LOCAL SPEECH RATE PERCEPTION IN GERMAN SPEECH

Hartmut R. Pfitzinger
Department of Phonetics, University of Munich, Germany

ABSTRACT

A model for deriving perceived local speech rate directly out of the speech signal is developed based on perception experiments. Since local speech rate modifies acoustic cues (e.g. transitions and voice-onset time [5]), phones [1, 4, 7, 19], syllables [3, 15], 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 local syllable rate and the local phone rate, since earlier investigations [11, 12] strongly suggest that neither the syllable rate nor the phone rate on its own represent the speech rate sufficiently. In the literature [6] effects of F_0 level and F_0 movement on speech rate perception have been reported. Therefore we included these cues in our linear combination model.

Our results show 1) that the duration of speech stimuli has a strong influence on the perception of speech rate, 2) that the linear combination of local syllable rate and phone rate is well-correlated with perceptual local speech rate ($r = 0.91$), 3) that F_0-measurements could not increase the accuracy of the model, and 4) that our method is able to calculate the perceptual local speech rate and the relative local speech rate between two utterances.

1 INTRODUCTION

Since it is common knowledge that automatic speech recognition degrades significantly with fast speakers there is a growing interest in speech rate research (Cedergren&Perreault [2, 1994], Siegler&Stern [17, 1995], Verhasselt&Martens [18, 1996], Morgan et al. [8, 1997], Samudravijaya et al. [16, 1998]). However no homogeneous opinion exists of what speech rate actually is. Undoubtedly, a high speech rate is characterized by above-average syllable rates and phone rates, but previous research has shown a middlingly high correlation between local syllable rate and local phone rate ($r \approx 0.6$) indicating that the information contents of both differ (Pfitzinger [11, 1996]). 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, suggests the hypothesis that syllable rate as well as phone rate are involved in speech rate perception.

To test this hypothesis a reference is needed. This we obtained by conducting an interactive perception experiment in which twenty-nine subjects were instructed to place 141 short speech stimuli along an interval scale, and were also asked to ensure that perceived speech rate differences corresponded to distances on the interval scale. Using this reference enables us to calculate the linear correlation coefficients of local syllable rate or local phone rate with the mean perception results. Additionally we are able to approximate a linear combination of syllable rate and phone rate, thus obtaining a model for deriving perceived local speech rate directly out of the speech signal with sufficient accuracy.

In the literature (Kohler [6, 1986]) effects of F_0 level and F_0 movement on speech rate perception have been reported. Therefore we included these cues in our linear combination model. However, this did not lead to a significantly ($\hat{t}_{141} = 1.641 < Z_{0.1} = 1.645, n.s.$) increased linear correlation coefficient.

2 COMMON PERCEPTION TEST PROCEDURE

During each perception experiment the subjects had to carry out a computer-aided interactive discrimination test using a desktop metaphor 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. 1). 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 first 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. 1). These stimuli served as a reference for the subjects to orientate to.

Every stimulus was extracted from the speech database *PhonDatII* dealing with railway information queries, read aloud by 10 male and 6 female German speakers (resulting in 40 minutes of labelled speech). The beginning and the end of each stimulus were faded linearly over 10 ms to avoid click sounds. Additionally, the level of all stimuli were equalized to eliminate this possible source of error.

3 FIRST PERCEPTION EXPERIMENT

The first perception experiment deals with the influence of the stimulus duration on the perception of speech rate. 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).

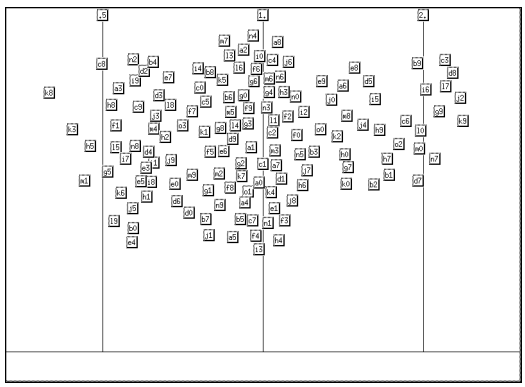

Figure 1: The completed answer sheet of one of the 29 subjects in the second perception experiment.

