Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech

Jacques Koreman

Citation: The Journal of the Acoustical Society of America **119**, 582 (2006); doi: 10.1121/1.2133436

View online: http://dx.doi.org/10.1121/1.2133436

View Table of Contents: http://asa.scitation.org/toc/jas/119/1

Published by the Acoustical Society of America

Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech

Jacques Koreman
Institute of Phonetics, Saarland University, P.O. Box 151150, D-66041 Saarbrücken, Germany

(Received 14 June 2004; revised 11 October 2005; accepted 11 October 2005)

In this study, the effect of articulation rate and speaking style on the perceived speech rate is investigated. The articulation rate is measured both in terms of the *intended* phones, i.e., phones present in the assumed canonical form, and as the number of actual, *realized* phones per second. The combination of these measures reflects the deletion of phones, which is related to speaking style. The effect of the two rate measures on the perceived speech rate is compared in two listening experiments on the basis of a set of intonation phrases with carefully balanced intended and realized phone rates, selected from a German database of spontaneous speech. Because the balance between input-oriented (effort) and output-oriented (communicative) constraints may be different at fast versus slow speech rates, the effect of articulation rate is compared both for fast and for slow phrases from the database. The effect of the listeners' own speaking habits is also investigated to evaluate if listeners' perception is based on a projection of their own behavior as a speaker. It is shown that listener judgments reflect both the intended and realized phone rates, and that their judgments are independent of the constraint balance and their own speaking habits. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2133436]

PACS number(s): 43.71.Es, 43.71.Gv, 43.71.Bp, 43.72.Kb [DOS] Pages: 582–596

I. INTRODUCTION

The segmental and prosodic characteristics of speech rate have been investigated quite extensively. The prosodic characteristics mainly concern pausing behavior and phonetic and phonological intonational properties. Fast speech characteristically has fewer and shorter pauses (Goldman-Eisler, 1968; see Butcher, 1981, and Lass, 1970, for the effect of pauses on the perceived speech rate) and fewer F0 resets (Trouvain and Grice, 1999). These observations support the idea that speech is divided into fewer prosodic units at fast rates than at normal or slow rates. Further, the complexity of pitch accents varies between speech rates, with fewer bitonal pitch accents and more monotonal ones in fast than in slow speech (cf. Rietveld and Gussenhoven, 1987, for its effect on the perceived speech rate). Also, the F0 range is generally reduced in fast speech, although there is substantial variation across speakers (Fougeron and Jun, 1998; Kohler, 1986).

Segmental effects of speech rate are reflected in the articulation rate, which by definition excludes pauses (Crystal and House, 1990) and is defined as the number of production units (often phones or phonemes, but also syllables, morae, or words) per unit time. In the present article, the relationship between the perceived speech rate and two different measures of segmental articulation rate—"intended" and "realized"—is under investigation. Due to the importance of the definition of these terms for the interpretation of the results, the two measures need to be considered in some detail prior to the presentation of our experiments. This discussion cannot offer an ultimate solution to all terminological issues, but it aims to identify the underlying theoretical problems and to clarify the operationalization of the two rate measures used here. These are derived from the hypothesized "canonical" and the actual, realized forms of the words in an utterance. The resulting calculable speech rate measures will be termed *intended* and *realized* rate, respectively (cf. "linguistic" and "phonetic" segments in Den Os, 1985).

The experimental issue under consideration is whether the perceived speech rate is determined by the number of actually realized speech sounds per unit time or whether it is dependent on the number of potentially realizable units according to some underlying, ideal(ized) form of the words, presumably defined in the speaker-listener's mental lexicon. For the latter concept the terms "canonical" and "intended" are used. There can be objections to the word "canonical" to refer to a more abstract representation. *Canonical* itself means "regular," "accepted or approved" (Funk and Wagnalls, 1973). In the present context the term could be interpreted as the accepted norm for the realization of a word in its everyday use, but this is not necessarily an indication of its hypothesized, underlying structure. The canonical form is often also referred to as the lexicon form, but it is not clear whether it can also be assumed to be the segmental representation in the speaker-listeners' *mental* lexicon, nor whether this representation is identical for each individual speaker (cf. Bromberger and Halle, 2000). To elicit a word in its canonical form, speakers may be requested to produce the word in citation form, i.e., the segment sequence which is produced when the word is articulated clearly and in isolation. But at the same time it is not certain that the citation form is really a good indicator of the canonical form: it may be overarticulated in the sense of a "spelling pronunciation." For German, the language in which the experiments presented here were carried out, the status of schwa is particularly interesting in this respect and has been discussed extensively by Kohler and his colleagues (Wesener, 1999; Kohler, 2001; Kohler and Rodgers, 2001). The present-day realization of schwa, for instance, which is historically present in the *-en* suffix of many verbal and adjectival forms, depends on the dialect region of the speaker. In most dialect regions it is realized as a syllabic nasal (this is the norm in Standard German—cf. Duden, 2000, p. 38 ff.), but in some dialects it is realized as schwa, in others still as schwa plus nasal. To decide whether schwa is part of the (non speaker-specific) canonical form, Kohler and Rodgers (2001) statistically analyze the (Northern) German Kiel Corpus (IPDS, 1994; IPDS, 1995-1997), setting a threshold on the relative frequency of its realization in different contexts. In contexts in which schwa only occurs rarely in the actual, realized form, as in poststress "plosive+schwa+apical nasal" syllables, Kohler concludes that "schwa-less and at the same time place-assimilated forms have become the canonical entries for the speaker group as a whole" (Kohler, 2001, p. 10; see also Wesener, 1999, p. 332). Kohler and his co-workers clearly commit themselves, therefore, to a definition of the canonical form in terms of its observable everyday, concrete use. Of course, a statistical measure such as used by Kohler is a useful descriptive tool, but evidence for the psychological status of schwa needs to be derived (if this is possible at all) from psycholinguistic experiments. This problem can be generalized to the status of other phones as part of the canonical form. The reliance on a statistical measure based on the direct phonetic context further ignores the role of other factors like the situational and dialog context on the realization of utterances, which would probably lead to different canonical forms for many words. In the present paper, the *intended* form is defined as the abstract, "full" canonical form that may be hypothesized to be stored in language users' mental lexicon, including schwa and other rarely realized segments. Since these segmentally more elaborate forms do occur in our data, their realizations represent *de facto* the more careful end of the articulatory continuum from "clear" to "sloppy."[1] Whether they belong to others' idea of "canonical," "underlying," or "intended" forms is not critical for the present study.

The realized form can differ substantially from the underlying form. This is demonstrated by the possible realizations of the phrase "I do not know"—from very clear realizations of the intended form to more sloppy realizations typical of a conversational speaking style, including "dunno" or even more strongly reduced, mainly vocalic, realizations (Hawkins, 2003). As Hawkins convincingly demonstrates, the realization depends on the contextual setting in which a conversation takes place. In his H&H ("hyper"- and "hypo"-articulation) theory, Lindblom (1990) describes the interaction between articulation rate and reduction as the result of system- and output-oriented constraints on speech production. The speaker is presented as striving for minimal effort leading to sloppy or hypospeech (cf. Lindblom, 1963; Gay, 1981), but also willing to resort to clear or hyperspeech if it is required by the situation to get his message across. The example from Hawkins shows that the underlying, canonical form cannot be taken to be "intended" in a literal sense, i.e., it is not claimed here that it is always the speaker's intention to realize this form. In fact, system-oriented control as discussed in H&H theory implies that the speaker does *not* always intend to invest all the effort necessary to produce an unreduced form of the word.

Looking at the example from a speech perception standpoint, Hawkins claims that "speech perception does not demand early reference to abstract, linguistic units" (p. 373). This, too, calls the status of the so-called "intended form" into question and takes exemplars as the basis for speech perception. Although the terminology in the main part of this article is taken from the abstract-form view ("intended," "canonical"), it should be stressed that the experimental results are compatible with both views. The controversy between the abstract-form versus exemplar-based models is, however, relevant to the interpretation of the results and will therefore be taken up again in Sec. V.

