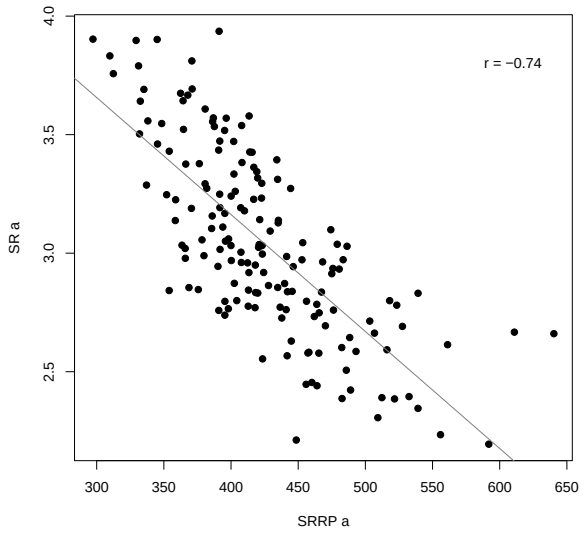

Figure 2: *Correlation between the mean of the time distances between maxima (SRRP) in ms and the syllable rate (SR) of alcoholized speech (a) in syllables per sec.*

The scatter plots of SR and SRRP can be seen in Figure 2 for intoxicated speech and Figure 3 for sober speech. As expected the speech rate provided by the mean of the time distances between maxima (SRRP) inversely correlates with the average syllable rate (SR).
A Repeated Measures ANOVA shows that SRRP exhibits significant differences ($p < 0.0001$) between the three speech styles and also between intoxication levels.
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).

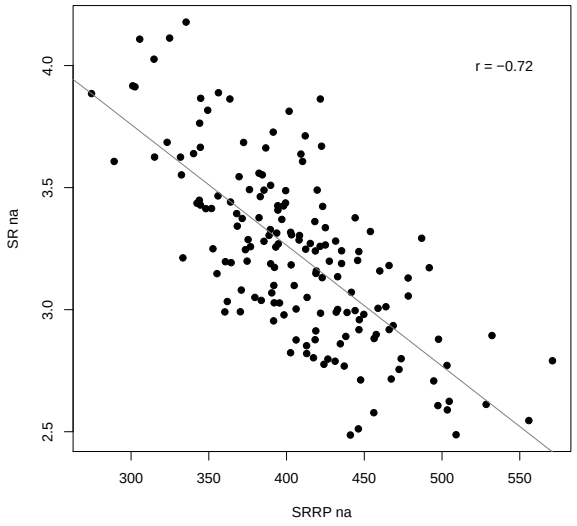

Figure 3: *Correlation between the mean of the time distances between maxima (SRRP) in ms and the syllable rate (SR) of non-alcoholized speech (na) in syllables per sec.*

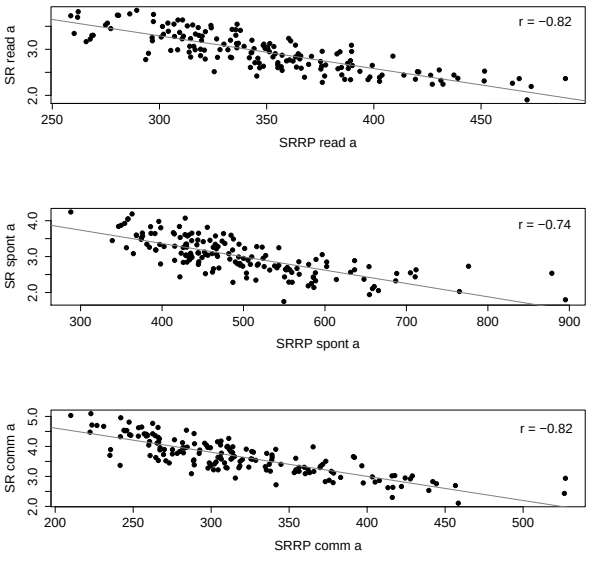

Figure 4: *Correlation between the mean of the time distances between maxima (SRRP) in ms and the syllable rate (SR) of alcoholized speech (a) in syllables per sec. shown separately for the three speech styles read, spontaneous and command speech.*

Figure 4 presents the correlations between SRRP and SR for the three speech styles and alcoholized speech, Figure 5 for non-alcoholized speech. For read speech as well as command speech the correlation coefficients are relatively high whereas for spontaneous speech they are marginally lower.

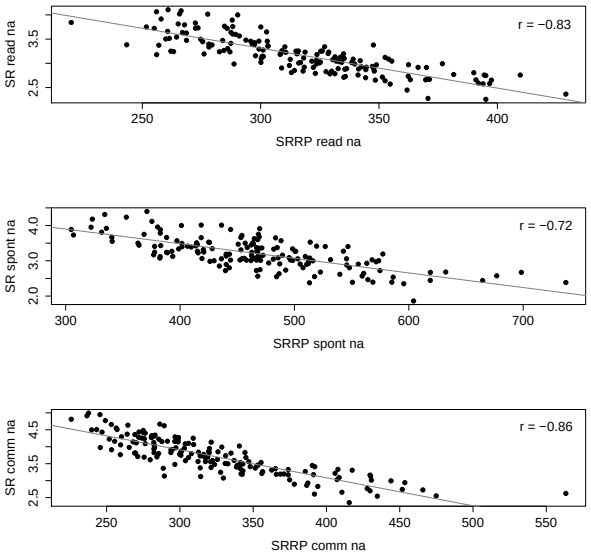

Figure 5: *Correlation between the mean of the time distances between maxima (SRRP) in ms and the syllable rate (SR) of non-alcoholized speech (na) in syllables per sec. shown separately for the three speech styles read, spontaneous and command speech.*

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. It can be seen that alcoholic intoxication strongly affects speaking rate: people tend to speak slower under the influence of alcohol.
In addition the standard deviation of the mean SRRP exhibits higher values for alcoholized speech. Referring to Schiel et al. [15] where all the presented rhythmicity features (including SRRP which is there introduced as rhythm feature B) were reported to rise with alcoholization, as a general result it thus can be said that speech under alcoholic intoxication is more irregular than sober speech.

