Local Speech Rate: Relationships between Articulation and Speech Acoustics

Hans G. Tillmann & Hartmut R. Pfitzinger

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

ABSTRACT

The relationship between the displacements, velocities, and accelerations of the articulators and speech rate is investigated experimentally. Especially, we address the question as to whether it is possible to predict the speech rate from articulatory kinematics by means of a linear regression model. Two different kinds of corpora have been recorded. The first consists of only two short sentences differing in their vowel/consonant quotient and read aloud with seven different speech rates. The second was two readings of a 5-minute short story. While the first corpus supported a close relationship, the second revealed a correlation coefficient of only $r = 0.5$ which provides evidence against a close relationship between kinematic variables and speech rate.

1 INTRODUCTION

A central but not yet really solved problem of spoken language processing is to predict the prosodic modifications of the phonetic form of words in connected speech. The variety of such modifications becomes especially clear if we look at the differences between (a) those phonetic forms which a speaker explicitly demonstrates in isolated word pronunciations and (b) all the other forms which we obtain as soon as the same speaker uses these words in connected speech, read or spontaneously produced.

Many investigations focus on the connection of speech rate variation and articulatory coordination [1–9]. What do the articulators do when we speak faster? Is there a reduction of the displacement of articulators or an increase of their velocities [10, 11]? And is there an increased overlap and/or a shorter duration of articulatory gestures [12]? Both would increase the mean velocities over all active articulators and hence presumably the articulatory effort [13] which is supposed to be a combination of velocities and accelerations.

The present study examines how the prosodic variation of speech rate within an utterance is related to the underlying articulatory kinematics. The focus is on "momentary" or local speech rate which can, for German speech, be determined by a new method developed by the second author [14, 15, 16]. It is based on a linear combination of both local syllable rate and local phone rate, and correlates well with perceived local speech rate ($r = 0.91$). A moving Hanning window with a duration of 625 ms produces a continuous speech rate representation, prosodically varying between local maxima and minima (being comparable to F0-contours).

2 FIRST EXPERIMENT

This investigation was aimed at collecting articulatory and acoustic speech data providing controlled speech rate variation and thus enabling us to develop and test first hypotheses about the relation between kinematic variables and speech rate.

A male speaker produced two German sentences with seven different metronomically controlled syllable rates between 72 and 200 BPM (2.4–6.67 syl/s). One sentence consists of 6 syllables and 12 phones having a syllable/phone ratio of 1:2 and a vowel/consonant quotient of V/C=6/6=1.0 while the other consists of 6 syllables and 22 phones having a syllable/phone ratio of 1:3.67 and a V/C quotient of 6/16=0.375.

The articulatory data were provided by an EMMA measurement (electro-magnetic mid-sagittal articulograph, Carstens AG-100 at 500 Hz sampling rate) of 7 articulators: jaw, both lips, tongue tip, blade, dorsum, and back (see Fig. 1).

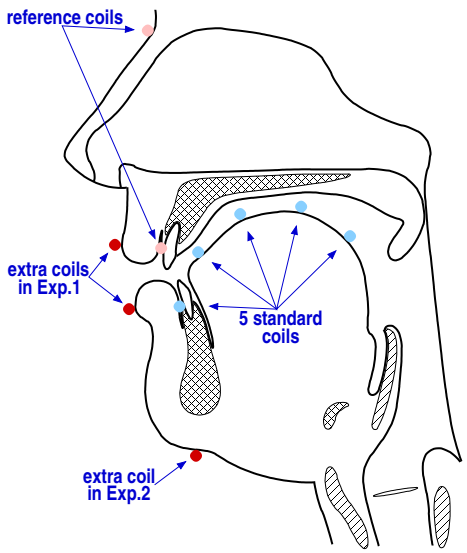

Figure 1: Standard positions of the EMMA receiver coils and additional positions in the two experiments.