Clearly, articulation rate cannot be dealt with without also considering speaking style (which will be used here in the restricted sense of articulatory precision or clarity). It should be pointed out that in our operationalization of the realized rate only the number of actual, realized segments per second is counted. Incomplete reductions, i.e., quantitative (e.g., vowel shortening) and qualitative (e.g., the substitution of long, tense by short, lax vowels) reductions which do not entail the deletion of a complete phone segment, do not affect the realized phone rate and are therefore not taken into consideration here. It is expected that realizations with more numerous deletions also show a greater number of incomplete reductions, since these can be considered as a less extreme but otherwise similar effect of a sloppy articulation—but their effect on the perceived speech rate is not investigated here.

Although it is clear that articulation rate and speaking style are related phenomena, their joint effects on the perceived speech rate have, to our knowledge, not been investigated in controlled experiments using unmodified, natural speech. In this paper, the following hypotheses are tested:

First, because segmental reduction is often related to a higher speech rate, it is hypothesized that utterances with many deletions (as the most extreme form of reduction) are interpreted by listeners as spoken at a higher speech rate than utterances of an equivalent realized articulation rate with no or very few deletions.

Second, the balance of input- and output-oriented constraints is different for fast and slow speech. In contrast to clear speech at lower intended articulation rates, the lack of reductions at fast intended articulation rates may be interpreted as hyperspeech and the greater perceived articulatory effort may be interpreted as an indication of faster speech. Conversely, the presence of reductions at lower intended articulation rates, where input-oriented constraints are weak, may be interpreted as hypospeech or even "slurred" articulation, and may therefore be perceived as slower. Exploring this extension of H&H theory, the perceived speech rates of two sets of carefully selected stimuli with high and low intended articulation rates are compared.

A third hypothesis tested here is that the perceived speech rate depends on the listeners' own speaking habits. Speech which is judged as fast by slow speakers may be

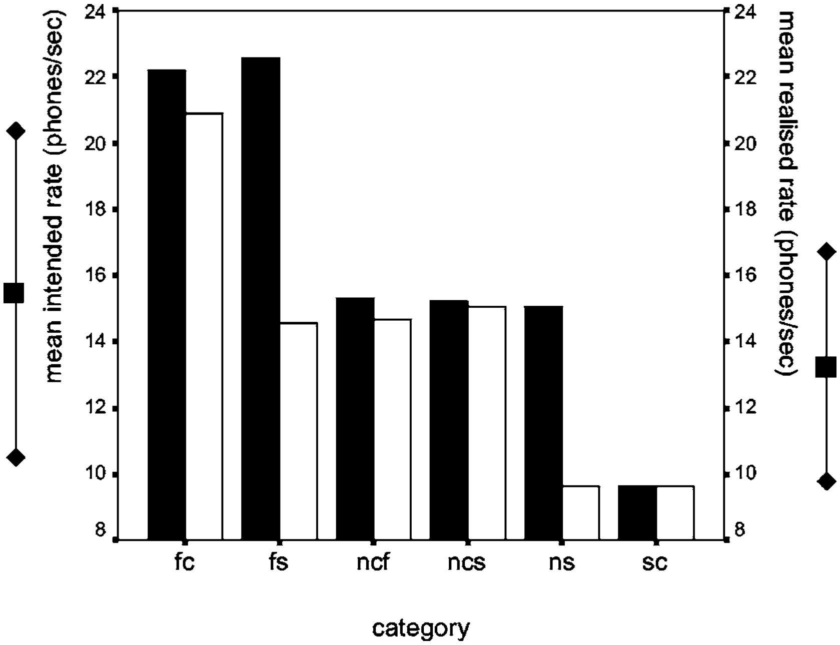

FIG. 1. Average intended (filled bars) and realized phone rates (white bars) in phones/s of stimuli from six rate categories (with means and standard deviations for intended and realized rates in the prosodically labeled part of the Kiel Corpus of Spontaneous Speech indicated on the vertical axes).

judged differently by fast speakers and vice versa. Also, clear and sloppy speakers may differ in their evaluation of phone deletions in speech utterances.

The above hypotheses are investigated in two listening experiments using intonation phrases from the German Kiel Corpus of Spontaneous Speech (IPDS, 1995-1997). The first experiment (*comparison* experiment) evaluates the effect of measured articulation rate and speaking style (clear versus sloppy) on the perceived speech rate in direct comparisons between the two intonation phrases in each stimulus pair. Listeners' ability to judge the speech rate of the individual phrases is investigated in a second experiment (*scaling* experiment). Further, the evaluation of speech rate in this experiment is related to the listeners' own speaking habits. Because it can be hypothesized that other factors such as hesitations, the relative number of function versus content words in an utterance, and the intonation patterns affect the perceived speech rate of the specific stimuli used in the experiments, the generalizability of the results from our controlled experiments is examined. This is done by comparison with the perceived speech rates labeled in the entire database from which the stimuli are selected.

II. METHOD

In two experiments, the influence of the intended and realized articulation rates on the perceived speech rate is investigated. In the first experiment, listeners compare which of two stimuli is faster (or whether they are equally fast). In a scaling experiment carried out immediately after the first experiment, each of the stimuli is then judged separately on a continuous scale for perceived speech rate. Before presenting the two experiments, the stimuli are described in some detail. The participants in both tasks were selected on the basis of their own speech rate as subjectively perceived by the experimenters. Speakers who were judged as either particularly fast or as particularly slow were invited to participate in the listening experiments.

A. Stimuli

The stimuli consist of intonation phrases (without pauses) selected from the German Kiel Corpus of Spontaneous Speech (IPDS, 1995–1997). The corpus contains high-quality recordings of conversations in which two speakers schedule one or more appointments. Despite the recording setup, in which the speakers have to press a button to obtain the floor, the speaking style is very natural. The reason for choosing intonation phrases is that the articulation rate within an intonation phrase can be assumed to have a relatively stable pattern (Dankovičová, 1999).

The intonation phrases and the segments they contain are manually labeled for a large part of the Kiel Corpus (see Sec. IV below). The segmental labels of the intonation phrases take the canonical form as a point of departure and indicate the changes to the canonical form which lead to the observed, realized form (see the Appendix). The intended and realized articulation rates can therefore be derived from the transcriptions by counting the number of intended and realized phones in an intonation phrase, respectively, and dividing them by its duration. In order to control for the effect of phrase duration (cf. Pfitzinger, 1999), only phrases with a duration between 1 and 1.5 s were selected. They were carefully matched within each set of phrases for which pairwise comparisons were made.

The selected phrases vary systematically in their intended and realized articulation rates, which allows us to separately evaluate the effect of articulation rate (fast versus slow) and speaking style (clear versus sloppy) on the perceived speech rate. Two sets of stimuli were selected (with five phrases for each of the six stimulus categories described below), one set from the fast and the other from the slow half of the Kiel Corpus (cf. Fig. 1).

Set 1: Comparison of fast with normal speech

FC: *Fast, clear* phrases have both high intended and high realized phone rates. Intended phone rates are between 1 and 2 standard deviations above the sample mean computed across the prosodically labeled part of the database from which our stimuli were selected. The realized phone rates are close to the intended phone rates, with few phone deletions (maximally 8%). The realized phone rates of these phrases are about 2–3 standard deviations above the sample mean. This category is considered as typical hyperspeech, because the phrases are spoken fast both from the point of view of their intended rate as well as the actual articulator speed, as reflected in the realized phone rate.

FS: The intended phone rates for the *fast but sloppy* phrases are similar to those of the clear phrases, but 35%–40% of the intended phones are not realized. The realized phone rate is therefore much lower than in the fast, clear category, namely within 1 standard deviation from the sample mean.

NCf: A third category of *normal, clear phrases for comparison with fast intended phrases* was selected from the database, for which both the intended and the realized phone rates are similar to the realized phone rates of fast, sloppy speech. The phrases are spoken clearly, with a low percentage of phone deletions, comparable to those in the FC category. Their intended phone rates are within about 0.5 standard deviations of the sample mean. As is the case for the other categories, each of the phrases in the NCf category was carefully matched, also in terms of its duration, with the comparison phrases from the other two categories.

