The role of articulation rate in distinguishing fast and slow speakers

Jacques Koreman

Institute of Phonetics
Saarland University, Germany
jkoreman@coli.uni-saarland.de

Abstract

This article discusses differences in articulation rate between fast and slow speakers in a production experiment. It is shown that fast and slow speakers differ in their articulation rates, both in terms of the number of phones in the canonical form (intended rate) as well as the number of phones present in the actual realization (realized rate). The articulatory precision index, which indicates the relative deletion rate, also differs for these speakers. The same differences are observed for fast and slow inter-pause stretches in a large German database of spontaneous speech. Both in the database and for the production experiment, however, there is considerable overlap between the measurements for fast and slow speakers. This shows that other factors also play a role in distinguishing fast and slow speakers or inter-pause stretches. The relationship between these factors and the articulation rates is discussed.

1. Introduction

The variation of speaking rate *within* speakers, for instance as one of the speech properties used for signalling prosodic units [1], has received considerable attention. Crystal and House [2] showed that the average syllable duration within an inter-pause stretch (ips) or run correlates strongly across speakers for comparable runs, and mainly depends on the number of phones per syllable and the proportion of stressed to unstressed phones or syllables.

Despite the clear relationship between within-speaker variation in speech rate and phonetic structure, there also seem to be differences between speakers which cannot be explained by other phonetic properties. As Crystal and House [2] (pp. 107-8) note, fast and slow speakers differ in their average syllable duration (ASDs). This difference cannot be explained by the phonetic properties they measured, as is shown by comparison of the speakers in their Table VII. This is interpreted as an indication of individual differences in speech rate between speakers.

This paper compares the articulation rates of fast and slow speakers, where the labels *fast* and *slow* are based on the perception of the speaker in an informal consensus judgment by six advanced students of phonetics. These subjects also participated in two perception experiments focusing on perceived rate for different intended and realized phone rates, the results of which were reported in [3].

2. Method

2.1. Speakers

Twelve subjects were invited from the experimenters' circle of acquaintances to participate in the recordings because the experimenters, who are all advanced students of phonetics, considered them to have an extreme speech rate. Eight of the speakers were judged to be particularly fast, four were considered particularly slow speakers. The speech rate judgment by the experimenter who invited a subject to participate was confirmed in a consensus judgment by the other five experimenters. Five of the speakers were male and seven were female. The speakers were in the age range 20-80, with an average age of 33.

2.2. Recordings

Each subject first read a German text aloud ("Die Buttergeschichte") and then retold the story in his/her own words. Both versions of the text (read and retold) were recorded onto audio tape and then digitized at a sampling frequency of 10 kHz and with a 16-bit amplitude resolution.

2.3. Measurements

The recordings were divided into runs or inter-pause stretches (ips) and both the intended and the realized phone rates were computed for each ips. The intended phone rate (ipr) was determined as the number of realisable segments per ips, taking the full canonical form as a starting point. The realized phone rate (rpr) was also computed. This is the number of phones actually present in the pronounced form of the ips. The intended and realized forms can often differ quite substantially, due to reductions leading to the complete deletion of phones or even syllables. For further discussion of the terms intended and realised form, see [3].

Since only ips with a duration of more than 0.5 seconds were analysed, no normalisation for differences in intrinsic phone durations was carried out [1], on the assumption that the different phone types are represented roughly equally within each ips. Furthermore, the measures are averaged across all ips for each speaker, further reducing the effect of differences between the ips in terms of the phones they consist of.

In addition to the intended and realized phone rates, an articulatory precision index (api) was computed by dividing the realized phone rate by the intended phone rate. A lower api value indicates more deletions. Reductions which do not entail the complete deletion of a phone are not taken into account (although we expect they correlate with full deletions, which represent one extreme end of the reduction scale), and were not part of the investigations reported here. The duration of each ips was also measured.

