TWO APPROACHES TO SPEECH RATE ESTIMATION

Hartmut R. Pfitzinger

Institut für Phonetik und Sprachliche Kommunikation
University of Munich, Germany

ABSTRACT – This paper introduces two approaches to speech rate estimation: one is based on automatic syllable detection and the other on automatic phone segmentation. For evaluation of both approaches we used manually segmented syllables and phones as a reference. Although the used segment detectors are not perfect it is possible to automatically estimate the local rate of phones and the local rate of syllables reliably. We argue that neither the rate of phones nor the rate of syllables suffices for estimating actual speech rate.

INTRODUCTION

Over recent years, efforts have increased to improve automatic speech recognition by making use of prosodic cues. Undoubtedly, speech rate is one of the most important prosodic cues because it modifies acoustic cues (e.g. transitions), phones, and even words (Crystal & House, 1990), and therefore plays also an important role in speech synthesis (Campbell, 1991). Speech rate detectors reported in the literature are based either on information provided by the recognition process, or on phone rate (Siegler & Stern, 1995; Verhasselt & Martens, 1996). Both are strongly influenced by speech rate resulting in a lack of validity of the estimated speech rates, especially for very fast or slow speech.

The approach proposed here also uses the number of speech units like words, syllables or phones per second to calculate speech rate. Since word boundaries cannot be detected without a recognition system they will not be taken into account (Jones & Woodland, 1993). Contrary to words, syllables and phones can be detected automatically by efficient algorithms with reasonable error rates (Hunt, 1992; Kipp et al., 1996; Pfitzinger et al., 1996). We will therefore investigate how speech rate can be estimated either by automatically detected syllables or by automatically segmented phones. Manually labelled phones and syllables served as a reference.

GLOBAL, LOCAL, AND RELATIVE SPEECH RATE

It is of great importance to differentiate between global, local, and relative speech rate. The *global speech rate* yields from dividing the number of segments by the sum of their durations for a complete utterance.

The *local speech rate* is of central interest in this study being a prosodic feature and allowing for drawing conclusions, which constituents in a given sentence are stressed and therefore being articulated with a slower rate, or which constituents have a higher speech rate and as a consequence show more assimilation. The local speech rate is estimated by a moving average.

The distinction between *gross and net* measures of rate (Wood, 1973) has to be taken into account: counting segments actually represented in the speech wave leads to the net speech rate that is focused here. The gross speech rate results from counting segments deemed to have been present in some phonological or linguistic abstraction of the utterance.

One possibility to calculate the *relative speech rate* was proposed in (Ohno & Fujisaki, 1995) and consists of DTW-based time alignment of a test and a reference utterance, both spoken by the same speaker at different speech rates.

