LOCAL SPEECH RATE AS A COMBINATION OF SYLLABLE AND PHONE RATE

Hartmut R. Pfitzinger

Department of Phonetics, University of Munich
Schellingstr. 3, 80799 München, Germany
[hpt@phonetik.uni-muenchen.de]

ABSTRACT

This investigation focuses on deriving local speech rate directly out of the speech signal, which differs from syllable rate and from phone rate. Since local speech rate modifies acoustic cues (e.g. transitions), phones, and even words, it is one of the most important prosodic cues.

Our local speech rate estimation method is based on a linear combination of the syllable rate and of the phone rate, since this investigation strongly suggests that neither the syllable rate nor the phone rate on its own represent the speech rate sufficiently.

Our results show (a) that perceptual local speech rate correlates better with local syllable rate than with local phone rate ($r = 0.81 > r = 0.73$), (b) that the linear combination of both is well-correlated with perceptual local speech rate ($r = 0.88$), and (c) that it is now possible to calculate the perceptual local speech rate with the aid of automatic phone boundary detectors and syllable nuclei detectors directly from the speech signal.

1 INTRODUCTION

Speech rate is the subject of many investigations, however no homogeneous opinion exists of what actual speech rate is. In Pfitzinger [4, 1996] we attempted to clarify the nomenclature by distinguishing between local, global, and relative speech rate, and between gross and net measures of speech rate. The common practice in recent literature of using the term *speech rate* when actually phone rate is meant, follows from the fact that errors in speech recognition are more dependent upon phone rate than upon word or syllable rate (Siegler&Stern [7, 1995]).

Syllable rate and phone rate show a quite high linear correlation coefficient ($r = 0.6$), but obviously the information content of both differ (see fig. 1). Nevertheless, each of them is understood as an equivalent to the speech rate. Undoubtedly, a high speech rate is characterized by above-average syllable rates and phone rates, but the existence of words such as *banana*, showing twice as many phones compared to the syllables, in contrast to the word *stretchmarks*, having five times more phones than syllables sufficiently explains the results in fig. 1.

Our hypothesis is that syllable rate as well as phone rate are involved in the perception of speech rate. To test this hypothesis a reference is needed. This we receive by conducting a perception experiment in which the subjects were instructed to compare and assess the speech rate of short speech signals. Using this reference enables us to calculate the linear correlation coefficients of syllable rate with perceptual speech rate, and likewise of phone rate with perceptual speech rate. Additionally we are able to approximate a linear combination of syllable rate and phone rate, thus obtaining a model for determining speech rate from syllable and phone rate.

Although there are automatic methods for the extraction of phone boundaries (Verhasselt&Martens [8, 1996]) and of syllable nuclei (Pfitzinger et al. [5, 1996]) that are not dependent on a speech recognition process, there are no algorithms for word boundary extraction, because every decision about word boundaries must fail if the language or the words spoken are unknown. We therefore concentrate on syllable and phone rate. Even though it was shown that it is possible to automatically estimate a good approximation to the local syllable rate and phone rate [4] our investigation is based on hand-labelled speech signals to evade this source of errors. We used the *PhonDatII* speech database dealing with railway information queries, read aloud by 10 male and 6 female German speakers, resulting in 40 minutes of labelled speech.

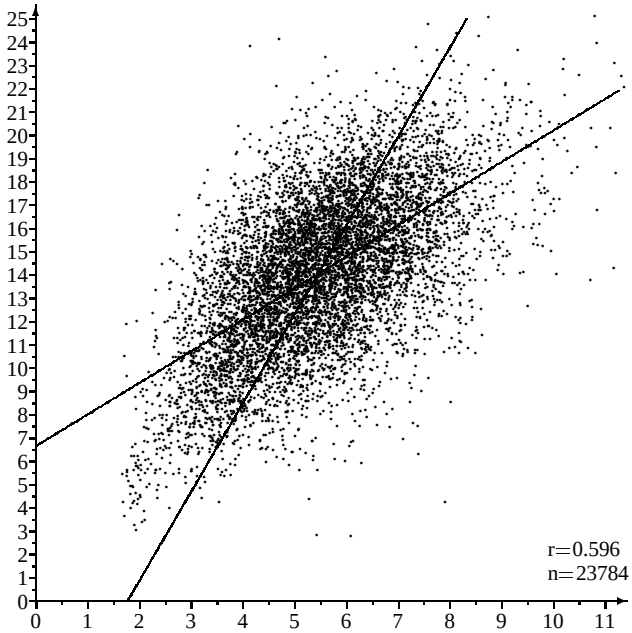

Figure 1: Scatter plot of the local phone rate based on manually segmented phone boundaries versus the local syllable rate based on manually segmented syllable nuclei. The ordinate illustrates the phone rate measured in phones per second and the abscissa illustrates the syllable rate measured in syllables per second.