3. Analysis of the ips

This section tries to answer the question whether there are consistent differences between fast and slow speakers in the three measures used in this article (ipr, rpr, api). This does not of course imply that there are no other differences between fast and slow speakers.

3.1. Disfluent and short ips

Because some of the ips contain disfluencies which would affect the articulation rate measures, these are analyzed first, before continuing to evaluate the intended and realized phone rates as well as the articulatory precision index for the fluently spoken ips.

In the 683 ips, a total of 76 disfluencies were observed. They are divided into stutters, slips of the tongue, interruptions, hesitations, lengthenings, filled pauses and laughters. Only 20% of the disfluencies occur in the read stories, the rest occur in the retold stories. This despite the fact that the read stories consist of more ips (382) than the retold stories (301), and have an average total duration (66 seconds) which is about 1.5 times that of the retold ones (fast speakers: 43 seconds). The higher frequency of disfluencies in the retold stories confirms our expectation. Although statistical analysis of the data is not possible given the small number of observations in most of the cells, table I suggests that subjectively slow speech is characterized by relatively (i.e. after division by the total number of disfluencies for each column) more stutters and hesitations or filled pauses, while subjectively fast speech contains more slips of the tongue, lengthenings and laughters. To prevent these disfluencies from biasing our measures, only fluent ips are used for further analysis.

Table I. *Disfluencies in ips of speakers subjectively perceived as fast and slow for read and retold stories*

	read		retold	
disfluency	fast	slow	fast	slow
stutter	0	0	0	2
slip of the tongue	3	2	8	0
interruption	4	0	1	0
hesitation	1	1	5	4
lengthening	2	0	23	4
filled pause	1	0	6	4
laughter	1	0	4	0
Total	**12**	**3**	**47**	**14**

Short ips with a duration of less than half a second are also excluded (2.2% of the read and 9.6% of the retold ips), because the small number of phones in these ips can lead to unreliable phone rate estimates, particularly because final lengthening has a strong effect on the average number of phones per second when the ips is short.

3.2. Speech rate

t-Tests for the fluent ips with a duration of at least half a second show that fast and slow speakers have ips of approximately the same duration when reading aloud, but when they retell the story the durations of the ips for fast speakers (1.7 seconds on average) are significantly shorter (t=2.4, df=217, p<0.05) than for slow speakers (average: 2.1 seconds). In both the reading and retelling tasks, the intended (reading: t=10.3, df=356, p<0.001; retelling: t=4.0, df=217, p<0.001) and realized phone rates (reading: t=8.3, df=230.6, p<0.001; retelling: t=5.9; df=217; p<0.001) differ significantly for fast and slow speakers. As expected, the fast speakers have higher phone rates than slower ones. The articulatory precision index, which is highly significant in the reading task (t=4.3, df=356, p<0.001; with lower articulatory precision, i.e. more deletions, for fast than for slow speakers) is not significant in the retelling task. In general, these results confirm that intended and realized, i.e. measured, articulation rates go hand in hand with the (inter-)subjectively perceived speech rate. The results are visualized in Figure 1.

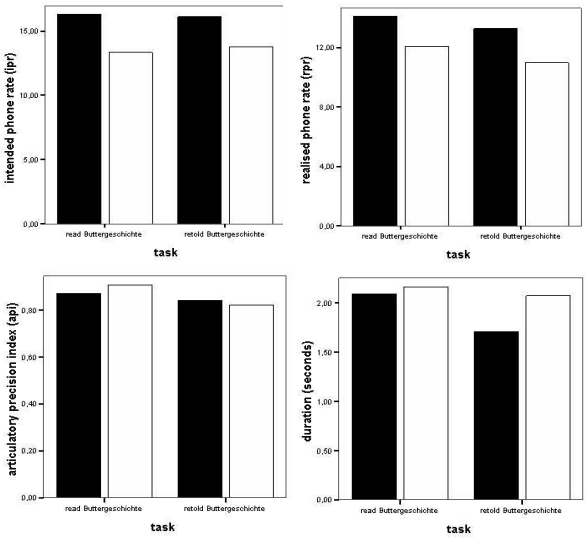

Fig. 1. *Average intended and realized phone rate, articulatory precision index and duration in ips produced by fast (black bars) and slow (white bars) speakers while reading and retelling a German text*

4. Automatic speaker classification

If the subjectively perceived speech rate of the speakers is determined entirely by the phone rates, automatic classification of the speakers on the basis of only these measurements should confirm this.