The stimuli in the FS category were selected on the basis of the numerous deletions, as in the phrase “Nee, da habe ich schon einen anderen Termin,” which is realized as [ne:, da: a:b ɪ ʃõ: n a̰nɐn tɛɐmi:n] (with only 20 out of 32 intended phones being realized). In many of the FS stimuli, schwa deletion (plus assimilation) occurs: “eig*e*ntlich,” “würd*e* es,” “Ihn*e*n,” “pass*e*n,” “ein*e*n,” “mach*e*n,” “vorschlag*e*n,” and “hab*e* ich” (the stimuli are listed in the Appendix, together with the transcriptions of the intended and realized phones). As expected, many of the schwas in the stimuli which were selected for the FC category on the basis of their low deletion rate are actually realized (high articulatory precision), even in clear reduction contexts like in “Revisionstreff*e*n,” “mach*e*n” (cf. Kohler and Rodgers, 2001). But, in general, far fewer intuitive deletion contexts are found in the FC stimuli. One could hypothesize that if there is no opportunity for deletion, this should not affect the perceived speech rate, i.e., given the identical intended rates for the FC and FS categories, they should be judged equal. On the other hand, following Kohler’s argument, it could also be argued that the lack of deletions, particularly in some of the contexts mentioned above, can be considered as hyperarticulation at a fast speech rate. Like the fast, clear stimuli (FC), the normal, clear (NCf) stimuli contain only few deletion contexts, and there are also a few contexts like “gan*z* *sch*lecht” and “d*e*n” where no deletion occurs, although the context seems appropriate for it. If the listeners derive their speech rate judgments solely from the actual speech rate without taking the canonical form into consideration, then the stimuli in the NCf category should be judged as equally fast as those in the FS category, which have the same realized phone rate.

Set 2: Comparison of slow with normal speech

To mirror the intonation phrases selected from the faster half of the database, another set of stimuli were selected from the slower half of the database. Comparison of the results for the two stimulus sets allows us to test the hypothesis that the different balance between input- and output-oriented constraints at fast versus slow rates affects the perceived speech rate. The clearly spoken phrases at slower rates are not necessarily considered as hyperspeech and may therefore have a different effect on the perceived speech rate compared to fast, clear utterances. Conversely, the pressure from the point of view of ease of articulation to reduce or delete segments is less at slower articulation rates, because there is more time to realize the segments at normal and slow than at high articulation rates. The set of stimuli from the slower half of the intended articulation rate range (normal to slow) offers an opportunity to examine whether the same regularities in speech rate judgments are found at faster and slower measured rates. The following stimulus categories were selected:

NCs: *Normal, clearly* spoken phrases were selected for comparison with other phrases with intended rates in the *slow half* of the articulation rate range. The intended and realized phone rates are similar to those of the NCf category, but the phrases are different ones. This is done in order to provide an optimal match with the phrases in the two following categories (also in terms of their durations).

NS: *Normal but sloppy* intonation phrases have intended phone rates close to the sample mean and 35%–40% of the intended phones are not realized (as in the FS category). This results in fairly slow realized phone rates between −0.5 and −1.5 standard deviations from the sample mean. Of all the categories in the experiment, this is the clearest instantiation of hypospeech, because at a normal intended articulation rate the pressure to reduce or delete segments is less than in the FS category.

SC: Finally, a category of *slow*, *clear* phrases is defined. The phrases are matched to the NS phrases in their realized phone rate. The intended phone rates are similar to the realized phone rates and lie between −1 and −2 standard deviations from the sample mean.

As in the NCf category, the stimuli in the NCs category hardly show any deletions. All schwas are realized, even in deletion contexts like in “wär*e*,” “ließ*e* sich,” or “mach*e*n,” except for a single schwa deletion in the final syllable of “dreizehnt*e*n” (cf. Kohler and Rodgers, 2001). Even the prevocalic glottal stops in “und” and “ich” are fully realized, with additional laryngealization of the vowel (Kohler, 1994; Wesener, 1999). As in the fast half of the database, the sloppy stimuli in the slower half (NS) contain many deletion contexts, like “hab*e*n,” “klein*e*n,” “Aug*e*nblick,” “morg*e*n,” "ein*e*m," "hätt*e*n," and "erst*e*n," all of which are realized without a schwa, and with assimilation where applicable. The stimuli in the SC category, which are very slow, contain few deletion contexts, and when they do the deletable segments are phonetically realized, like the schwa in "nehm*e*n" or the prevocalic glottal stops in "am," "und," and "auf." The presence of a clear glottal stop is quite usual for any of the normal and slow clear categories, whereas they are not realized (or, rarely, as creak in the following segment) in fast clear or in sloppy speech. Figure 1 displays the intended and realized speech rates of the six stimulus categories.

B. Listeners and their speaking habits

Since it was hypothesized that listeners' own speaking habits may affect their perception of speech rate, the 12 subjects for our listening experiments (five male and seven female in the age range 20–80, with an average age of 33) were selected on the basis of the subjective impression of their speech rate (subjective speech rate: fast or slow) in a consensus judgment by six advanced students in the Institute of Phonetics. There were thus two subjectively allocated listener groups: eight fast speakers and four slow speakers.

In order to also obtain a more objective estimate of the listeners' speech behavior, and particularly in order to see whether the subjective judgments (which may also be based on speech properties such as pausing and intonation) are related to the intended and realized *articulation* rates, each subject first read aloud a German text ("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. The recordings were subsequently divided into interpause stretches (ips). These were used instead of intonation phrases (used as stimuli in the listening experiments) because they are easier to determine. In the main, they correspond to intonation phrases. The differentiation of the speakers in terms of their speaking habits in a spontaneous speech task was based on the retold stories.

The intended and realized phone rates were computed for each ips and an *articulatory precision index* (api) was computed by dividing the realized phone rate by the intended phone rate (low api values indicate more deletions). The duration of each ips was also measured. Subsequently, the data were analyzed to see whether the subjective speech rate (as perceived by the judges) is reflected in the subjects' measured speech behavior. Finally, the subjects were also divided into objective articulation rate/speaking style categories on the basis of the measures taken from their speech.

1. Objective basis of subjective speech rate groups

As was pointed out in the Introduction, many factors beside articulation rate are likely to affect the perception of speech rate, so it is necessary to verify whether the division into fast and slow speakers on the basis of the judges' subjective impression of speech rate is (also) reflected in the measures used in this study. 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.

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

A total of 76 disfluencies were observed in the 683 ips. They are divided into seven categories: stutters, slips of the tongue, interruptions, hesitations, lengthenings, filled pauses, and laughters. As expected, fewer disfluencies occur for read than for retold stories. Given the low number of disfluencies, chi square tests are not appropriate, but the pattern shown for the retold stories in 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. 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 estimates, particularly because final lengthening will have a strong effect on the average number of phones per second.

For the read and retold stories, the duration of each ips is measured. Further, their intended and realized phone rates and the articulatory precision index are computed. 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 s on average) are significantly shorter ($t=2.4$, $df=217$, $p<0.05$) than for slow speakers (average: 2.1 s). 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. These results in general confirm that articulation rate properties, both intended and realized, go hand in hand with the subjectively perceived speech rate.

TABLE II. Three groupings of the speakers into four categories on the basis of the measured intended and realized rate and the articulatory precision in a retold text (dashes indicate combinations which do not occur).

		Category			
Grouping		1	2	3	4
1	intended rate	fast	fast	slow	slow
	realized rate	fast	slow	fast	slow
2	intended rate	–	fast	slow	–
	articulatory precision		sloppy	clear	
3	realized rate	fast	fast	slow	slow
	articulatory precision	clear	sloppy	clear	sloppy

The subjective speech rate will be used later to analyze the results from the listening experiments for the effect of speaking habits on the perceived speech rate.

