The Perception of Articulation Rate

Jacques Koreman

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

ABSTRACT

Segmental as well as prosodic utterance characteristics determine listeners' perception of speaking rate. The segmental characteristics are usually captured by the articulation rate, which is sometimes based on the canonical, underlying representation and at other times on the actual, surface realisation of phones in an utterance. The relation between the two is defined by the phone deletion rate. Both the underlying and the surface structure are shown to play a role in the perception of speaking rate and must (in addition to other related rate phenomena) be taken into account when modelling speaking rate.

1. INTRODUCTION

Differences in speaking rate can be reflected by a variety of phonetic and phonological properties. The number and duration of pauses [1] as well as the number and strength of prosodic phrase boundaries, the use of F0 resets to signal them, the complexity of pitch accents (bitonal versus monotonal) as well as F0 range can vary with speaking rate [2-4]. Some of these characteristics have been shown to play a role in the *perception* of speaking rate [5]. Between pauses, articulation rate also varies, though to a lesser extent than overall speaking rate – and presumably contributes to the general impression of speaking rate. Intonation phrases (IP's) have been identified as the domain of speaking rate variation [6] and were selected to evaluate speaking rate in the study reported here.

Lindblom [7] describes speaking rate as a result of the interplay between output-oriented control, i.e. the need for a speaker to (be able to assume he will) reach his communicative goal, and system-oriented control, i.e. his tendency to "default to some low-cost form of behavior". If output-oriented constraints are strong, the speaker can either speak slowly or move his articulators from one target position to the next at greater than normal speed, resulting in a clear speaking style (hyperspeech), even at high speaking rates. Otherwise, system-constraints, e.g. due to inertia of the articulators, will lead him to economise on his articulatory effort and adopt a relaxed, conversational speaking style (hypospeech).

The difference between the two articulatory strategies implies that the *perception* of speaking rate may be a product of actual articulatory events *and* knowledge of what articulations are implied by a particular utterance. This is not Motor Theory in the strict sense [8], but does appeal to the listener's knowledge of the required motor patterns. The two articulatory strategies are indirectly reflected in the realised phone rates and in the corresponding rates for the intended or canonical phones (which we shall call intended rates). In the German Kiel Corpus of Spontaneous Speech [9] the intended and realised articulation rates of IP's show a strong overall correlation ($r = 0.928$, $p < 0.001$), cf. [10]. With increasing intended phone rates, more phones are deleted from the canonical pronunciation (Pearson's correlation between intended phone rate and the ratio of realised/intended phone rate: $r = –0.515$, $p < 0.001$), so that both the measured phone rate and the phone deletion rate are possible cues for perceived speaking rate. To test this, we used stimuli with carefully selected intended and realised phone rates.

2. METHOD

Stimuli were selected from the segmentally and prosodically manually labelled German Kiel Corpus of spontaneous speech. The corpus consists of high-quality recordings of conversations in which two speakers schedule one or more appointments. Despite the recording set-up, in which the speakers had to press a button to obtain the floor, the speech is very natural.

Intonation phrases (IP's) were selected on the basis of their intended and realised articulation rates computed from the canonical phone representation of each phrase and the actually realised phones, respectively. Incomplete reductions (i.e. quantitative and qualitative reductions which do not entail the deletion of a complete phone segment) do not affect the realised phone rate and were therefore not taken into consideration. It is expected that phrases with more numerous deletions also show a greater number of incomplete reductions (and vice versa), since these can be considered as a less extreme but otherwise similar effect of a sloppy articulation, but their effect on perception is not investigated here. Since the phrases were to be judged in pairs, we only selected phrases of similar duration (1–1.5 seconds) in order to control for a possible effects of phrase duration [11]. The selected phrases varied systematically in their intended and realised articulation rates, allowing us to evaluate the effect of articulation rate per se and differentiate it from that of speaking style (clear versus sloppy) on listeners' perception. Six rate categories were selected (cf. Fig. 1):

FC *Fast, clear* IP's had high intended and realised phone rates. Intended phone rates were between 1 and 2 standard deviations above the mean. The realised phone rates were close to the intended

ISBN 1-876346-48-5

phone rates, with a low phone deletion rate (maximally 8%). The realised phone rates of these phrases were about 2-3 standard deviations above the mean.