ISBN 1-876346-48-5

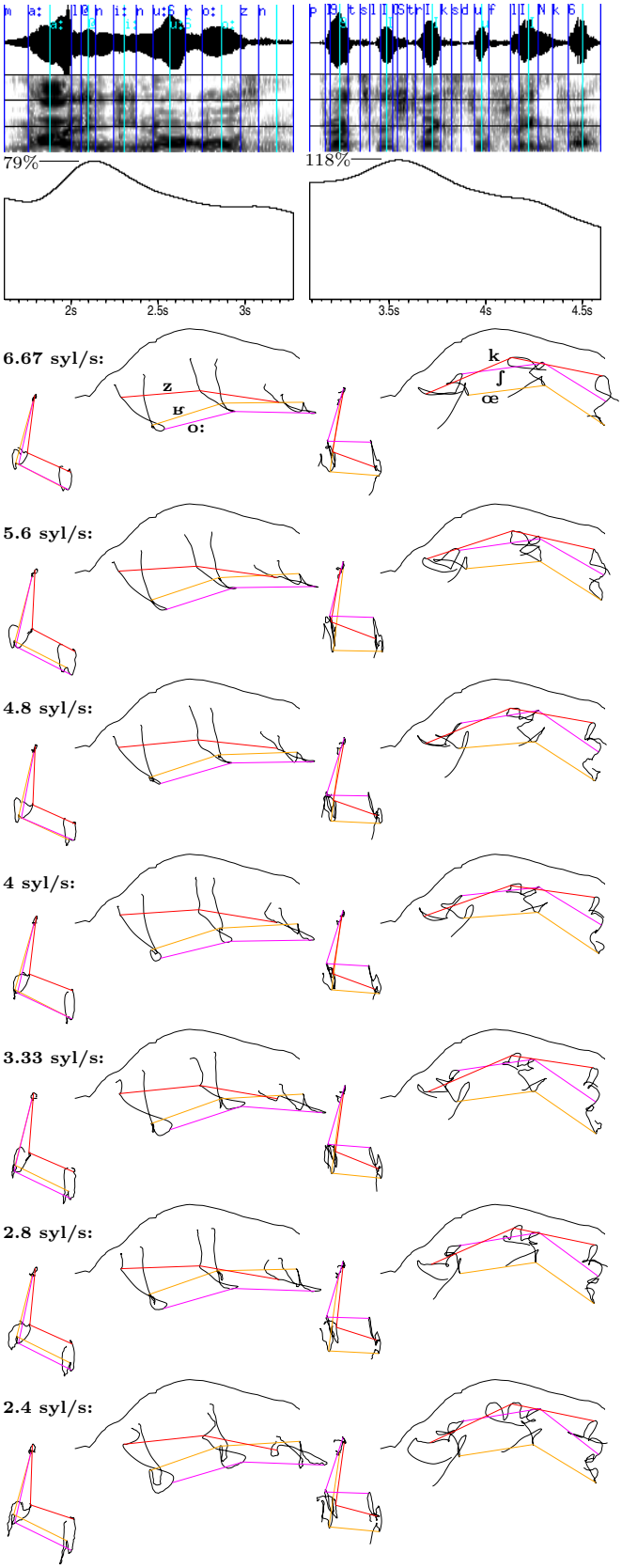

Figure 2: *Left:* utterance [ma:ləni:nuɐʁo:zn], *right:* utterance [plœtslɪçʃtχɪksdu:flɱkɐ]. *From top to bottom:* oscillogram and spectrogram (with hand-labelled phone boundaries and syllable nuclei), estimated perceptual local speech rate (PLSR), and seven sagittal images of the movements of both lips, jaw, tongue tip, blade, and dorsum (see Fig. 1) during the underlined stretches of the utterances.

2.1 RESULTS AND DISCUSSION

A comparison of the sagittal images in Fig. 2 shows that on the whole the range in which each pellet is moved remains almost constant compared with the large amount of speech rate variation which is present in the utterances. On closer examination, the shapes of the trajectories become more complex (i.e. they show more bends and loops), the slower the speech rate is, or vice versa, a high speech rate leads to simplified trajectories. This allows the conclusion that displacement is at least partially influenced by speech rate.

These observations lead to the hypothesis that the trajectories of the articulators, when producing fast speech, are characterized by shorter total distances traversed by the pellets. But since, in the case of fast speech, the duration of the trajectories decreases more than the total distance (see Fig. 2), the velocity of the articulators and likewise their acceleration increases. This suggests the hypothesis that speech rate is mainly related to the velocity and acceleration of the articulators. The correlation coefficients between sentence duration and mean rectified velocity ($r = 0.946$) or acceleration ($r = 0.929$), estimated over all articulators, support the hypothesis.

The limited amount of data as well as the obvious unnaturalness of speaking to the rhythm of a metronome may severely degrade the validity of these preliminary results. Therefore further investigation of the hypotheses is required.

3 SECOND EXPERIMENT

This study explores the hypotheses raised in the previous experiment that the articulatory movements, their velocities and accelerations are in close relation with the speech rate. Especially, the question as to whether a local approach is suited to predict the estimated perceptual local speech rate (PLSR) by means of a linear regression model is addressed.