2. Objective rate and style

Because the subjective impression of speech rate (viz. as either slow or fast) is in all likelihood not just based on articulation rate, but is probably also influenced by pausing behavior and other prosodic characteristics of the speakers as well as by disfluencies, an objective grouping of the subjects is obtained on the basis of cluster center analyses performed on measures reflecting the subjects' articulation behavior. Two groups (fast, slow) are created on the basis of the median values of the measured intended phone rates in the retold story; the same is done for the realized phone rates (fast, slow) as well as for the articulatory precision index values (clear, sloppy). Each two out of the three variables are combined to derive subject categories which reflect different articulation rates and speaking styles in a manner which is similar to the way the stimuli selected from the Kiel Corpus were categorized. These speaker groupings, which are summarized in Table II, are used in some of the later analyses. The clustering confirms that factors other than the intended and realized rates also influence the subjectively perceived speech rate, because one of the speakers with a slow subjectively perceived speech rate belongs to category 1 of grouping 1 (category 2 of groupings 2 and 3), while one of the speakers with a fast subjectively perceived speech rate belongs to category 4 of grouping 1 (category 3 of groupings 2 and 3).

C. Comparison experiment

Two experiments were carried out. The goal of the first, comparison experiment is to see how the different articulation rates and speaking styles in stimulus pairs affect the listeners' judgments of speech rate. The Experimenter software (Altman, 1992) was used to carry out the comparison experiment. In order to obtain a judgment of perceived speech rate which is not affected by the intelligibility of the stimuli, the two intonation phrases were first orthographically displayed on a computer screen for 3.5 s, followed by a beep and a silent pause of 0.5 s. Then the stimulus pair was played, separated by a silence interval of 0.5 s. The stimuli, which were set to equal loudness levels and adjusted for each listener to a comfortable volume, were played to the listeners over loudspeakers in a sound-treated room. After presentation of a stimulus pair the subject had 5 s to respond by pressing one of three keys on the keyboard for "first stimulus faster" (key: ←), "second stimulus faster" (key: →) or "both stimuli equally fast" (key: ↓). Both the response and the response time were registered.

All possible combinations of the six categories shown in Fig. 1 were compared, giving 15 comparisons. Five sets of six phrases were selected from the database. The total number of stimulus pairs was therefore 5 (sets) × 15 (comparisons) = 75 per listener. The word content of the phrases was always different and with a few exceptions the phrases were all produced by different speakers. The stimulus pairs were offered in pseudo-randomized order. 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, as well as interspersed with, filler items. Because the responses of the comparison experiment constitute nominal data which is mainly suitable for qualitative analysis, metrical data were collected in a scaling experiment, which is described in the following section.

D. Scaling experiment

The scaling experiment serves a double aim. First, it tests whether listeners are able to make consistent judgments of speech rate when they are required to judge each stimulus phrase by itself. Second, the speech rate judgments of the individual intonation phrases can be related to the listeners' own speech behavior. The same 30 intonation phrases as in the comparison experiment (5 sets × 6 categories) were used for the scaling experiment. The stimuli were offered in two blocks, each containing all the stimuli, but in a different randomized order. The blocks were preceded by five and followed by four filler items. Each phrase was shown on the computer screen for 1.75 s before it was played to the listener. As in the comparison experiment, the stimuli were played over loudspeakers in a sound-treated room. After hearing the phrase, the listener scored the perceived speech rate on a continuous scale from −3 ("too slow") to 3 ("too fast") by marking the position on the scale with a vertical line. The intermediate full values were labeled with "very slow" (−2), "quite slow" (−1), "normal" (0), "quite fast" (1), and "very fast" (2). The positions of the vertical lines were later measured with a precision of one decimal—thus metrical data are obtained, which are more suited to quantitative analysis than the data obtained from the comparison experiment. The extremes of the scale, which contain a subjective judgment, were not used by any of the listeners. After scoring a stimulus, the listener hit return to continue with the next stimulus.

III. RESULTS

A. Comparison experiment

For ease of interpretation, the results from the multiple comparisons between all stimulus categories are condensed into four types of stimulus pairs. In set A, both the intended

TABLE III. Comparison of intended (ipr) and realized phone rates (rpr) within the stimulus pairs for four conditions (see text), with the number of listener judgments and the compared rate conditions in their intrinsic order (cf. Fig. 1).

Condition	ipr	rpr	No. of judgments	Rate categories compared
A	different	different	480	FC-NCf, FC-NCs, FC-NS, FC-SC, FS-NS, FS-SC, NCf-SC, NCs-SC
B	same	different	180	FC-FS, NCf-NS, NCs-NS
C	different	same	180	FS-NCf, FS-NCs, NS-SC
D	same	same	60	NCf-NCs

and the realized articulation rates of the two intonation phrases in a stimulus pair are different. In set B, only the realized articulation rate is different, while in set C there is only a difference in intended rate. In set D, finally, both the intended and realized articulation rates of the two intonation phrases are the same. Table III lists the sets to which each of the comparisons between the categories in Fig. 1 belongs. Irrespective of actual order of presentation, results are given as percentage "stimulus faster" with respect to the *intrinsic order* of categories given in the column "rate categories compared" in Table III.

The perceived speech rate differences for each of the four sets are shown in Fig. 2. As expected, when both the intended and realized articulation rates of the two phrases are different, a large number of "first stimulus faster" responses is found [Fig. 2(A)], with relatively few "equal" responses. Figures 2(B) and 2(C) show that the effect of the intended and the realized rate is virtually the same: Relatively more "equal" responses are given when only one of the two measured rates differs for the two phrases, but the first stimulus of a pair is considered faster three to four times as often as the second. This is not the case when both the intended and the realized rates of the two phrases in the stimulus pair are the same [Fig. 2(D)]: Although the percentage of "equal" responses is similar to that in Figs. 2(B) and 2(C), there are fewer "first stimulus faster" than "second stimulus faster" responses (intrinsic order). The results confirm those reported in Koreman (2003), where a different set of 20 listeners carried out the same experiment.

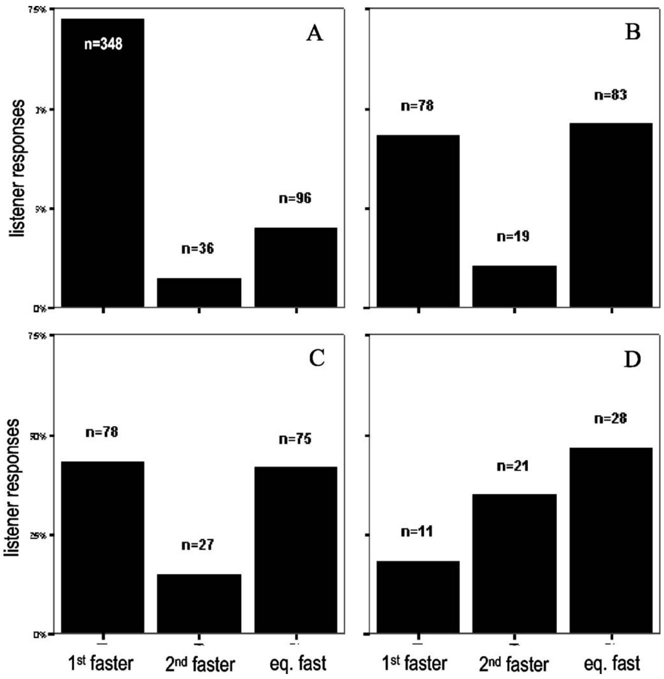

FIG. 2. Listener responses (percentages) to stimuli with different intended and realized rates (A), only different realized rates (B) or intended rates (C), or with both equal (D).

Of course, the effect for the stimuli in set A, shown in Fig. 2, is not only large because both the intended and realized rates of the stimuli in each pair differ, but also because this stimulus set contains comparisons between phrases with extreme articulation rate differences (e.g., FC-SC), while this is not true for the other stimulus sets. But the clear perceptual effect of differences in intended and realized rates is consistently reproduced when each two stimulus categories are compared separately (Fig. 3). If there is a difference in either the intended or the realized rate between the two stimuli in a pair, this difference is used by the listeners to determine which stimulus is faster (cf. Table III). The greater the distance from the top-left to bottom-right diagonal (the greater the differences in intended and/or realized articulation rates), the clearer the difference between the stimuli is perceived. This is true for the difference between the "first stimulus faster" and "second stimulus faster" responses, where an increasing preference towards "first stimulus faster" responses can be observed (black bars), as well as for the reduction in "equal" responses (dark gray bars). In one condition, namely NCf-NCs, phone rates are compared for stimuli which are similar in their intended as well as in their realized rates. As expected, the listeners are not able to decide which of the two stimuli was spoken faster, resulting in a large number of "equal" responses and even a reversal of "first stimulus faster" and "second stimulus faster" responses. Possible reasons why there are many "first stimulus faster" and "second stimulus faster" responses instead of only "equal" responses are that the listeners are encouraged by the task to judge one of the two stimuli as faster, but also that in the absence of segmental cues the listeners rely on stimulus properties which were not controlled in the present experiment, such as local speech rate, disfluencies, and intonational characteristics. There is no qualitative difference between stimulus comparisons from the fast half (dashed graphs) and the slow half of the database (dotted graphs).