Cluster centre analyses were performed on the measures reflecting the subjects' articulation behavior to obtain an objective grouping of the subjects. The analyses were only carried out for the retold stories, because for these the speakers' speech behaviour is expected to be more similar to their natural, spontaneous speech behaviour (on which the experimenters' judgment of speech rate was based) than the same measures for the read stories. The cluster centre analyses are only carried out for ipr and rpr, since the difference in api between fast and slow speakers was not significant for the retold stories. Two groups (fast, slow) were created on the basis of the median values of the measured intended phone rates in the ips of the retold story; the same was done for the realized phone rates (fast, slow). The cluster belongingness of the speakers to each two classes for the two variables are combined to derive subject categories which reflect different articulation rates and speaking styles. The speaker groupings are shown in Table II. Speaker CT, who has a slow subjectively perceived speech rate, belongs to category 1 (fast intended and realized articulation rates), while speaker ST, with a fast subjectively perceived speech rate, belongs to category 4 (slow intended and realized articulation rates).

Table II. Grouping of the speakers into four categories on the basis of their measured intended (ipr) and realized rates (rpr); slow speakers are italicized

	cat. 1	cat. 2	cat. 3	cat. 4
ipr	fast	fast	slow	slow
rpr	fast	slow	fast	slow
speaker	AM	*BB*	EK	*CS*
	AS			*MM*
	CK			ST
	CM			
	CT			
	EM			
	HR			

This shows that although fast and slow subjects were found to differ in their measured speech rates (section 3), the subjective impression of the speakers as either slow or fast is not only based on articulation rate, but is clearly also influenced by other factors.

5. Analysis of a large database

The number of subjects used in the production experiment is only small. This raises the question whether the findings can be generalized to a larger database. With this aim, the intended and realized phone rates as well as the articulatory precision index were measured for the intonation phrases (for this purpose comparable to inter-pause stretches used in the production study) in the German Kiel Corpus of Spontaneous Speech. At the time of analysis a total of 1329 files from the Kiel Corpus of Spontaneous Speech had been prosodically labeled. The prosodic labels were used to divide the speech files up into intonation phrases (only one phrasing level was used), resulting in 5779 intonation phrases after exclusion of intonation phrases with only non-speech material or hesitations. For intonation phrases with extreme perceived speech rates, the perceived rate was indicated by the labelers. Note that these labels do not indicate a judgment of the speaker as fast or slow, but only of the intonation phrases. The labels RM for rate minus and RP for rate plus were used when (part of) an intonation phrase was considered to be spoken particularly slowly or fast, respectively. The rest of the intonation phrases were labeled by us as rate normal (RN) – though this does not preclude variation in articulation rates. As in the analysis of the subjects' speaking habits, only intonation phrases with a duration of at least half a second are analyzed (n=4736). This reduces the data by 18 per cent, but only 8 out of 137 fast and 2 out of 116 slow intonation phrases are less than half a second long. The fact that the labelers only rarely use the labels RM and RP for short intonation phrases supports Pfitzinger's observation that "using speech signal segments of less than 500 ms hindered the assessment of speech rate" [4].