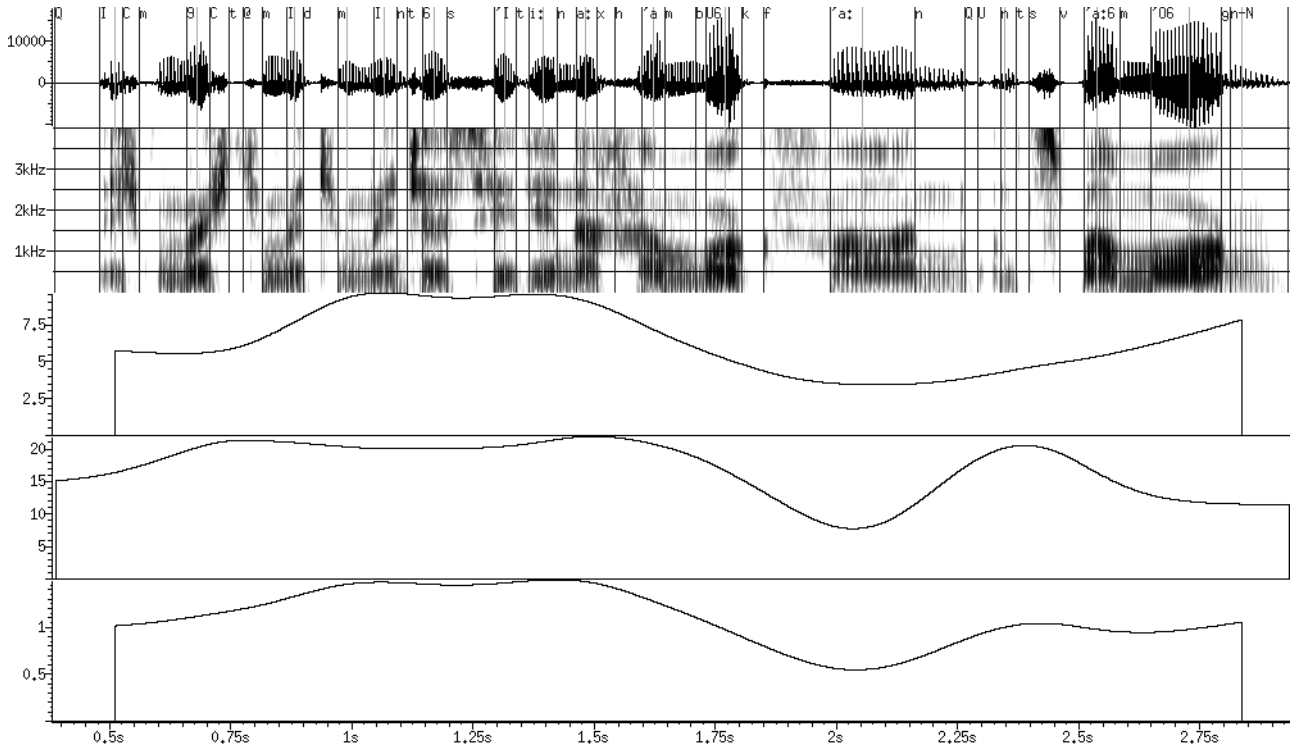

Figure 2: PhonDat sentence *dlma5270*, phone boundaries, syllable nucleus marks (light), syllable rate, phone rate, and speech rate.

2 ESTIMATING SYLLABLE RATE AND PHONE RATE

When estimating local syllable rate or phone rate three problems arise: (i) speech pauses have to be detectable because they should yield a zero rate, (ii) the analysis window size from which the rate is accumulated has to be larger than the longest speech segment and short enough to follow fast rate changes, and (iii) the choice of window type. We decided upon a Hanning-window because it down-weights the marginal segments and leads to less outliers.