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 very robust across different speech styles and different speaker states such as intoxicated and sober speech. Clearly correlating with the segment-based syllable rate (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

[1] Dekens, T., Demol, M., Verhelst, W., Verhoeve, P. (2007), "A Comparative Study of Speech Rate Estimation Techniques". In: Proc. of the INTERSPEECH 2007, Antwerp, Belgium, pp. 510-513.

[2] Cassidy, S., Harrington, J. (2001), "Multi-level annotation in the EMU speech database management system". Speech Communication 33(1-2), pp. 61-77.

[3] Kitaazawa, S., Ichikawa, H., Kobayashi, S., Nishinuma, Y. (1997), "Extraction and Representation Rhythmic Components of Spontaneous Speech". In: Proc. of the EUROSPEECH 1997, Rhodes, Greece, pp. 641-644.

[4] Martinez, F., Tapias, D., Alvarez, J. (1998), "Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition", In: Proc. of the ICASSP 1998, New York, USA, vol. 2, pp. 725-728.

[5] Morgan, N., Fosler, E., Mirghafori, N. (1997), "Speech Recognition Using On-Line Estimation of Speaking Rate". In: Proc. of EUROSPEECH 1997, Rhodes, Greece, vol.4, pp. 2079-2082.

[6] Morgan, N., Fosler-Lussier, E. (1998), "Combining Multiple Estimations of Speaking Rate". In: Proc. of the ICASSP 1998, Seattle, Washington, pp. 729-732.

[7] Narayanan, S., Wang, D. (2005), "Speech Rate Estimation via Temporal Correlation and Selected Sub-Band Correlation". In: Proc. of the ICASSP 2005, Philadelphia, PA, pp. 413-416.

[8] Ohno, S., Fujisaki, H. (1995), "A Method for Quantitative Analysis of the Local Speech Rate". In: Proc. of EUROSPEECH 1995, Madrid, Spain, vol. 1, pp. 421-424.

[9] Pellegrino, F., Farinas, J., Rouas, J.-L. (2004), "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech". Speech Prosody 2004, Nara, Japan, pp. 517-520.

[10] Pfau, T., Ruske, G. (1998), "Estimating the Speaking Rate by Vowel Detection". In: Proc. of the ICASSP 1998, Seattle, Washington, p. 945-948.

[11] Pfitzinger, H. R. (1996), "Two Approaches to Speech Rate Estimation". In: Proc. of SST 1996, Adelaide, Australia, pp. 421-426.

[12] Pfitzinger, H. R. (1998), "Local Speech Rate as a Combination of Syllable and Phone Rate". In: Proceedings of ICSLP 1998, Sydney, Australia, vol. 3, pp. 1087-1090.

[13] Pfitzinger, H. R. (1999), "Local Speech Rate Perception in German Speech". In: Proc. of the ICPhS 1999, San Francisco, California, pp. 893-896.

[14] Schiel, F. (1992), "Phonetically Seeded SCHMM of Variable Length for Speaker independent Recognition of Isolated Words". In: Proc. of the Forth Australian International Conference on Speech Science and Technology, Brisbane, Australia, pp. 92-97.

[14] Schiel, F. (1999), "Automatic Phonetic Transcription of Non-Prompted Speech". In: Proc. of the ICPhS 1999. San Francisco, California, pp. 607-610.

[15] Schiel, F., Heinrich, C. (2010), "Rhythm and Formant Features for Automatic Alcohol Detection". In: Proceedings of the Interspeech 2010, Tokio, Japan, pp. 458-461.

[16] Schiel, F., Heinrich, C., Barfüßer, S. (2011), "Alcohol Language Corpus". In: Language Resources and Evaluation, Springer, Berlin, New York, in print.

[17] Verhasselt, J. P., Martens, J.-P. (1996), "A Fast and Reliable Rate of Speech Detector". In: Proc. of ICSLP 1996, Philadelphia, Pennsylvania, vol.4, pp. 2258-2261.

[18] Wang, D., Narayanan, S. (2007), "Robust Speech Rate Estimation for Spontaneous Speech". In: IEEE Transactions on Audio, Speech, and Language Processing, vol. 15, 2007, pp. 2190-2201.

[19] Yuan, J., Liberman, M. (2010), "Robust Speaking Rate Estimation Using Broad Phonetic Class Recognition". In: Proc. of the ICASSP 2010, Dallas, Texas, pp. 4222-4225.

[20] Zhang, Y., Glass, J. R. (2009), "Speech Rhythm Guided Syllable Nuclei Detection". In: Proc. of the ICASSP 2009, Taipei, Taiwan, pp. 3797-3800.