The underlying speech data consists of a 5-minute short story with strong natural speech rate variations. It was read aloud by a male German speaker twice: with normal and with loud voice. An EMMA measurement (250 Hz sampling rate) provided articulatory data of 6 articulators: the jaw, the chin, and the tongue tip, blade, dorsum, and back (see Fig. 1).

The local syllable rate, phone rate, and perceptual local speech rate are derived directly from the speech signal [14, 15, 16]. The rectified velocities and accelerations are calculated from the X- and Y-movements of each articulator. Finally, the locally smoothed velocities and accelerations are estimated using a Hanning window with 625 ms duration. This duration is most suitable to account for speech rate changes as discussed in [14, 15, 16].

3.1 RESULTS AND DISCUSSION

To test the interdependence among the articulatory data, the correlations between the movements of the articulators and their rectified first and second deriva-

ISBN 1-876346-48-5

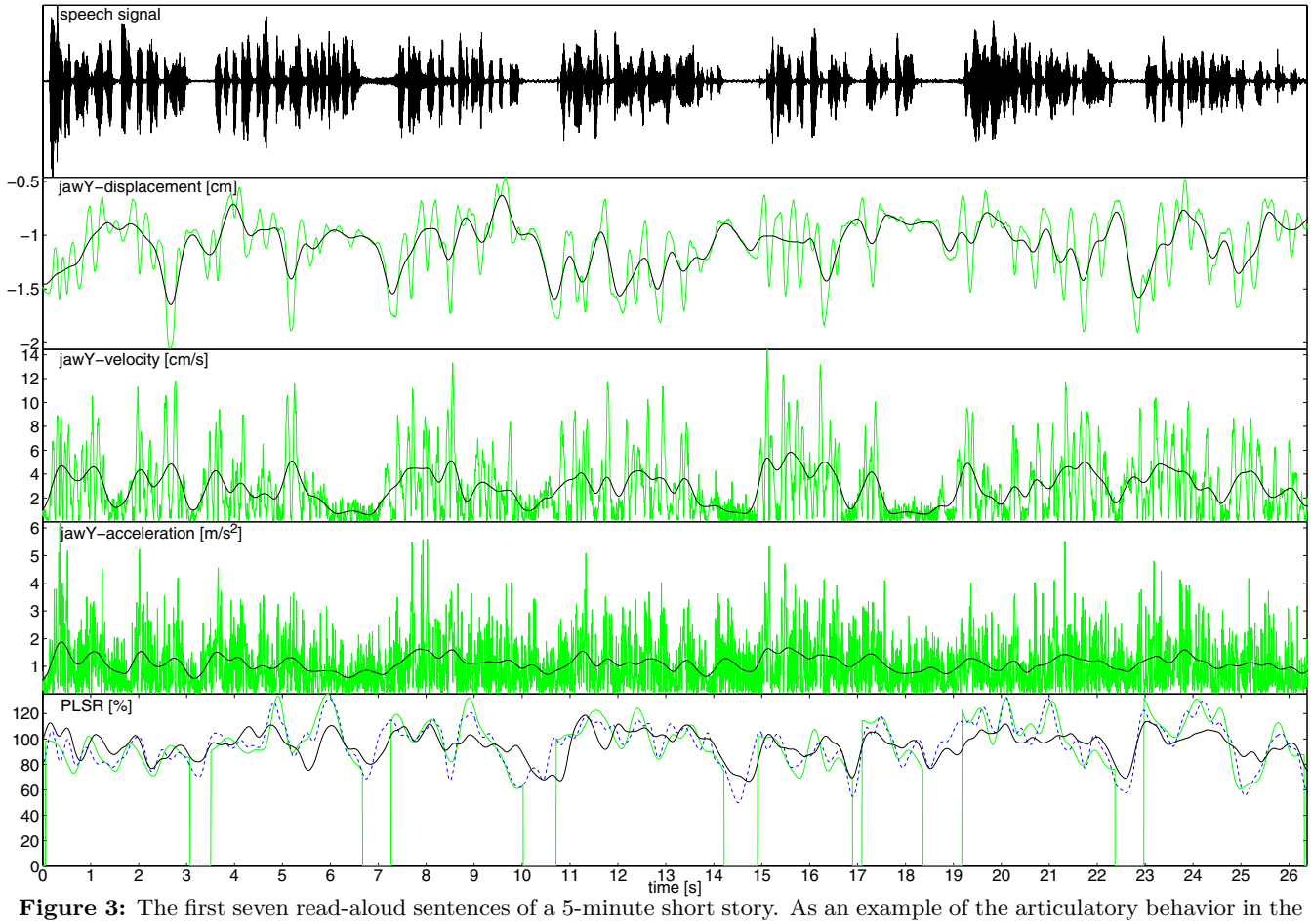

Figure 3: The first seven read-aloud sentences of a 5-minute short story. As an example of the articulatory behavior in the time domain, the Y-displacements of the jaw and its velocities and accelerations are shown in the second, third, and forth panel, respectively. The black contours show their smoothed versions, respectively. The bottom panel shows the estimated perceptual local speech rate and the predicted (*black*, $r = 0.504$) and over-adapted local speech rate (*dotted*, $r = 0.907$).

tives (velocities and accelerations) as well as between the smoothed versions were estimated:

Correlations (r)	Velocities	Accelerations
Displacements	-0.03	-0.04
Velocities		0.252

Correlations (r)	Smoothed vel.	Smoothed accel.
Smoothed displacements	-0.04	-0.07
Smoothed velocities		**0.744**

Only the smoothed rectified velocity correlates remarkably high ($r = 0.744$) with smoothed rectified acceleration. Nevertheless, the information contents of these six data sources differ. Therefore, their correlation coefficients with local phone rate, local syllable rate, and estimated perceptual local speech rate (PLSR) were estimated by means of linear regression:

Correlations (r)	phone rate	syllable rate	PLSR
Phone rate		0.562	0.898
Syllable rate			0.869
Displacements	**0.304**	0.251	0.285
Smoothed displ.	**0.364**	0.314	0.349
Velocities	0.124	0.108	0.127
Smoothed vel.	0.277	0.272	**0.300**
Accelerations	0.079	0.080	0.086
Smoothed accel.	0.196	**0.254**	0.243