Duration (ms)	225	425	625	825	1025
$\mu_2 - \mu_1$	32.8	14.6	0.102	5.99	-8.14
σ_1	14.4	13.5	13.7	12.0	12.3
σ_2	20.3	20.6	21.8	17.4	18.8
F	1.42	1.53	1.59	1.45	1.52
$F_{0.05;119/119} = 1.36$	$\star$	$\star$	$\star$	$\star$	$\star$
$\hat{t}$	14.4	6.50	0.043	3.11	3.96
f	214	205	200	211	205
α	0.001	0.001	0.1	0.005	0.001
$t_{f;\alpha}$	3.34	3.34	1.65	2.84	3.34
Significance	$\star\star\star$	$\star\star\star$	n.s.	$\star\star$	$\star\star\star$

Table 1: Differences in speech rate perception between stimuli with 625 ms duration and with variable length. F-test for homogeneity of two samples and t-test for inhomogeneous variances with correction of the degrees of freedom (after Welch).

3.1 Procedure

Ten subjects participated in two paired listening experiments. In each of them they had to carry out the above mentioned computer-aided interactive discrimination test. In the first perception experiment 60 stimuli with an equal duration of 625 ms spoken by 16 speakers at different speech rates were presented. In the second test these stimuli were grouped into five classes of twelve stimuli, while ensuring that the classes were composed comparably. Six months intervened between the two tests.

3.2 Results and discussion

Table 1 lists the mean differences between corresponding classes of the paired listening tests, and the standard deviations for each class. Figure 2 shows these data graphically. The third classes, both consisting of the same stimuli with 625 ms duration, are suitable for testing the consistency of the subjects. There is a significantly greater standard deviation of the class of the variable duration experiment. This might be explained by the increased difficulty in identifying a speech rate when there are various stimulus durations available for discrimination. But there is no mean difference between the two classes indicating that the subjects are capable of, and consistent in, perceiving and assessing local speech rate. Shortening of the same stimuli leads to a highly significantly increased difference. This means the shorter the segment duration the faster the perceived speech rate. For durations longer than 625 ms this does not hold.

4 SECOND PERCEPTION EXPERIMENT

The aim of the second perception experiment was to obtain a perceptual reference for local speech rate. Twenty-nine subjects participated in the listening test. They had to carry out the above mentioned computer-aided interactive discrimination test. 141 stimuli each having a duration of 625 ms were presented. We were careful to use a broad distribution of speech rates and to sufficiently cover all interesting cases (for details see [12, 1998] and [13, 1999]).

4.1 Acoustic analysis

There are automatic methods for the extraction of phone boundaries (Verhasselt&Martens [18, 1996]) and of syllable nuclei (Pfitzinger et al. [15, 1996]), and it was shown that it is possible to automatically estimate a good approximation to the local syllable rate and phone rate [11]. Nevertheless our investigation is based on manually labelled speech signals to evade this source of errors.

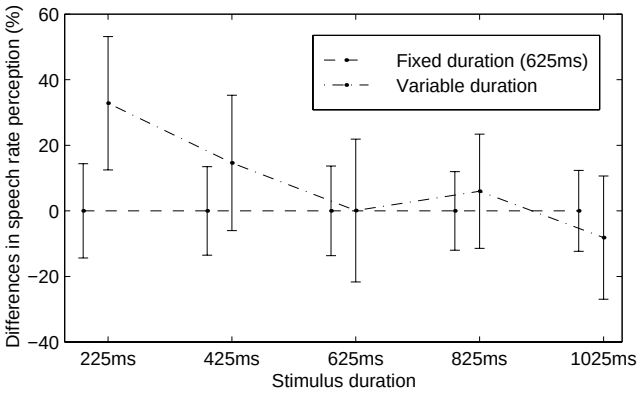

Figure 2: Differences in speech rate perception between stimuli with 625 ms duration and with variable length.

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 is difficult. We decided upon a Hanning-window because it down-weights the marginal segments and leads to less outliers. The window size was set to 625 ms. The selection of the window size and type is discussed in detail in Pfitzinger [12, 14].