For a more quantitative evaluation of how sure the listeners are that their response is correct, a certainty measure is derived for each listener across the five comparisons between each two categories. The measure is a Bayesian estimate of

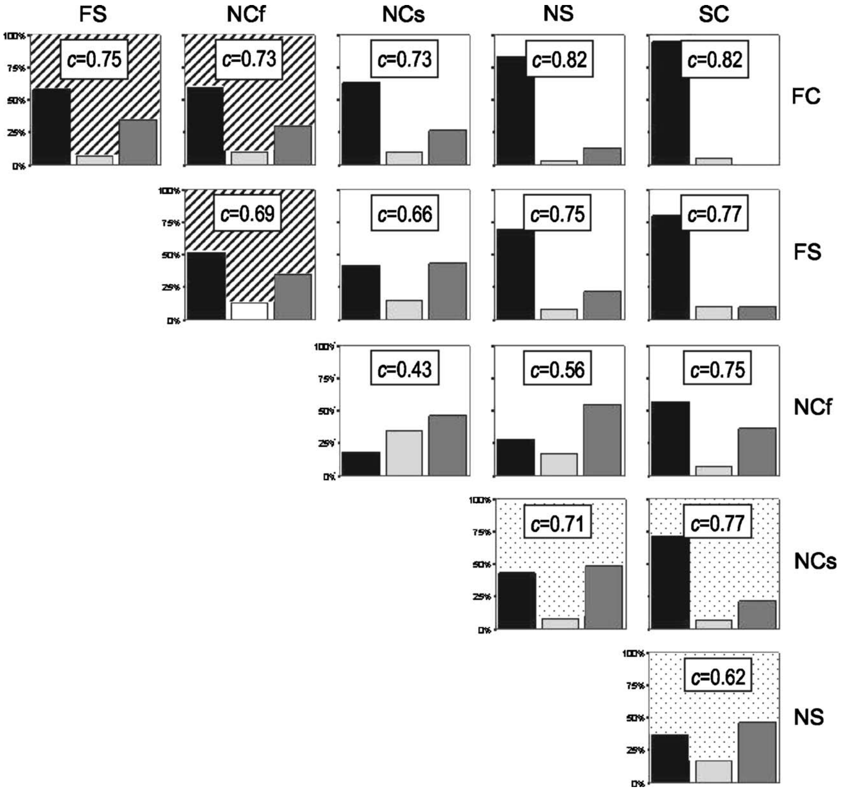

FIG. 3. Number of "first stimulus faster" (black bars), "second stimulus faster" (light gray bar), and "both stimuli equal" (dark gray bars) averaged for each stimulus pair. Rows indicate the category of the first stimulus in a pair, while columns show the category of the second stimulus (intrinsic order). The graphs with the dashed backgrounds represent comparisons between stimuli from the fast half of the database, while the ones with a dotted background are for stimulus comparisons for the slow half of the database. Certainty values c in each graph are derived from the size of the bars and indicate how certain the listeners are that their response is correct (see text).

the Bernoulli probability (Collings, 1997, p. 26), assuming a prior probability of 0.5; "equal" responses are considered as failed trials (but do affect the certainty measure), so that the certainty of giving the correct response can be computed as $c=(1+s1)/(2+s1+s2)$, where s1 is the number of "first stimulus faster" and s2 the number of "second stimulus faster" responses.[2] The lowest certainty value ($c=0.44$) is obtained for the comparison between NCf and NCs categories, where no difference ($c=0.50$) is expected. An analysis of variance ($F(14,165)=7.3$, $p<0.001$) with Tukey-HSD *posthoc* tests shows that this value is significantly lower ($p<0.05$) than that for all categories which differ in *either* intended or realized rate (or both)—except for the comparison of the NCf and NS categories, from which it does not differ significantly. The certainty values are indicated in Fig. 3. As can be seen from the figure, the certainty values increase as the distance between two compared categories (as recognizable from Fig. 1) increases. The same analysis for the four conditions in Fig. 2 showed a significant main effect ($F(3,176)=25.8$, $p<0.001$), with a significantly decreasing certainty value (Tukey-HSD *posthoc* tests at $p<0.05$) from condition A-D, except for the comparison between conditions B (only realized rates of the two stimuli different) and C (only the intended rates different). The results from the scaling experiment will show further differences between the perceived rates for the stimulus categories.

B. Scaling experiment

1. Comparison with the results of the first experiment

The stimuli were judged by the listeners in two blocks in the scaling experiments. The identified rates for repetitions of identical stimuli showed a high and significant correlation ($r=0.88$, $n=360$, $p<0.001$). An analysis of variance of the *average* rate scores across the two stimulus repetitions, with the five stimuli in each stimulus category as repeated measures, shows that the stimulus category affects the identified rate ($F(20,264)=11.1$, $p<0.001$). Each of the stimulus categories is significantly different from each other category, as shown by *posthoc* Tukey-HSD tests (all $p<0.001$), with two exceptions: (1) the two categories NCf and NCs are not significantly different, as expected, and (2) the average identified rates for stimuli from category FS are not significantly different from those for NCs stimuli, but this is mainly caused by one stimulus (they do differ significantly from the NCf stimuli). The results are in perfect agreement with those from the comparison experiment. The farther apart two categories are in their intrinsic order (as shown in Fig. 1), the greater the difference in average perceived speech rate (see Table IV). The slightly (but not significantly) higher average identified rate for the NCs stimuli than for the NCf stimuli supports the results from the comparison experiment, where comparison between the two categories led to more "second stimulus faster" (NCs faster than NCf) responses.

For the three response categories, significant differences are found both in the identified rates ($F(2,879)=109.5$, $p<0.001$) and in the reaction times for the responses ($F(2,879)=15.0$, $p<0.001$). In most stimulus comparisons, the choice for "first stimulus faster" is clear and the differ-

TABLE IV. Means and standard deviations of the identified rate for each stimulus category in a scaling experiment

	FC	FS	NCf	NCs	NS	SC
mean	1.12	0.40	−0.08	0.03	−0.58	−1.14
sd	0.58	0.65	0.74	0.78	0.75	0.66

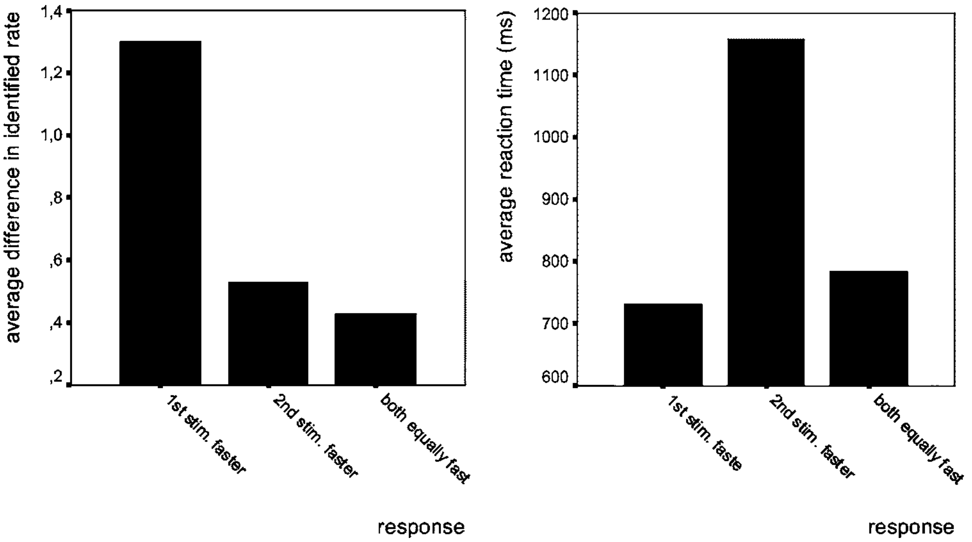

FIG. 4. Average differences in identified rate (left) and average response times (right) for each judgment.