Displ.+vel.+accel.	**0.351**	0.276	0.327
Smoothed d.+v.+a.	**0.504**	0.453	**0.504**

The tangential velocities and accelerations yielded correlation coefficients which were less than or equal to those listed above and therefore were skipped.

Obviously, no articulatory-based data or transformation showed a higher correlation coefficient than $r = 0.504$. Especially, the un-smoothed versions hardly correlate, even when combining them in a linear regression model ($r = 0.351$). The bottom panel of Fig. 3 shows that the predicted perceptual local speech rate differs considerably from the estimated contour. A correlation coefficient of 0.504 explains only 25% of the observed variation and is therefore not applicable in a prediction task. These results indicate that perceptual local speech rate, phone rate or syllable rate seem to be unpredictable from articulatory kinematics.

By contrast, each 26-second stretch out of the 10 minutes of read speech leads to a correlation coefficient between 0.907 and 0.996. But since in this case the paradigm of separate training and evaluation corpora was violated and, additionally, the data was too small for developing a valid prediction model the typical scenario of over-adaption took place leading to extremely high correlations. This indicates that 26 seconds of even a phonetically rich corpus are not sufficient to generalize the results. At least when using 5 minutes of read speech or more the correlation coefficients became stable on the evaluation data.

ISBN 1-876346-48-5

4 DISCUSSION

The local syllable rate might have been expected to be better predicted from kinematic variables than the local phone rate. But as shown in Fig. 4 local phone rate is slightly but significantly better correlated with the articulatory data. A probable explanation is that articulatory gestures are organized rather on a phone level than on a syllable level. Even the jaw displacement is weakly correlated with the supposed opening and closing gestures of syllables. This may at first seem to be surprising. But the jaw actually is lower for an [ɑ] or [ɔ] than for an inter-syllabic [f] which in turn might be produced with a lower jaw than an [i] or [e]. Thus, the jaw displacement is largely dominated by the 'jaw height'-constraints of the underlying phonemes.

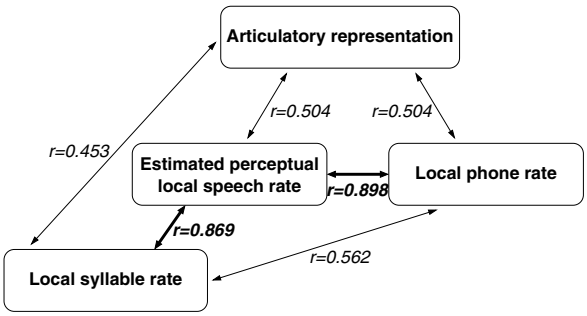

Figure 4: Correlations between a linear combination of the smoothed displacement, velocity, and acceleration of the articulators and three representations of speech rate.