2.1 How slow is speech?

The hand-labelled phone segmentation allowed us to find the phone with the longest duration. It was an /a/ in the German word *ja* (engl. *yes*) with a duration of 444 ms. Hence for segment durations greater than 450 ms the estimation procedure yields a zero phone rate.

With the hand-labelled syllable nuclei no decision over speech or pause between to marks was possible. Therefore a more extensive procedure to find the adequate maximum duration is required. We evaluated all speech segments, starting with syllable distances greater than 800 ms, then going to 700–800 ms and finally to 600–700 ms. We found speech signals even in the group with segments shorter than 700 ms, so we chose 625 ms as an optimum maximum syllable duration between accepting speech pauses and rejecting speech signals.

2.2 Some words on the window size

An informal perception experiment showed that using speech signal segments of less than 500 ms hindered the assessment of speech rate, while segments of more than 700 ms could contain strong changes in speech rate (e.g. the beginning of the segment was slow but the end was fast). Both effects dramatically increased the degree of difficulty of the assessment task.

In addition the analysis window must be of equal or greater length than the maximum syllable duration because otherwise it could

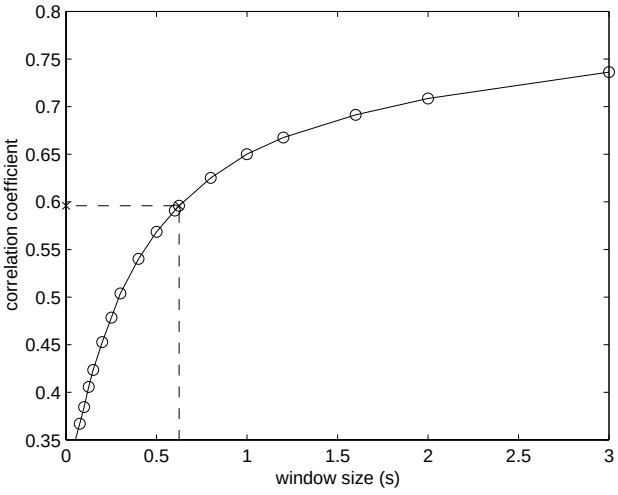

Figure 3: Correlation coefficient as a function of window size.

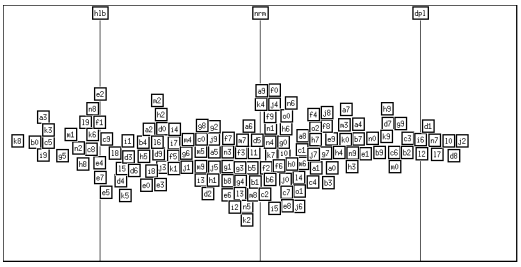

Figure 4: The completed answer sheet of one subject in the perception test.

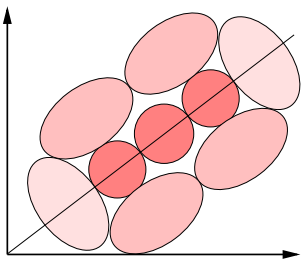

Figure 6: From each of the 16 speakers nine stimuli have been chosen according to the nine areas (compare to fig. 1).

not accumulate very large syllable nuclei from slow speech and would produce more outliers. If the analysis window were chosen too large, local extrema would become temporally shifted and evened out. Fig. 3 shows the linear correlation coefficient of the data in fig. 1 as a function of the analysis window size. We chose an analysis and perception window size of 625 ms. For estimating syllable rate and phone rate we used the same analysis window size ensuring that the results are based on the same analysis window content.

As expected the step size has no influence on the linear correlation coefficient. 100 ms seemed to be a good choice with sufficient temporal resolution. Fig. 2 illustrates an example signal analysis with a reduced step size of 1 ms to increase graphical resolution.

3 PERCEPTION EXPERIMENT

Five subjects participated in the listening experiment. They had to carry out a computer-aided interactive discrimination test using a desktop on which they could place and reorganize the labels of the stimuli and compare the acoustics of the stimuli as often as they wished (fig. 4). The subjects were instructed to place all stimuli along a rate-scale according to the speech rate and to finally check all labels for their correct order, and all perceptual speech rate differences between them for corresponding distances on the rate scale.

Three anchor stimuli were selected auditorily before the perception experiment to guarantee that the subjects would use the desktop space comparably. One of the three anchor stimuli was placed in the middle of the desktop, having a normal speech rate, the second, having roughly half of the normal speech rate, was placed on the left, and the third, having a doubled normal speech rate, was placed on the right (fig. 4). These stimuli served as a reference for the subjects to orientate to.