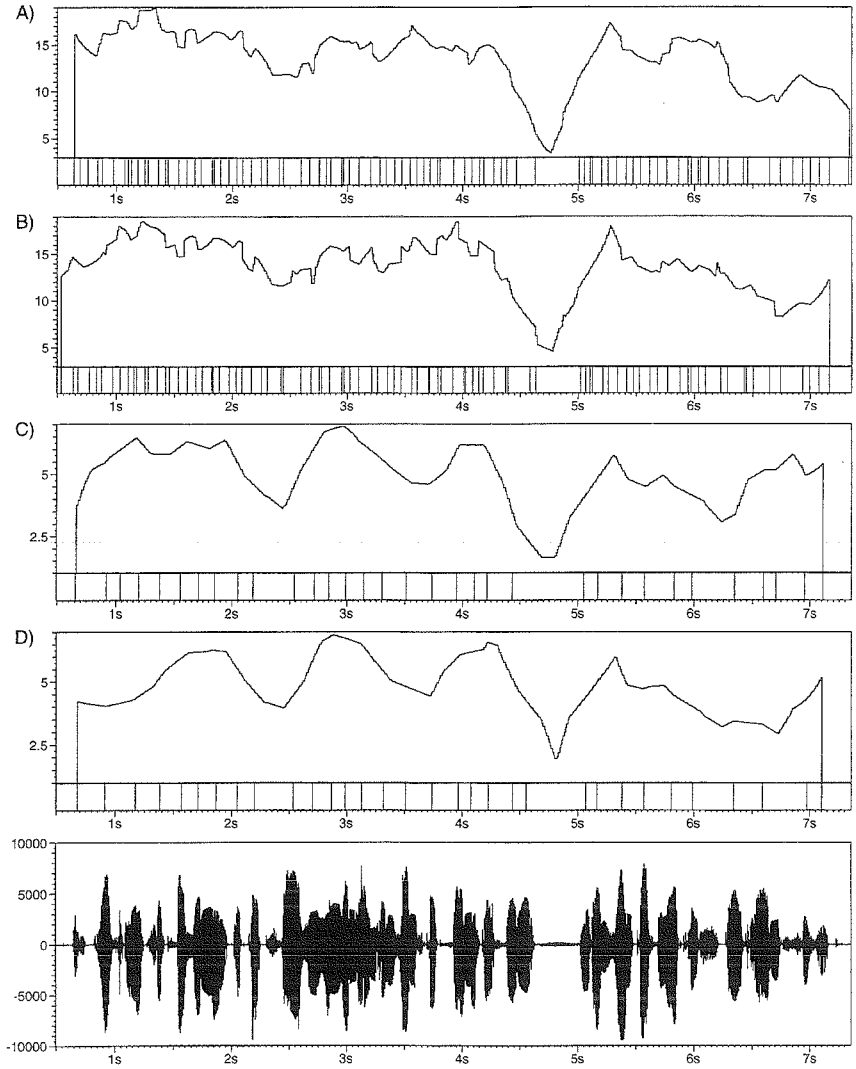

Figure 1: Speech signal of the PhonDat sentence no. *tpon7050* and four different local speech rate functions. Function (A) is based on manually labelled phones, (B) on automatically segmented phones, (C) on manually labelled syllables, and (D) on automatically detected syllables. The ordinates of each of the four panels represent segments/second, and the lower part illustrates the segment boundaries.

This investigation is focused on local speech rate. We used 30 sentences spoken by eight speakers (5 male, 3 female) giving a total of 240 sentences. Each 100 ms the distances between subsequent segmentation marks falling in a window of the length of 500 ms were accumulated and divided by their number. The reciprocal of the quotient is a measure for the speech rate (ROS):

$$ROS_{LR} = \frac{\frac{S_{l+1}-w_L}{S_{l+1}-S_l} + \frac{w_R-S_r}{S_{r+1}-S_r} + r - l - 1}{S_{l+1} - w_L + w_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i},$$

were w_L is the left and w_R is the right window boundary. Since the left (S_l) and the right (S_r) segment most frequently are covered only partially by the accumulation window, they have to be accumulated proportionally to guarantee a constant window length. This procedure is mathematically sound and leads to less outliers. Fig. 1 illustrates some examples calculated with a reduced step size of 10 ms increasing graphical resolution.