FS The intended phone rates for the *fast but sloppy* phrases were similar to those of the clear phrases, but 35–40% of the intended phones were not realised. The realised phone rate was therefore much lower than in the fast, clear category, namely within one standard deviation of the mean (all above the mean, except for two phrases).

NCf A third category of clearly spoken phrases was selected, with intended and realised phone rates similar to the realised phone rates of the fast, sloppy speech. The percentage of phone deletions was comparable to that in the FC category. The intended phone rates were within about 0.5 standard deviation from the mean. This category is named *normal, clear phrases* for comparison with fast intended phrases (NCf).

As we noted above, output-oriented constraints can lead a speaker to slow his speech down or increase articulator speed in order to prevent undershoot [12]. Since at fast intended articulation rates the inertia of the articulators (which itself is constant) creates a strong force to slow down or reduce phones, the lack of phone deletions in category FC is taken as a sign of strong output-oriented constraints and therefore a clear case of hyperspeech. Conversely, the deletion of phones at slower speaking rates is a sign of extremely lax output constraints, allowing system-oriented constraints to induce hypospeech (category NS below). In order to evaluate whether the different relative weight of output and system constraints is taken into account by the listener in his judgment of speaking rate, another set of stimuli was selected from the slower half of the intended articulation rate range (normal to slow):

NCs *Normal, clearly* spoken phrases were selected with similar intended and realised phone rates to those of the NCf category (but with *all* intended phones being realised), but the phrases were different ones. This was done in order to provide an optimal match with the phrases in the two categories below.

NS *Normal but sloppy* intonation phrases showed phone deletion rates between 35–40% (as in the FS category). This resulted in fairly slow realised phone rates between –0.5 and –1.5 standard deviations from the mean.

SC Finally, *slow, but clear* phrases were selected. The phrases were matched to the *NS* phrases in their realised articulation rates. The intended phone rates are between –1 and –2 standard deviations from the mean.

All possible combinations of the 6 categories were compared, giving 15 comparisons. Five sets of 6 phrases

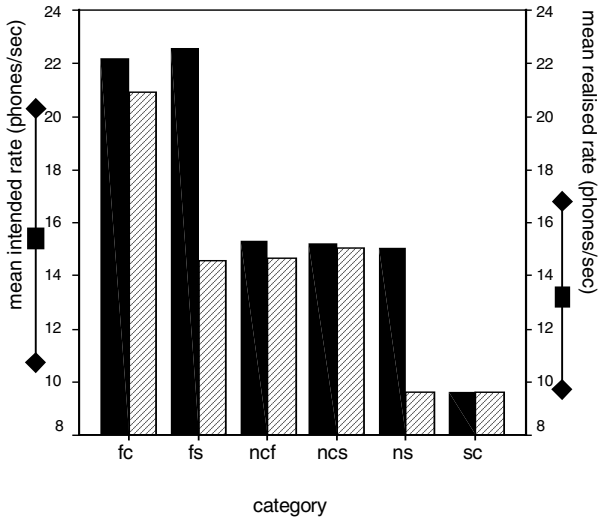

Figure 1: Average intended (filled bars) and realised phone rates (dashed bars) in phones/sec of stimuli from six rate categories (see text) – with mean and standard deviations for intended and realised rates indicated on the vertical axes

were selected from the database. The total number of stimulus pairs was therefore 5 (sets) x 15 (comparisons) = 75 per listener. The word content of the phrases was always different and with a few exceptions they were all produced by different speakers. The stimulus pairs were offered in pseudo-randomised order. Ten female and ten male listeners with no hearing or language deficiencies aged between 20 and 59 (average: 30) participated in the listening test. The listeners were divided into two equal groups, which heard the stimuli within each pair in opposite orders. The stimuli were preceded and followed by filler items. In addition, they were mixed with a small number of similar phrases not used for this study.