ence in perceived speech rate of the stimuli is large, as shown by the first column in Fig. 4 (left). The differences in identified rate are small when listeners perceive the second stimulus as faster than or equally fast as the first one (second and third columns). The average differences in identified rate are significantly smaller for these two response categories than for the "first stimulus faster" category (Tukey-HSD *post hoc* test, $p < 0.001$). There are also obvious differences in response time to the stimulus pairs in the comparison experiment (Fig. 4, right). If the listeners base their judgments on the measured rates only, the second stimulus should never be judged as faster (because the results are always presented for the intrinsic order of the stimulus categories). When listeners nevertheless judge the second stimulus as faster than the first one, they take a long time for their response (second column). This is a clear indication of the uncertainty about their judgment. The difference with the other two categories is significant at $p < 0.001$ in a Tukey-HSD *posthoc* test.

2. Comparison with the listeners' speaking habits

An analysis of variance of the average identified rate of the stimuli with the stimulus categories (with NCf and NCs merged into a single category NC) and the subjective speech rate of the subjects (fast or slow, as judged by the experimenters) as independent variables confirms the influence of stimulus category on the identified rate ($F(16,248) = 6.3$, $p < 0.001$). The subjective speech rate, however, does not affect the perceived rate of the stimuli, nor is there an interaction between subjective speech rate and stimulus category. Also after standardizing the identified rate values for each speaker (by computing their z-score values) in order to compensate for a different use of the perception scale among individual listeners, there is no interaction between subjective speech rate and stimulus category.

The subjective speech rate of the subjects, which was used above, is partly based on other cues beside articulation rate (although this, too, was shown to differ). In order to use a more objective grouping of the subjects according to their speaking habits, the groupings shown in Table II (which are based on the speakers' *measured* articulation rates) were used to replace the subjective speech rate in the previous analysis of variance (but note that with only 12 subjects in total the tests are not very powerful). Several of the suggested groupings lead to significant differences for these "objective rate" clusters, but in all *posthoc* analyses fast, clear speakers group with slow, sloppy speakers. As with the subjective speech rate, there are no significant interactions between stimulus category and objective rate, either for the raw identified rate values or for their standardized (z-score) values.

IV. COMPARISON WITH DATABASE ANALYSIS

Using spontaneous speech stimuli in perception experiments has the advantage that the results are representative of conversational speech. But it also has the disadvantage that other variables cannot always be completely controlled. It was assumed that for the selection of our stimuli they constitute random variables in our experimental design. Careful inspection of the stimuli in terms of the hesitations, intonational properties (number and type of accents as well as the contours which connect them; boundary tones), and the relative number of function and content words did not show any *systematic* differences between the stimulus categories which could constitute an alternative explanation for our results.

The rationale for the analysis of the database from which the stimuli were selected in terms of the intended and realized rates of all the intonation phrases is that a confirmation of the differences found in the perception experiments gives further support for their perceptual relevance. At the time of analysis a total of 1329 files from the Kiel Corpus of Spontaneous Speech had been prosodically labeled. The files were divided up into intonation phrases on the basis of the prosodic labels (only one phrasing level was used), resulting in 5779 intonation phrases after exclusion of intonation phrases with only nonspeech material or hesitations. For intonation phrases with extreme perceived speech rates, the perceived rate was indicated by the labelers. 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

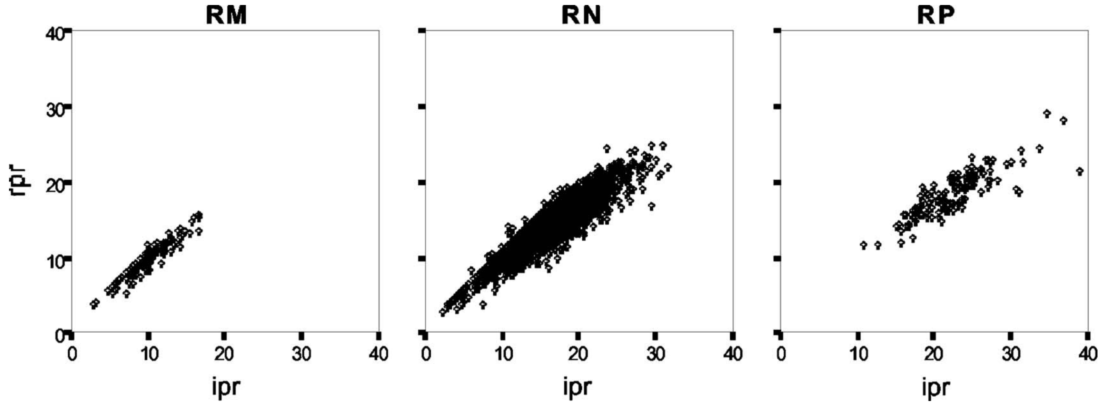

FIG. 5. Scatterplot of intended (ipr) versus realized rate (rpr) in phones per second for all intonation phrases with a minimum duration of 0.5 s perceived as slow (RM), normal (RN), and fast (RP).

were labeled by us as rate normal (RN)—though this does not preclude variation in rate, since it was from these phrases that all the stimuli in the two listening experiments were taken. 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%, 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" (Pfitzinger, 1999, p. 893).

For each intonation phrase, the number of intended and realized phones per second was determined.[3] There is a strong and highly significant correlation between the two rates (overall correlation: $r=0.93$, $n=4736$, $p<0.001$). Figure 5 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.

The considerable overlap in the measured phone rates of the different perceived rate categories indicates that factors other than articulation rate also determine the perceived speech rate. Nevertheless, an analysis of variance with subsequent Tukey-HSD *posthoc* tests shows 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. Figure 6 shows that, as expected, both the intended and the realized speech rate increase with perceived speech rate, while the articulatory precision index decreases.

V. DISCUSSION

As the results from the comparison experiment have shown, differences between the stimuli in intended and realized phone rates are clearly perceived [Fig. 2(A)]. Even if only the intended, canonical rate of two phrases differs (with identical realized rates), listeners perceive a difference in speech rate. In this case, the phrase with the faster intended rate is perceived as faster [Fig. 2(C)]. The experiment shows therefore that the perceived speech rate is not solely determined by actual articulatory events, but also by the listener's knowledge of what articulations are implied by a particular utterance.

On the other hand, knowledge of the implied articulations is not used to normalize differences in the realized ar-

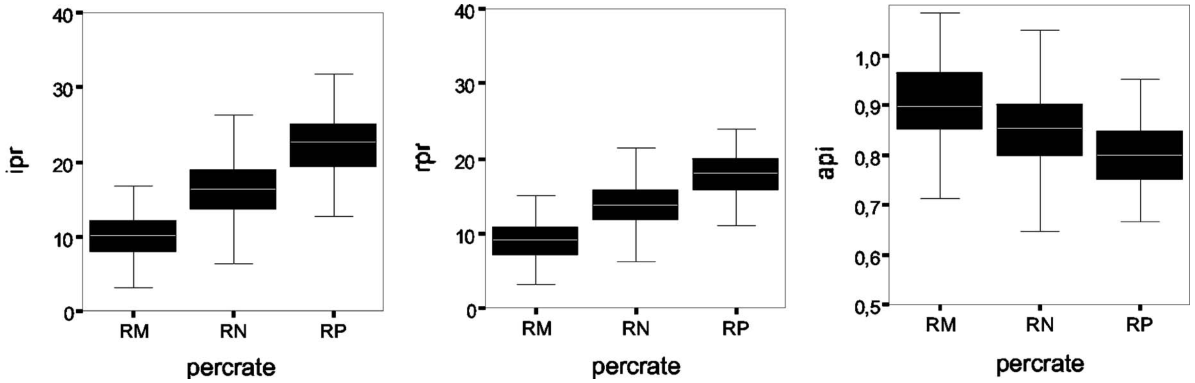

FIG. 6. Boxplots of intended (ipr) and realized phone rates (rpr) in phones per second as well as articulatory precision index (api) values for intonation phrases perceived as slow (RM), normal (RN), and fast (RP).

ticulation rates away. If this were the case, the perceived speech rate would have been entirely determined by the intended rate. As was noted in the description of the stimuli, the difference in realized rate between the FC and FS categories, and also between the NCs and NS categories (also NCf-NS), is at least in part caused by the lack of reduction contexts in the clear phrases. If listeners were to compensate for this when they judge speech rate, the stimuli should be considered equally fast. In reality, the higher realized articulation rates in clearly spoken phrases is taken at face value and interpreted as an indication of faster speech [Fig. 2(B)].