4.2 Refining the local speech rate prediction model

As shown earlier (Pfitzinger [12, 1998]) perceptual local speech rate is predictable by an acoustic model with fair accuracy. The proposed model consisted of a linear combination of local syllable rate and local phone rate. To further reduce the deviation of the prediction, utilizing results from the literature, we now introduce measurements of F_0. This is known to be a cue for speech rate perception (Kohler [6, 1986]). In addition, we allow for rotation of the perception data. Solving the following overdetermined linear equation system

$$\begin{bmatrix} pr(1) & sr(1) & f0(1) & 1 \\ pr(2) & sr(2) & f0(2) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ pr(M) & sr(M) & f0(M) & 1 \end{bmatrix} \begin{bmatrix} p \\ s \\ f \\ c \end{bmatrix} = \begin{bmatrix} pz(1) \\ pz(2) \\ \vdots \\ pz(M) \end{bmatrix},$$

where M is the number of stimuli, pr is the phone rate, sr is the syllable rate, $f0$ is the mean F_0, and pz is the mean perception result for each stimulus from M, leads to the coefficients s, p, and f which respectively weight the syllable rate, the phone rate, and the mean fundamental frequency. c compensates for rotation.

4.3 Results of the perception experiment and the refinement

In a one-to-many comparison the twenty-nine subjects showed linear correlation coefficients between 0.83 and 0.96 with a mean of 0.91. This leads to the assumption that all subjects have a homogeneous intuition of how to assess speech rate. As a reference for the following investigations the mean perception result is obtained by averaging the judgements of the twenty-nine subjects for each of the 141 stimuli. Table 2 shows the linear correlation coefficients of this reference with several acoustic analyses. The very low values

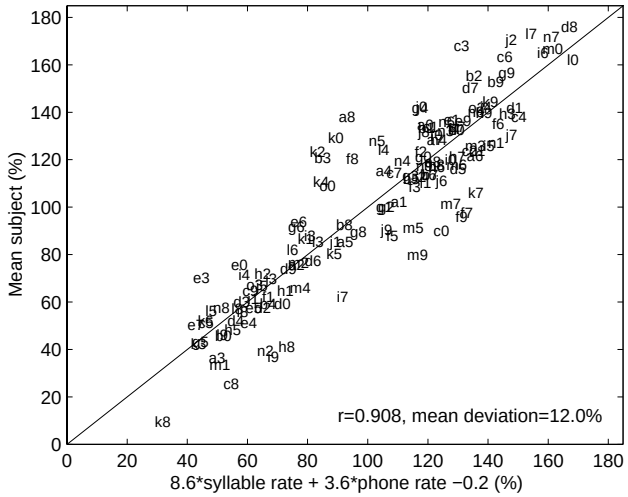

Figure 3: Scatter plot of the mean local speech rate perception results vs. the results of the acoustic model excluding F_0.

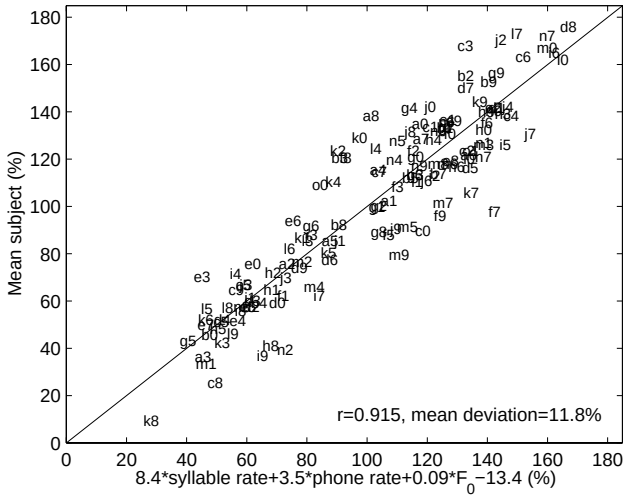

Figure 4: Scatter plot of the mean local speech rate perception results vs. the results of the acoustic model including F_0.

occuring with F_0-measurements are worth mentioning. Obviously, subjects take F_0 into account only if other acoustic cues discriminating speech rate are absent.

The linear correlation coefficient $r = 0.796$ of the syllable rate with the mean perception results over all twenty-nine subjects is not significantly different from the linear correlation coefficient $r = 0.792$ of the phone rate with the mean perception results ($\hat{t}_{141} = 0.13 < Z_{0.05} = 1.96, n.s.$, see table 2). In contrast to earlier investigations [12] we must state that syllable rate and phone rate are equally badly-suited to prediction of perceptual local speech rate. Since $\hat{t}$ is very close to zero it seems unlikely that the difference would become significant with a larger number of subjects.