Every stimulus was cut from the speech database and had a duration of 625 ms. The beginning and the end of each stimulus were faded linearly during 10 ms to avoid click sounds. The duration of 625 ms lead to the perception of speech and not solely rhythmic sound structures which is important when assessing speech rate. Additionally, the level of all stimuli was equalized to eliminate this source of errors.

$9 \cdot 16 = 144$ stimuli (3.8% on the speech database) were manually selected paying attention to get the same distribution as shown in fig. 1, and to sufficiently cover all interesting cases (see fig. 6).

3.1 Approximation

The perception experiment yields assessment values only between 0 ($\approx$ less than half as fast as normal speech rate) and 2 ($\approx$ more than twice as fast as normal speech rate). Particularly the value 1 denotes a normal speech rate.

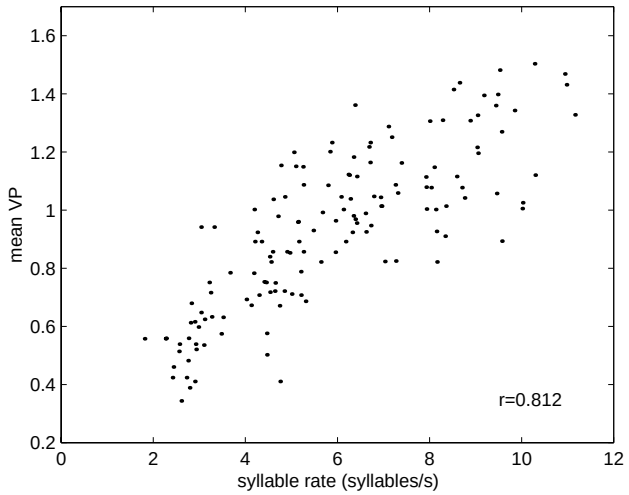

Figure 5: Scatter plot of the perception results vs. syllable rate.

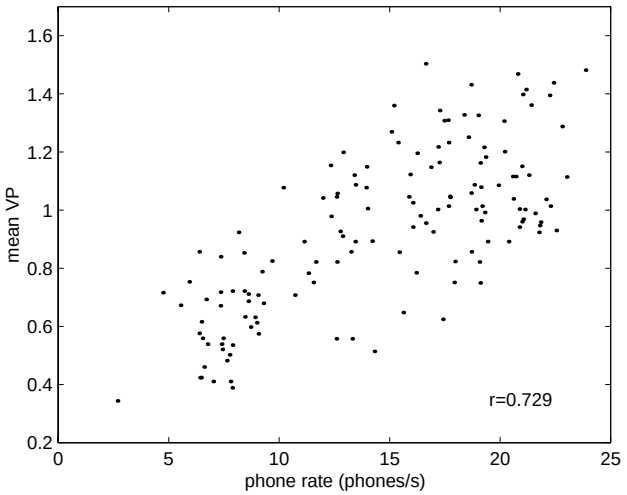

Figure 7: Scatter plot of the perception results vs. phone rate.

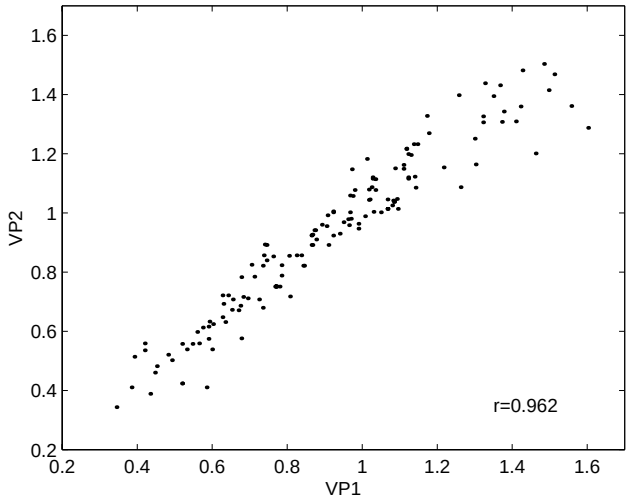

Figure 8: Scatter plot of the perception results of two subjects.