The "Experimenter" software [13] was used to carry out the listening experiment. The stimuli, which had been set to equal loudness levels[1], were played to the listeners over headphones in a sound-treated room at a self-selected comfortable loudness level. The stimuli were displayed orthographically on a computer screen for 3.5 seconds, followed by a beep and a silent pause of 0.5 seconds. Then the stimulus pair was played, separated by a silence interval of 0.5 seconds. The subject then had 5 seconds to respond by pressing one of three keys on the keyboard for "first phrase faster", "second phrase faster" or "both phrases are equally fast". Response times were also registered, but not used in the present analysis.

[1] It may be argued that setting the loudness to an equal level for all stimuli destroys the possible relationship between rate and loudness. Since the speakers were (nearly) all different and therefore have their own preferred loudness this was considered to be of minor importance. We cannot, however, exclude the possibility that the factors studied in this experiment interact with loudness.

ISBN 1-876346-48-5

3. RESULTS AND DISCUSSION

In order to evaluate whether rate perception is affected by the measured rates, the 15 stimulus pair categories are divided into four sets according to their intended and realised phone rates (cf. Fig. 1). In set A, both the intended and the realised phone rates of the two phrases in a stimulus pair differ. In set B, the two phrases differ only in their realised rate, while in set C there is only a difference in intended rate. In set D, finally, both the intended and realised phone rates are the same (see Table 1).

cond.	int. rate	real. rate	#judge-ments	compared rate categories
A	diff	diff	800	FC-NCf, FC-NCs, FC-NS, FC-SC, FS-NS, FS-SC, NCf-SC, NCs-SC
B	same	diff	300	FC-FS, NCf-NS, NCs-NS
C	diff	same	300	FS-NCf, FS-NCs, NS-SC
D	same	same	100	NCf-NCs

Table 1: Intended and realised rates for 4 conditions (see text), with the number of listener judgements and the compared rate conditions (cf. Fig. 1)

The perceived rate distinctions are shown in Fig. 2 for each of the four sets. The responses were recoded so that the first stimulus of each pair always represents a faster intended and/or realised phone rate, although the actual order was always cross-balanced in the experiment. The response "first stimulus faster" (black columns) should therefore be expected if one of the measured phone rates determines rate perception, except in Fig. 2D, where there is no rate difference between the stimuli. Fig. 2A shows that listeners clearly perceive the expected rate difference if both the intended and realised rates of the two IP's differ (82% "stim1 faster" responses), with relatively few "equal" responses (16%). More "equal" and fewer "stim1 faster" responses are given when only one of the two measured phone rates differs for the two IP's (Fig. 2B and 2C), with somewhat more "equal" and fewer "stim1 faster" responses when there is a difference in realised rather than intended rate. The lower realised rate of the (recoded) second stimulus in each of the pairs represented in Fig. 2B causes it to be perceived as slower than the first stimulus, despite equal intended rates. This shows that there is an effect of surface rate. In Fig. 2C, the first stimulus has a higher intended phone rate than the second stimulus and is perceived as faster, despite equal realised rates. This therefore shows that there is also an underlying rate effect. Articulator speed, reflected in the surface or realised rate, and knowledge of the underlying or canonical form, reflected in the intended rate, therefore play a role in the perception of speaking rate.