5 CONCLUSIONS

When using only a small amount of speech data this investigation was able to reproduce close relationships between articulatory data and speech rate which have been reported in the literature. But as soon as a phonetically rich corpus of at least 5 minutes of read speech is used no relationship could be observed any longer. We would not expect the prediction model to considerably improve through additionally including lip or velum measurements. The present correlation coefficient of $r = 0.504$ between perceptual local speech rate and a linear combination of smoothed displacements of six articulators and their velocities and accelerations is too far from achieving an acceptable local speech rate prediction quality.

Two additional analyses of the two short stories read with normal and loud voice revealed no significant effect of loud speech on the correlation coefficients shown in Fig. 4. Due to the high individuality of the production strategies for speaking faster [10] we would expect even smaller correlation coefficients in a multi-speaker and spontaneous speech articulatory speech database. These experiments as well as a detailed phone-specific analysis of the data remain to be done.

ACKNOWLEDGMENTS

We thank Phil Hoole for conducting the EMMA recordings and Christian Kroos for his carefully designed phonetically rich short story "*Moorsoldaten*".

REFERENCES

[1] T. Gay, T. Ushijima, H. Hirose, and F. S. Cooper, "Effect of speaking rate on labial consonant-vowel articulation," *J. of Phonetics*, vol. 2:47–63, 1974.

[2] B. Tuller, K. S. Harris, and J. A. S. Kelso, "Stress and rate: Differential transformations of articulation," *J. of the Acoustical Society of America*, vol. 71, no. 6, pp. 1534–1543, 1982.

[3] Y. Sonoda, "Effect of speaking rate on articulatory dynamics and motor event," *J. of Phonetics*, vol. 15:145–156, 1987.

[4] G. Wieneke, P. Janssen, and H. Belderbos, "The influence of speaking rate on the duration of jaw movements," *J. of Phonetics*, vol. 15:111–126, 1987.

[5] O. Engstrand, "Articulatory correlates of stress and speaking rate in Swedish VCV utterances," *J. of the Acoustical Society of America*, vol. 83, no. 5, pp. 1863–1875, 1988.

[6] J. E. Flege, "Effects of speaking rate on tongue position and velocity of movement in vowel production," *J. of the Acoustical Society of America*, vol. 84, no. 3, pp. 901–916, 1988.

[7] S. G. Adams, G. Weismer, and R. D. Kent, "Speaking rate and speech movement velocity profiles," *J. of Speech and Hearing Research*, vol. 36, pp. 41–54, 1993.

[8] S. Shaiman, S. G. Adams, and M. D. Z. Kimelman, "Timing relationships of the upper lip and jaw across changes in speaking rate," *J. of Phonetics*, vol. 23:119–128, 1995.

[9] T. Okadome, T. Kaburagi, and M. Honda, "Relations between utterance speed and articulatory movements," in *Proc. of EUROSPEECH '99*, Budapest, 1999, vol. 1, pp. 137–140.

[10] D. P. Kuehn and K. L. Moll, "A cineradiographic study of VC and CV articulatory velocities," *J. of Phonetics*, vol. 4:303–320, 1976.

[11] K. G. Munhall, D. J. Ostry, and A. Parush, "Characteristics of velocity profiles of speech movements," *J. of Experimental Psychology: Human Perception and Performance*, vol. 11, no. 4, pp. 457–474, 1985.

[12] D. Byrd and C. C. Tan, "Saying consonant clusters quickly," *J. of Phonetics*, vol. 24:263–282, 1996.

[13] W. L. Nelson, "Physical principles for economies of skilled movements," *Biological Cybernetics*, vol. 46, pp. 135–147, 1983.

[14] H. R. Pfitzinger, "Local speech rate as a combination of syllable and phone rate," in *Proc. of ICSLP '98*, Sydney, 1998, vol. 3, pp. 1087–1090.

[15] H. R. Pfitzinger, "Local speech rate perception in German speech," in *Proc. of the XIVth Int. Congress of Phonetic Sciences*, San Francisco, 1999, vol. 2, pp. 893–896.

[16] H. R. Pfitzinger, "Phonetische Analyse der Sprechgeschwindigkeit," FIPKM 38, Institut für Phonetik und Sprachliche Kommunikation, Munich, 2001.