For each intonation phrase, the number of intended and realized phones per second was determined. These can be derived from the labelling of the corpus, in which changes to the canonical form are indicated for the realisation of the utterances. The articulatory precision index was also derived. An analysis of variance with subsequent Tukey-HSD post-hoc tests show that both the intended (F(2,4733)=298.0, p<0.001) and realized phone rates (F(2,4733)=253.8, p<0.001) as well as the articulatory precision index values (F(2,4733)=45.1, p<0.001) differ significantly for the three perceived rates. Fig. 2 shows that, as expected, both the intended and the realized speech rate increase with perceived speech rate, while the articulatory precision index decreases.

There is a strong and highly significant correlation between the intended and realized articulation rates (overall correlation: r=0.93, n=4736, p<0.001). Fig. 3 shows scatterplots of the data for slow (RM: r=0.95, n=114, p<0.001), normal (RN: r=0.92, n=4493; p<0.001) and fast (RP: r=0.85, n=129, p<0.001) perceived speech rates. There is considerable overlap in the measured phone rates of the different perceived rate categories. This confirms that other factors than articulation rate also determine the perceived speech rate.

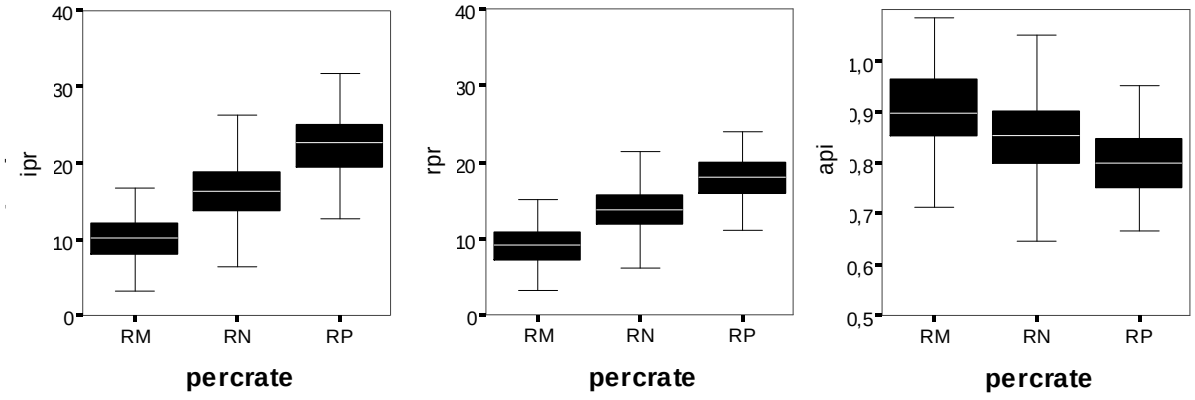

Figure 2. *Boxplots of intended (ipr) and realized phone rates (rpr) in phones per second as well as articulatory precision index (api) values for intonation phrases perceived as slow (RM), normal (RN) and fast (RP)*

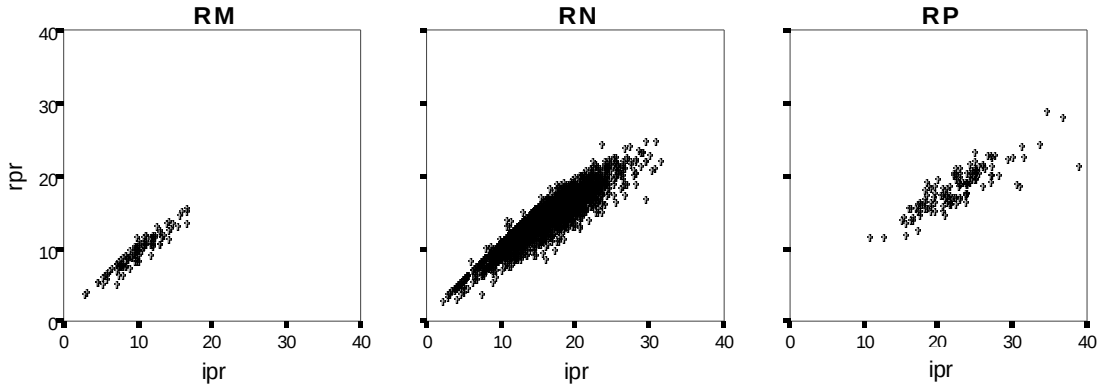

Figure 3. Scatterplot of intended (ipr) versus realized rate (rpr) in phones per second for all intonation phrases with a minimum duration of 0.5 seconds perceived as slow (RM), normal (RN) and fast (RP)