The relative differences between Figs. 2B and 2C (more "stim1 faster" and fewer "equal" responses in Fig. 2C) should not be taken as an indication that the intended

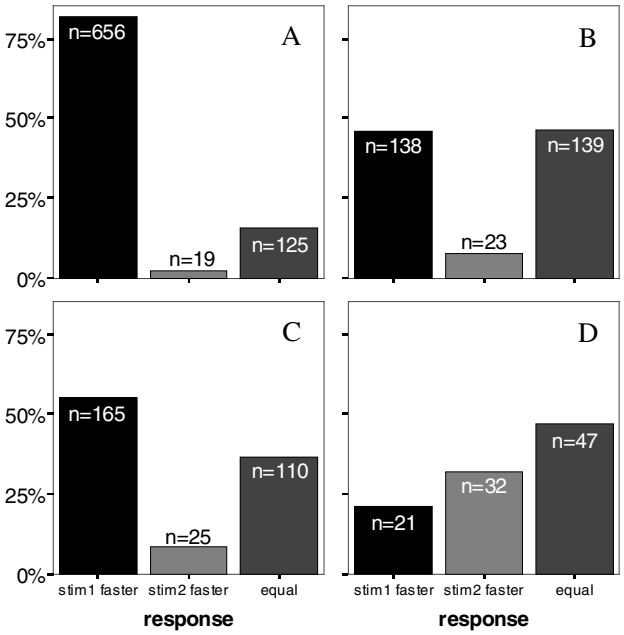

Figure 2: Listener responses (percentages) to stimuli with different intended and realised rates (A), only different realised rates (B) or intended rates (C), or with both equal (D).

rate is of greater importance for rate perception than the realised rate, since the differences in rate values in terms of their variation (i.e. in terms of z-scores) are not identical in this study. As expected, when both the intended and the realised rates of the two phrases in the stimulus pair are the same (Fig. 2D), the highest percentage of "equal" reponses are given[2] (comparable to when there is only a difference in realised rate), and there is no clear tendency for one of the stimuli to be perceived as faster.

The clear perceptual effect of differences in intended and realised rates is consistently reproduced when we compare each two rate categories in Fig. 3 (next page). There are no qualitative differences in rate perception in the upper half of the intended rate range (dashed graphs) compared to the lower half of the range (dotted graphs). This seems to indicate that the stronger system-oriented constraints in fast hyperspeech do not lead the listener to judge speaking rate differently at high articulation rates from clear speech at normal and slow intended rates. The deletion of phones at *normal* speaking rate, too, has no qualitatively different effect on the perception of speaking rate from phone deletion at *fast* speaking rates, although sloppiness at normal intended rates can be considered as a clear indication of extremely lax system constraints.

[2] The reason why there are many "stim1 faster" and "stim2 faster" responses instead of only "equal" responses is probably that the listeners are induced by the task to choose one of the two stimulus as faster.

ISBN 1-876346-48-5

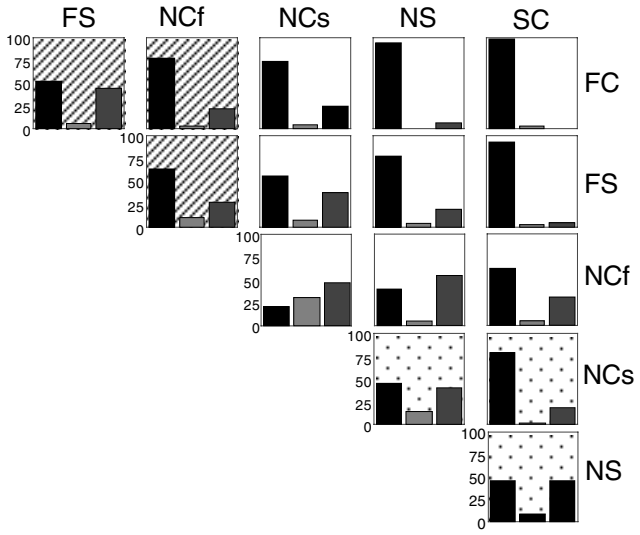

Figure 3: Percentages "stim1 faster" (black bars), "stim2 faster" (light gray bar) and "equal" (dark gray bars) averaged for each stimulus pair. Rows indicate the category of the first stimulus in a pair, while columns show the category of the second stimulus.