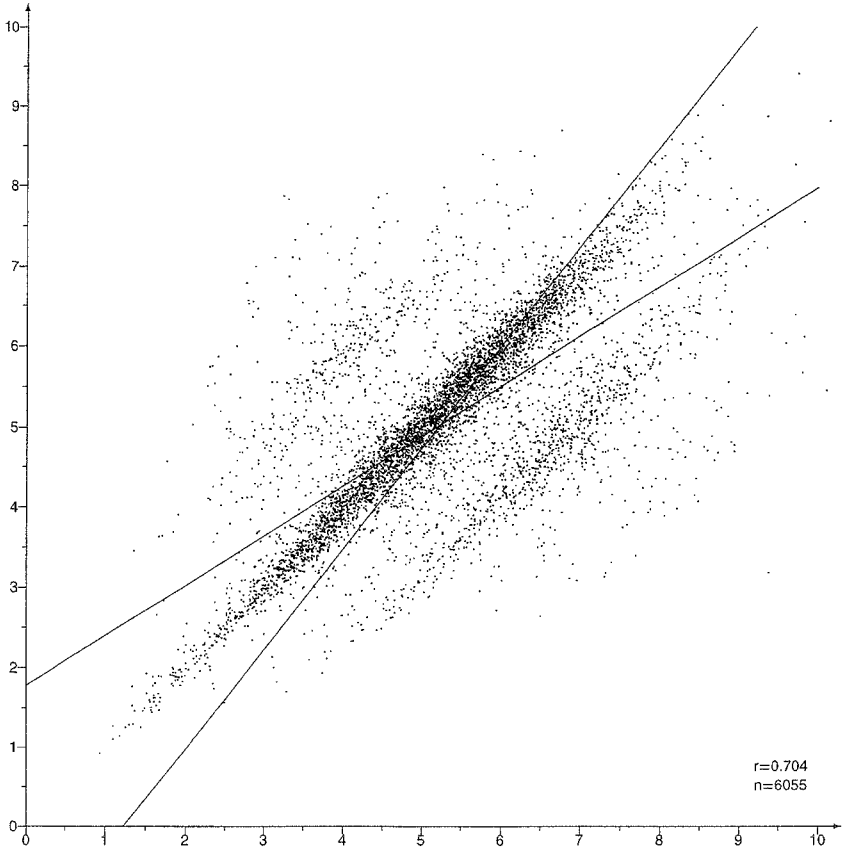

Figure 2: Scatter plot of the speech rate based on automatically detected syllables versus the speech rate based on manually segmented syllables. The speech rate is measured in syllables/second.

The automatic syllable detection method used in the first approach is a time domain algorithm: the logarithmic short-term amplitude of the bandpass filtered speech signal (250Hz–1kHz) is low pass filtered (8Hz) to suppress ripples caused by F0 or transient phonemes and to force the system to oscillate at the syllable frequencies. The peaks of the resulting energy contour represent the syllable nuclei. The current error rate being composed of false rejections and misses of reference syllables is 94% and the percentage of false insertions is 4%. For details refer to (Pfitzinger et al., 1996). Finally, the speech rate is estimated by a moving average over the syllable nuclei distances.

Fig. 2 illustrates a scatter plot of the automatically estimated rate of syllables versus the rate of syllables estimated from the manually segmented syllables. The number of frames is 6055. Two parallels to the central diagonal can be observed at which the density is significantly higher. The upper one results from false insertions of syllable marks and the lower one from false rejections.