6. Discussion

The perceived fast and slow speakers in section 3, and perceived fast and slow ips in section 5, differ significantly in their intended and realized articulation rates as well as in their articulatory precision index values. The results from automatic classification of the speakers as fast or slow on the basis of their measured articulation rates in section 4 shows a different picture. One of the speakers who was judged as a fast speaker by the experimenters was automatically classified as slow and another speaker who was perceived as slow was categorized as fast on the basis of the measured articulation rates. This clearly shows that the intended and realized phone rates per se cannot explain the perceived rates of the speakers. This is confirmed for individual ips in the database analysis in section 5.

It is possible that pausing or disfluencies also determined the perceived rate. Pausing was excluded from our experiments, since articulation rate as a measure does not include pauses, and also because our measures were computed for ips (and, for the Kiel Corpus, for intonation phrases, which usually also do not contain pauses). The same is true for disfluencies, which we explicitly excluded from our data. Both have been found to influence perceived rate.

It is also important to note that the observed articulation rates may "merely" be an automatic concomitant of other, prosodic properties of the speech. Therefore, we cannot assume a direct causal relationship between articulation rate and perceived rate, since the intended and realized rate may depend on other phonetic or phonological properties of the ips produced by the speakers.

It is not very probable that the differences between speakers in section 4 can be explained by differences in the proportion of content versus function words, since these are determined by the message (linguistic content of the utterances) and therefore expected to show similar variation across speakers. But it is certainly possible that some speakers for instance use more accents than others, which may cause differences in articulation rate between the speakers [2]. Other intonational differences between the speakers may have similar effects. To find out whether the perceived rate of the speakers could be explained completely by intonational properties, further investigations are needed.

In general, the considerable overlap in the measured phone rates of the different perceived rate categories in the database analysis confirms that other factors than articulation rate also determine the perceived speech rate. The analysis of a small subset of intonation phrases differing in perceived rate as well as articulation rate did not, however, lead to any consistent differences [3]. But this does not exclude the possibility that articulation rate is a concomitant of *several different* properties. That is, although no single property of the utterances could be found as the responsible factor for the observed differences, it is certainly possible that different properties influenced the articulation rate in different intonation phrases. In this view, the intended and realized phone rates are the surface result of a complex interaction between many phonological/linguistic factors in speech production. A further analysis of the retold texts in terms of intonational, pausing and other characteristics may help to clear up the possible relationship with articulation rate.

Despite interpretational problems of intended and realized phone rates with regard to their role as an independent factor in speech rate perception, it is clear that these measures cannot be ignored. For applications which make use of speech synthesis, it is probably not sufficient to control only intonational or other prosodic properties of the synthesized utterances, but the intended and realized phone rates (and probably other reductions which do not entail the deletion of a segment) must also be modelled appropriately to achieve a natural speech quality.

7. References

[1] Batliner, A.; Kießling, A.; Kompe, R.; Niemann, H.; Nöth, E., 1997. Tempo and its change in spontaneous speech. *Proc. Eurospeech,* Rhodes, vol. 2, 763-766.

[2] Crystal, T.H.; House, A.S., 1990. Articulation rate and the duration of syllables and stress groups in connected speech. *JASA* **88**(1), 101-112

[3] Koreman, J., 2006. Perceived speech rate: the effects of articulation rate and speaking style in spontaneous speech, *JASA* (to appear).

[4] Pfitzinger, H.R. (1999). "Local speech rate perception in German speech," *Proc. 14th ICPhS*, San Francisco, Vol. 2, 893-896.