4. GENERAL DISCUSSION

Comparing prosodic structures differing in their surface complexity but with equal assumed underlying structures, Rietveld and Gussenhoven [5] found that resynthesised utterances with linked intonation contours are perceived as faster than utterances with physically equal speaking rate but unlinked contours. They show this bias cannot be explained by the phonetic (number of rising or falling movements) or phonological complexity (number of tone segments) of the utterance. They conclude that the linked contour, implying absence of a tone domain boundary, causes a bias in the listener towards perceiving fast speech, beacuse linked intonation contours typically occur at faster speaking rates.

An interaction between underlying and surface realisation was also found on the segmental level in our experiment. Segment deletions as an extreme form of reduction were shown to affect the perception of speaking rate. A lower rate of surface events (realised phones) causes sloppy speech to be perceived as slower than clear speech when the intended rates of the compared IP's are identical. On the other hand, the perceived speaking rate of sloppy IP's is not reduced so much that it becomes equal to that of clear IP's with the same realised rate. I.e. the perception of speaking rate in IP's is the product of the interaction between surface and underlying rate.

Although the above effects are important in modelling the perceptual effects of natural speaking rate, the desirability of modelling natural speaking rate effects in speech technological applications using flexible-rate synthesis remains an issue for further investigation [14].

REFERENCES

[1] F. Goldman-Eisler, *Psycholinguistics*, London & New York: Academic Press, 1968.

[2] C. Fougeron and S.-A. Jun, "Rate effect on French intonation: prosodic organization and phonetic realization," *Journal of Phonetics*, vol. 26, pp. 45–69, 1998.

[3] K.J. Kohler, "Parameters of speech rate perception in German words and sentences: duration, F0 movement and F0 level," *Language & Speech*, vol. 29, pp. 115–139, 1986.

[4] J. Trouvain and M. Grice, "The effect of tempo on prosodic structure," *Proc. 14th Int'l Conf. of the Phonetic Sciences (ICPhS)*, San Francisco, pp. 1067-1070, 1999.

[5] T. Rietveld and C. Gussenhoven, "Perceived speech rate and intonation," *Journal of Phonetics*, vol. 15, pp. 273–285, 1987.

[6] J. Dankovičová, "Articulation rate variation within the intoantion phrase in Czech and English," *Proc. 14th Int'l Conf. of the Phonetic Sciences (ICPhS)*, San Francisco, pp. 269-272, 1999.

[7] B. Lindblom, "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal, Eds., pp. 403–439. Dordrecht/Boston/ London: Kluwer Academic Publishers, 1990

[8] A.M. Liberman and I.G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, pp. 1–36, 1985.

[9] IPDS, The Kiel Corpus of Spontaneous Speech, vols. 1–3 (CD-ROM #2–4). Kiel: Insitut für Phonetik und digitale Sprachverarbeitung, 1995–1997.

[10] J. Trouvain, J. Koreman, A. Erriquez and B. Braun, "Articulation rate measures and their relation to phone classification in spontaneous and read German," in: *Proc. of the Isca ITR-Workshop on Adaptation Methods for Speech Recognition,* Sophia-Antipolis, pp. 155-158, 2001.

[11] H.R. Pfitzinger, "Local speech rate perception in German speech," *Proc. 14th Int'l Conf. of the Phonetic Sciences (ICPhS)*, San Francisco, vol. 2, pp. 893–896, 1999

[12] B. Lindblom, "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, col. 35, nr. 11, pp. 1773–1781.

[13] G. Altman, *Experimenter. A Toolkit for Multi-Modal Psycholinguiistic Experimentation on the Apple Macintosh*, Sussex: Laboratory of Eperimental Psychology, University of Sussex, 1992.

[14] E. Janse, "Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech," submitted to *Speech Communication*.