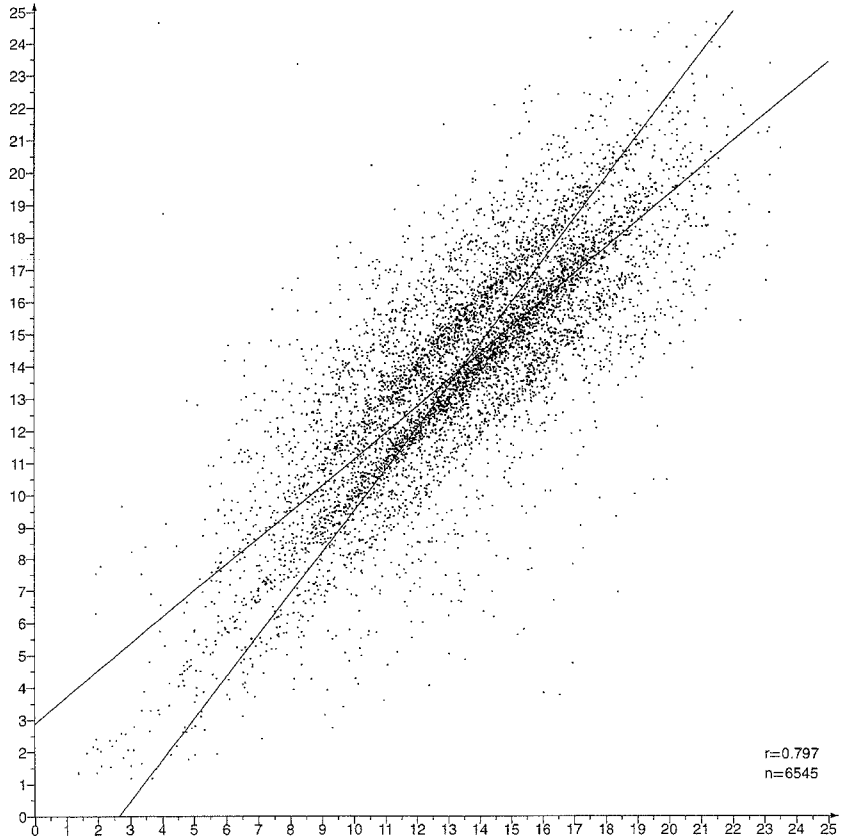

Figure 3: Scatter plot of the speech rate based on automatically detected phone boundaries versus the speech rate based on manually segmented phones. The speech rate is measured in phones/second.

PHONE-BASED APPROACH

This approach uses the automatic segmentation results produced by a hybrid statistical and rule-based segmentation system which takes into account phonetic variation of German (Kipp et al., 1996). At first the orthographic representation of an utterance is converted into phonetic transcription by lexicon lookup. Reduction rules comparable to those of generative phonology lead to all possible variations of the transcription. These are represented in a search graph structure. A Viterbi search time-aligns the speech signal of the utterance to the search graph and finds the best transcription and the segment boundaries.

Fig. 3 shows a scatter plot of the automatically estimated rate of phones versus the rate of phones estimated from the manually segmented phones. It is worth noting that as in fig. 2 the diagonals are obvious. The correlation coefficient ($r = 0.797$) is slightly higher than the correlation coefficient resulting from syllables.