These results, and corroborating evidence from the scaling experiment showing differences between all stimulus categories except NCf-NCs (but also FS-NCs), clearly support hypothesis 1 that both the intended and the realized rate affect the perceived speech rate (see the Introduction). The significant differences in intended and realized rates between fast (RP), normal (RN), and slow (RM) speech in the extended database analysis seem to support the role of both rates for the perception of speech rate.

There are two possible interpretations of the joint contribution of intended and realized rates to the perceived rate. In the first interpretation it is assumed that the intended form has a real, psychological status. Listeners combine the actual, realized phone rate with their knowledge of the intended, canonical form to determine the perceived speech rate. But instead of claiming direct reference to listeners' knowledge of the intended form, there is a second possible explanation of our results within exemplar-based approaches to speech perception. If we do not want to assume reference to abstract units in speech perception before or simultaneously with the construction of meaning (Hawkins, 2003)—and this certainly does not seem necessary for the task—one must assume that the perceived speech rate is determined after the meaning of an utterance has been established (on the basis of the realized speech signal and the situational context in which it is produced), possibly by referring to functional groupings for the context. These functional groupings must then be indexed with information about the speech rate directly, or with information from which speech rate judgments can be derived, e.g., information about the time pressure under which conversations take place in which the observed realized forms have been used before, or information about the level of formality. The results from the listening experiments presented in this study are fully compatible with both theories.

The differences in perceived rate between categories from the faster half of the database (FC-FS-NCf) are the same as those in comparisons for the slower half (NCs-NS-SC). In general, there is no evidence that a difference in the balance between input- and output-oriented constraints at faster versus slower articulation rates in any way affects listeners' perception of speech rate (cf. hypothesis 2 in the Introduction). Although it was hypothesized that hyperspeech may be a typical fast speech phenomenon, clear speech in the slower half of the database has the same effect of making the utterance sound faster than an utterance with the same intended, but a lower realized rate (cf. the dashed with the dotted graphs in Fig. 3). The same is true for hypospeech: although it was suggested this may be more typical for slow speech, sloppiness has the same perceptual effect at high intended rates, making the utterance sound slower (cf. Fig. 3). There is no evidence from the perception experiments, therefore, to support the second hypothesis.

The results from the scaling experiment show that listeners do not only distinguish between articulation rates when two stimuli are presented for comparison (first experiment), but they can also judge them in isolation (second experiment). There is, however, no evidence that the perception of speech rate depends on the speaker-listener's own speaking habits. Subjectively fast speakers do not judge stimuli any differently from subjectively slow speakers. And although the effect of the objective groupings for the listeners' speech behavior on the identified rate of the stimuli was significant, these groupings are not consistent with the hypothesized effects of articulation rate and articulatory precision on the perceived rate of the stimuli. In particular, slow speakers were expected to evaluate the stimuli as faster on average than fast speakers, and sloppy speakers were expected to consider clear stimuli as faster (more hyperarticulated) than clear speakers, particularly in the case of fast, clear speech. But, in fact, fast, clear speakers group with slow, sloppy speakers in all statistical tests. This clearly contradicts our hypothesis. Thus, perceived speech rate does not seem to be a projection of the listener's own speech behavior.

The results presented in this paper described the effects of articulation rate and speaking style on the perceived speech rate. Of course, the perceived speech rate is also affected by other factors which still need to be investigated. For instance, *local* variation within an intonation phrase was not taken into consideration, although this may influence the perceived speech rate. Also, quantitative and qualitative phone reductions which do not affect the number of phones (e.g., vowel shortening or reductions from long, tense to short, lax vowels) and are therefore not reflected in our measures are likely to influence the perceived speech rate. Many of the deletions and reductions can be described using a small set of well-known phonological rules, such as schwa deletion, (nasal) assimilation and lenition of pre-vocalic glottal stops into laryngealized vowel onsets, or complete deletion of glottal stops (cf. Zellner, 1998; Kohler, 2001; Brinckmann and Trouvain, 2002). Further phonetic analysis of the database for the type of deletions and reductions at the different rates and subsequent evaluation of their perceptual effects are needed. The fact that other signal properties besides the articulation rate measures investigated here do affect the perceived speech rate is reflected in the discrepancy between the subjective and objective speaker groupings (cf. section on objective rate and style) as well as in the considerable overlap (shown in the section on database analysis) between normal and slow measured rates (RN and RM) and also between normal and fast measured rates (RN and RP) in the database. This stresses the importance of a phonetic analysis of the effect of nonsegmental properties such as disfluencies, the relative number of function versus content words in an utterance, and the intonation patterns at different perceived rates in the whole database to arrive at a more complete understanding of the speech characteristics which affect the perceived speech rate.

ACKNOWLEDGMENTS

The author should like to thank Dominik Bauer, Fabian Brackhane, Tamara Hausmann, Anja Moos, Olga Trifsik, and Steven Webber for their perseverance in running the experiments. Further thanks are due to Bistra Andreeva, Bill Barry, Andrew Morris, and Markus Pospeschill for discussion, as well as to the anonymous reviewers for their helpful comments on an earlier version of this paper.

APPENDIX: STIMULI AND THEIR TRANSCRIPTIONS

Intonation phrases for each of the stimulus categories used in this study, selected from the files of Kiel Corpus of Spontaneous Speech, together with their orthographic, intended, and realized phone transcriptions as well as their duration (dur), intended (ipr), and realized phone rates (rpr) and their articulatory precision index values (api). Phone counts are based on the conventions used in the Kiel Corpus transcriptions.

FC (fast clear)

File:	g114a009.s1h	api=1.00
Orthographic:	"…Revisionstreffen zu machen?"	dur=1.00 s
Intended:	/ʁɛvɪzjo:nstʁɛfən tsu: maχən/	ipr=22.98 phones/s
Realized:	[ʁɛvɪzjo:nstʁɛfən tsu: maχən]	rpr=22.98 phones/s
File:	g312a012.s1h	api=0.92
Orthographic:	"Und wann sind Sie dann wieder zurück?"	dur=1.19 s
Intended:	/ˀʊnt van zɪnt zi: dan vi:dɐ tsʊʁʏk/	ipr=21.85 phones/s
Realized:	[ʊnd van zɪnt zi: dan vi:ɐ tsʊʁʏk]	rpr=20.17 phones/s
File:	g077a003.s1h	api=0.94
Orthographic:	"Ich weiß ja nicht direkt, wie lange das dauert…"	dur=1.49 s
Intended:	/ˀɪç vaɪs ja: nɪçt dɪʁɛkt, vi: laŋə das daʊɐt/	ipr=20.78 phones/s
Realized:	[ɪ̰ç vaɪs ja: niç dɪʁɛkt, vi: laŋə das daʊɐd]	rpr=19.44 phones/s
File:	g113a011.s1h	api=0.93
Orthographic:	"…nämlich von Montag, dem neunzehnten…"	dur=1.33 s
Intended:	/nɛ:mlɪç fɔn mo:nta:k, de:m nɔɪntse:ntən/	ipr=21.06 phones/s
Realized:	[nɪmɪç fɔn mo:nta:χ, de:m nɔɪntse:ntn]	rpr=19.55 phones/s
File:	g197a009.s1h	api=0.92
Orthographic:	"…würde mir ganz gut in den Kram…"	dur=1.12 s
Intended:	/vʏɐdə mi:ɐ gants gu:t ˀɪn de:n kʁa:m/	ipr=23.07 phones/s
Realized:	[vʏɐdə mi:ɐ gants gu:t ɪ̰n ne:ŋ kʁa:m]	rpr=21.29 phones/s