The local speech rate predictions correlate significantly ($\hat{t}_{141} = 6.31 > Z_{0.001} = 3.29, \star\star\star$) better with perceptual speech rate than syllable rate or phone rate do (table 2). The term *speech rate* should not be used if *syllable rate* or *phone rate* is meant because as we have shown, these rates are quite different. The model covering F_0 (fig. 4) is not significantly ($\hat{t}_{141} = 1.641 < Z_{0.1} = 1.645, n.s.$) better than the model ignoring F_0 (fig. 3), which possibly could alter under other experimental conditions.

The approximated coefficients of the speech rate models and their 95%-confidence intervals are shown in table 3. For the values of the coefficients p and s there is 95% certainty that their real values deviate less than 6.4% and 8.2%. As c compensates for rotation along the rate-scale, its confidence interval relates to the mean scale value, viz. 100. Hence its real value deviates less than 6%, 95% of the time. These results seem to be accurate enough to work with in future research.

Acoustic analysis for each stimulus	Linear correlation coefficient with mean local speech rate perception
Mean F_0	0.230
Mean F_0 − mean F_0 of corresponding speaker	0.344
Standard deviation of F_0	0.026
Linear approximation coefficient of F_0	-0.077
Syllable rate	0.796
Phone rate	0.792
Model 1: syllable rate & phone rate	0.908
Model 2: syllable rate & phone rate & mean F_0	0.915

Table 2: Linear correlation coefficients of the mean perception result with various acoustic analyses of the 141 speech stimuli.

5 THIRD EXPERIMENT

To substantiate our local speech rate prediction model we used a method introduced by Ohno and Fujisaki [9, 1995][10, 1996] for comparison, which is based on dynamic time warping. It approximates the *relative local speech rate* by the gradient of the warping function. The weakness of this method is that both underlying utterances have to contain the same words.

We can overcome this limitation of the method using our local speech rate prediction model. In figure 5 one example for this application is shown. We calculate the perceptual local speech rate of both utterances, time-warp one of the resulting functions to the other, and finally calculate the quotient of them (fig. 5-5.). Since the distortions occuring are due to the time-warping process and, furthermore, most of the differences are smoothing artifacts, it is to be expected that both methods are equivalent.

	Coefficient	Mean value	95% confidence interval
Model 1 (fig. 3)	p	8.59	0.52
	s	3.58	0.29
	c	-0.19	4.91
Model 2 (fig. 4)	p	8.41	0.54
	s	3.54	0.29
	f	0.091	0.028
	c	-13.35	5.97

Table 3: Mean value and 95% confidence interval of the coefficients obtained from 29 subjects for the two speech rate models (excluding and including F_0).