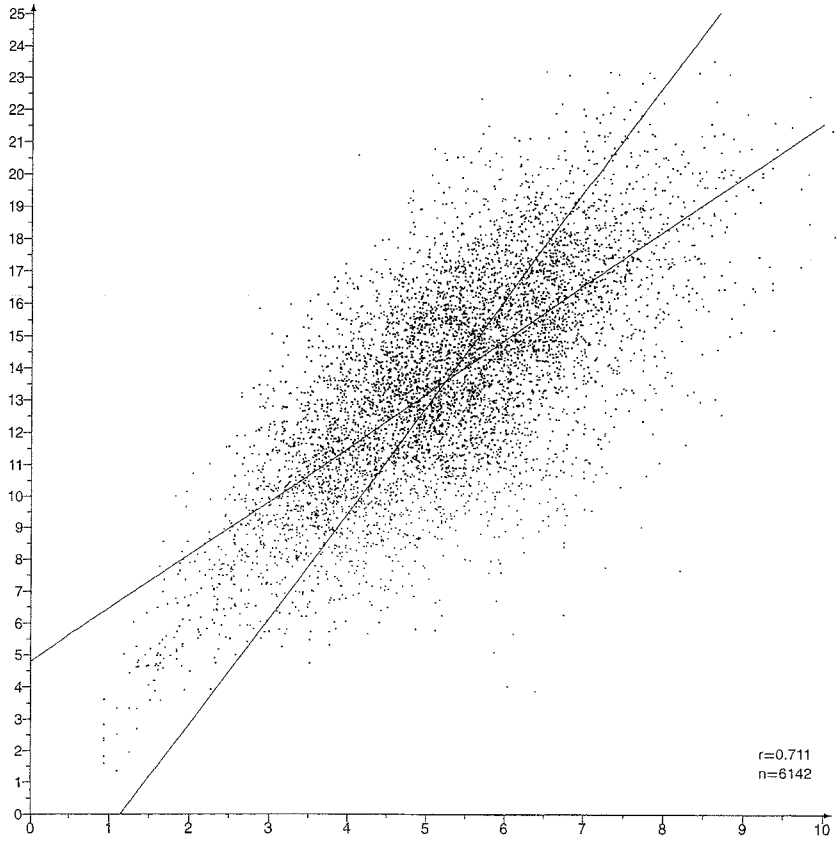

Figure 4: Scatter plot of the speech rate based on manually segmented phones versus the speech rate based on manually segmented syllable boundaries. The ordinate illustrates the speech rate measured in phones/second and the abscissa illustrates the speech rate measured in syllables/second.

This can be explained by the fact that one single error occurring in the syllable marks causes a higher deviation from the diagonal than one error in the phone marks since the mean number of phone marks per second is roughly 2.5 times higher.

RESULTS

This investigation has shown that it is possible to automatically estimate a good approximation to the local rate of phones and the local rate of syllables even though the underlying segment detectors are not perfect.

Fig. 4 illustrates a scatter plot of the rate of phones versus the rate of syllables both estimated from manually labelled speech. The correlation coefficient $r = 0.711$ is quite high, but it shows clearly, that the information content of both methods is not identical. It is evident that constituents containing a small number of phones and a mean rate of syllables, or words with many syllables and a mean rate of phones do not allow for deciding whether the rate of phones or the rate of syllables correspond to the speech rate. It is rather likely that speech rate is a combination of both. Consequently, it is not allowed to describe neither the rate of phones nor the rate of syllables as the speech rate.

Part of our future work is to perceptually evaluate speech rate and to check whether there is a combination of syllable and phone rate that correlates well with perceptual speech rate. Also the gross speech rate should play an important role in future research.

REFERENCES

[1] Campbell, W. N.; Isard, S. D. (1991) *Segment durations in a syllable frame*, J. Phon. 19, pp. 37–47.

[2] Cedergren, H. J.; Perreault, H. (1994) *Speech rate and syllable timing in spontaneous speech*, Proc. ICSLP 94, Yokohama, vol. 3, pp. 1087–1090.

[3] Crystal, T. H.; House, A. S. (1990) *Articulation rate and the duration of syllables and stress groups in connected speech*, JASA, vol. 88, pp. 101–112.

[4] Hunt, A. J. (1992) *Recurrent neural networks for syllabification*, Proc. SST, Brisbane, pp. 220–225.

[5] Jones, M.; Woodland, P. C. (1993) *Using relative duration in large vocabulary speech recognition*, Proc. EUROSPEECH'93, Berlin, vol. 1, pp. 311–314.

[6] Kipp, A.; Wesenick, M.-B.; Schiel, F. (1996) *Automatic detection and segmentation of pronunciation variants in German speech corpora*, Proc. ICSLP 96, Philadelphia.

[7] Mirghafori, N.; Fosler, E.; Morgan, N. (1995) *Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes*, Proc. EUROSPEECH'95, Madrid, vol. 1, pp. 491–494.

[8] Ohno, S.; Fujisaki, H. (1995) *A method for quantitative analysis of the local speech rate*, Proc. EUROSPEECH'95, Madrid, vol. 1, pp. 421–424.

[9] Pfitzinger, H. R.; Burger, S.; Heid, S. (1996) *Syllable detection in read and spontaneous speech*, Proc. ICSLP 96, Philadelphia.

[10] Siegler, M. A.; Stern, R. M. (1995) *On the effects of speech rate in large vocabulary speech recognition systems*, Proc. ICASSP-95, vol. 1, pp. 612–615.

[11] Verhasselt, J. P.; Martens, J.-P. (1996) *A fast and reliable rate of speech detector*, Proc. ICSLP 96, Philadelphia.

[12] Wood, S. (1973) *What happens to vowels and consonants when we speak faster?*, Working Papers Lund, Phonetics Laboratory Lund University, pp. 8–39.