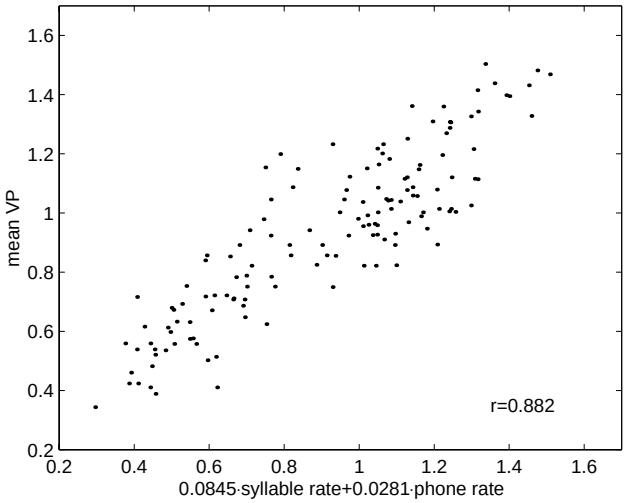

Figure 9: Scatter plot of the perception results vs. the estimated local speech rate.

We assume that local speech rate is representable by a linear combination of local syllable rate and local phone rate. Solving the following linear equation system

$$\begin{bmatrix} pr(1) & sr(1) \\ pr(2) & sr(2) \\ \vdots & \vdots \\ pr(M) & sr(M) \end{bmatrix} \begin{bmatrix} p \\ s \end{bmatrix} = \begin{bmatrix} pz(1) \\ pz(2) \\ \vdots \\ pz(M) \end{bmatrix},$$

where M is the number of stimuli, pr is the phone rate, sr is the syllable rate, and pz is the mean perception result for each stimulus from M, leads to the coefficients s, weighting the syllable rate, and p, which weights the phone rate.

4 RESULTS

The linear correlation coefficient $r = 0.81$ of the syllable rate with the mean perception results over all subjects (see fig. 5) is higher than the linear correlation coefficient $r = 0.73$ of the phone rate with the mean perception results (see fig. 7). Obviously the syllable rate is more suited to predict perceptual speech rate than the phone rate.

The perception results of the two subjects whose values correlate least are shown in fig. 8. The very high correlation coefficient $r = 0.96$ leads to the assumption that all subjects have a homogeneous intuition of how to assess speech rate.

The syllable rate is weighted by $s = 0.0845$ and the phone rate is weighted by $p = 0.0281$. If we take into account that the mean syllable rate (5.445 syllables/s) is 2.58 times slower than the mean phone rate (14.05 phones/s) then we conclude that perceptual speech rate is based on syllable rate by 54% and on phone rate by 46%. So phone rate has less influence on speech rate than syllable rate.

The linear combination method for estimating local speech rate correlates better ($r = 0.88$) with perceptual speech rate than syllable rate or phone rate do (fig. 9). The term *speech rate* should not be used if *syllable rate* or *phone rate* is meant because as we have shown now, these rates are quite different (see fig. 2).

REFERENCES

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

[2] Crystal, T. H.; House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, 88: 101–112.

[3] Ohno, S.; Fujisaki, H.; Taguchi, H. (1997). A method for analysis of the local speech rate using an inventory of reference units. In *Proceedings of EUROSPEECH '97*, vol. 1, pp. 461–464, Rhodes.

[4] Pfitzinger, H. R. (1996). Two approaches to speech rate estimation. In *Proceedings of SST '96*, pp. 421–426, Adelaide.

[5] Pfitzinger, H. R.; Burger, S.; Heid, S. (1996). Syllable detection in read and spontaneous speech. In *Proceedings of ICSLP '96*, vol. 2, pp. 1261–1264, Philadelphia.

[6] Samudravijaya, K.; Sanjeev, K. S.; Rao, P. (1998). Pre-recognition measures of speaking rate. *Speech Communication*, 24(1): 73–84.

[7] Siegler, M. A.; Stern, R. M. (1995). On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP95)*, vol. 1, pp. 612–615.

[8] Verhasselt, J. P.; Martens, J. P. (1996). A fast and reliable rate of speech detector. In *Proceedings of ICSLP '96*, vol. 4, pp. 2258–2261, Philadelphia.

[9] Wood, S. (1973). What happens to vowels and consonants when we speak faster? Working Papers 9, pp. 8–39, Phonetics Laboratory Lund University, Lund.