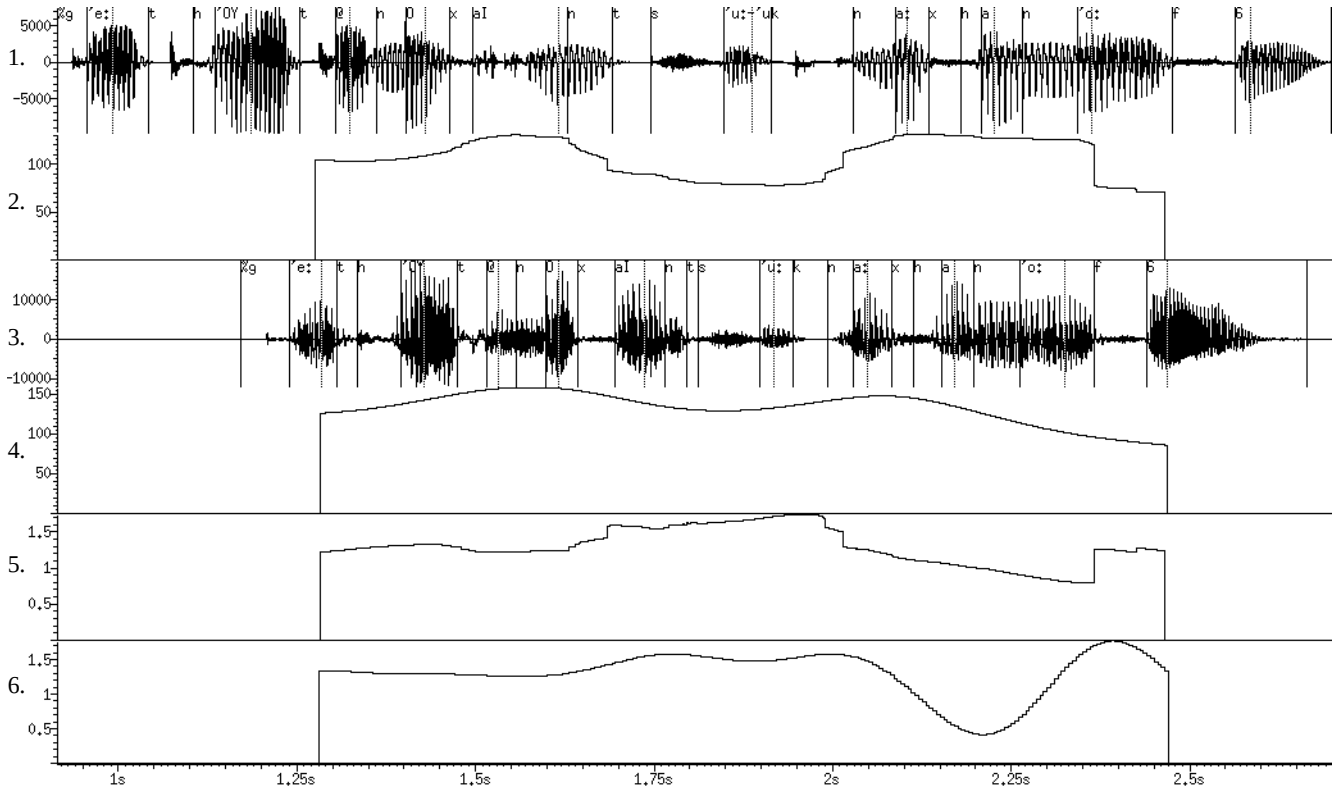

Figure 5: 1. PhonDatII sentence *wsen5050*, 2. perceptual local speech rate of 1 aligned to 3 through dynamic time warping, 3. PhonDatII sentence *hpta5050* having the same orthography as 1, 4. perceptual local speech rate of 3, 5. relative local speech rate calculated by dividing 4 by 2, and 6. relative local speech rate estimated using dynamic time warping of 1 and 3. Note the similarity of the signals 5 and 6.

REFERENCES

[1] Byrd, D.; Tan, C. C. (1996). Saying consonant clusters quickly. *Journal of Phonetics*, 24: 263–282.

[2] 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.

[3] 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.

[4] Gottfried, T. L.; Miller, J. L.; Payton, P. E. (1990). Effect of speaking rate on the perception of vowels. *Phonetica*, 47: 155–172.

[5] Kessinger, R. H.; Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: some implications for perception studies. *Journal of Phonetics*, 26: 117–128.

[6] Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: duration, F_0 movement, and F_0 level. *Language & Speech*, 29: 115–139.

[7] Miller, J. L.; O'Rourke, T. B.; Volaitis, L. E. (1997). Internal structure of phonetic categories: effects of speaking rate. *Phonetica*, 54: 121–137.

[8] Morgan, N.; Fosler, E.; Mirghafori, N. (1997). Speech recognition using on-line estimation of speaking rate. In *Proceedings of EUROSPEECH '97*, vol. 4, pp. 2079–2082, Rhodes.

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

[10] Ohno, S.; Fukumiya, M.; Fujisaki, H. (1996). Quantitative analysis of the local speech rate and its application to speech synthesis. In *Proceedings of ICSLP '96*, vol. 4, pp. 2254–2257, Philadelphia.

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

[12] Pfitzinger, H. R. (1998). Local speech rate as a combination of syllable and phone rate. In *Proceedings of ICSLP '98*, vol. 3, pp. 1087–1090, Sydney.

[13] Pfitzinger, H. R. (1999). Acoustic and perceptual measurements of local speech rate. Forschungsberichte (FIPKM) 39, Institut für Phonetik und Sprachliche Kommunikation der Universität München, to appear.

[14] Pfitzinger, H. R. (1999). Transformation of time-varying frequencies between frequency- and time-domain using non-rectangular windows. Forschungsberichte (FIPKM) 39, Institut für Phonetik und Sprachliche Kommunikation der Universität München, to appear.

[15] 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.

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

[17] 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.

[18] 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.

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