FS (fast sloppy)

File:	g077a005.s1h	api=0.65
Orthographic:	"und da habe ich wieder eigentlich…"	dur=1.09 s
Intended:	/ˀʊnt da: ha:bə ˀɪç vi:dɐ ˀaɪgəntlɪç/	ipr=23.78 phones/s
Realized:	[ʊn da: ha:b ɪ̰ç vi:ɐ a̰ɪŋl̰ɪç]	rpr=15.55 phones/s
File:	g147a005.s1h	api=0.64
Orthographic:	"Würde es Ihnen da es gut passen?"	dur=1.10 s
Intended:	/vʏɐdə ˀɛs ˀi:nən da: ˀɛs gu:t pasən/	ipr=22.78 phones/s
Realized:	[vʏɐ əs ḭ:n da: s gu:t pasn]	rpr=14.58 phones/s
File:	g211a012.s1h	api=0.63
Orthographic:	"Ich hätte hier einen Vorschlag zu machen…"	dir=1.49 s
Intended:	/ˀɪç hɛtə hi:ɐ ˀaɪnən fo:ɐʃla:k tsu: maχən/	ipr=20.19 phones/s
Realized:	[ç hɛtə hi:ɐ n fo:ɐʃla:χ ts maχŋ]	rpr=12.79 phones/s
File:	g086a018.s1h	api=0.65
Orthographic:	"Ich würde Ihnen daher vorschlagen…"	dur=1.24 s
Intended:	/ˀɪç vʏɐdə ˀi:nən da:he:ɐ fo:ɐʃla:gən/	ipr=20.19 phones/s
Realized:	[ɪç vʏɐdə i:n dahe:ɐ fɔɐʃla:ŋ]	rpr=13.66 phones/s
File:	g193a009.s1h	api=0.62
Orthographic:	"Nee, da habe ich schon einen anderen Termin."	dur=1.25 s
Intended:	/ne:, da: ha:bə ˀɪç ʃo:n ˀaɪnən ˀandɐʁən tɛɐmi:n/	ipr=25.69 phones/s
Realized:	[ne:, da: a:b ɪ ʃõ: n a̰nɐn tɛɐmi:n]	rpr=16.06 phones/s

NCf (normal clear, for comparison with fast stimuli)

File:	g115a011.s1h	api=1.00
Orthographic:	"Das wird ja ganz schlecht."	dur=1.15 s
Intended	/das vɪɐt ja: gants ʃlɛçt/	ipr=15.63 phones/s
Realized:	[das vɪɐt ja: gants ʃlɛçt]	rpr=15.63 phones/s

NCf (normal clear, for comparison with fast stimuli)

File:	g121a003.s1h	api=1.00
Orthographic:	“Samstag bis Montag.”	dur=1.09 s
Intended:	/zamsta:k bɪs mo:nta:k/	ipr=14.64 phones/s
Realized:	[zamssa:χ bɪs mo:nta:k]	rpr=14.64 phones/s
File:	g212a007.s1h	api=0.94
Orthographic:	“Bis Mittwoch, den achten?”	dur=1.33 s
Intended:	/bɪs mɪtvɔχ, de:n ʔaχtən/	ipr=13.52 phones/s
Realized:	[bɪs mɪtvɔχ, de:n ʔa̰χtn]	rpr=12.77 phones/s
File:	g195a001.s1h	api=0.94
Orthographic:	“Nee, Samstag bin ich schon…”	dur=1.22 s
Intended:	/ne:, zamsta:k bɪn ʔɪç ʃo:n/	ipr=14.71 phones/s
Realized:	[ne:, zamssa:χ bɪn ɪç ʃo:n]	rpr=13.89 phones/s
File:	g091a024.s1h	api=0.92
Orthographic:	“Wie wäre es mit Ende November?”	dur=1.34 s
Intended:	/vi: vɛ:ʁə ʔɛs mɪt ʔɛndə no:vɛmbɐ/	ipr=17.89 phones/s
Realized:	[vi: ve:ʁ əs mɪt ʔɛ̰ndə no:vɛmbɐ]	rpr=16.40 phones/s

NCs (normal clear, for comparison with slow stimuli)

File:	g312a019.s1h	api=1.00
Orthographic:	“…gleich drei Termine fertig.”	dur=1.14 s
Intended:	/glaɪç dʁaɪ tɛɐmi:nə fɛɐtɪç/	ipr=15.83 phones/s
Realized:	[glaɪç dʁaɪ tɛɐmi:nə fɛɐtɪç]	rpr=15.83 phones/s
File:	g105a010.s1h	api=1.00
Orthographic:	“Ja, das wäre günstig…”	dur=1.13 s
Intended:	/ja: das vɛ:ʁə gʏnstɪç/	ipr=14.17 phones/s
Realized:	[ja: das vɛ:ʁə gʏnstɪç]	rpr=14.17 phones/s
File:	g117a005.s1h	api=1.00
Orthographic:	“Und ich denke mal…”	dur=1.08 s
Intended:	/ʔʊnt ʔɪç dɛŋkə ma:l/	ipr=13.84 phones/s
Realized:	[ʔʊ̰nt ʔɪ̰ç dɛŋkə ma:l]	rpr=13.84 phones/s
File:	g254a007.s1h	api=0.95
Orthographic:	“Bis zum dreizehnten Juli?”	dur=1.46 s
Intended:	/bɪs tsʊm dʁaɪtse:ntən ju:li:/	ipr=14.40 phones/s
Realized:	[bɪs tsʊm dʁaɪtse:ntn ju:li:]	rpr=13.71 phones/s
File:	g075a002.s1h	api=1.00
Orthographic:	“…ließe sich dann schnell machen…”	dur=1.07 s
Intended:	/li:sə zɪç dan ʃnɛl maχən/	ipr=17.69 phones/s
Realized:	[li:sə zɪç dan ʃnɛl maχən]	rpr=17.69 phones/s

NS (normal sloppy)

File:	g215a001.s1h	api=0.67
Orthographic:	“Dann haben wir half nur…”	dur=1.17 s
Intended:	/dan ha:bən vi:ɐ halt nu:ɐ/	ipr=15.43 phones/s
Realized:	[dan ha:m vi:ɐ halt nu:ɐ]	rpr=10.29 phones/s
File:	g072a008.s1	api=0.67
Orthographic:	“Kleinen Augenblick…”	dur=1.06
Intended:	/klaɪnən ʔaʊgənblɪk/	ipr=14.10 phones/s
Realized:	[klaɪn a̰ʊŋblɪk]	rpr=9.40 phones/s
File:	g317a006.s1h	api=0.62
Orthographic:	“…den wir morgen haben…”	dur=1.16 s
Intended:	/de:n vi:ɐ mɔɐgən ha:bən/	ipr=13.82 phones/s
Realized:	[de:n vi:ɐ mɔɐŋ ha:m]	rpr=8.64 phones/s
File:	g105a005.s1h	api=0.59
Orthographic:	“…um neun Uhr in einem Hotel…”	dur=1.48 s
Intended:	/ʔʊm nɔɪn ʔu:ɐ ʔɪn ʔaɪnəm ho:tɛl/	ipr=14.86 phones/s
Realized:	[ʊ̰m nɔɪn ṵ:ɐ ɪ̰n m ho:tɛl]	rpr=8.78 phones/s
File:	g142a008.s1h	api=0.65
Orthographic:	“Hätten Sie in der ersten…?”	dur=1.18 s
Intended:	/hɛtən zi: ʔɪn de:ɐ ʔɛ:ɐstən/	ipr=17.03 phones/s
Realized:	[hɛtn zi: ɪn de:ɐ ɛ̰:ɐstn]	rpr=11.07 phones/s

SC (slow clear)

File:	g251a024.s1h	api=1.00
Orthographic:	"…zum Beispiel am…"	dur=1.28 s
Intended:	/tsʊm baɪʃpi:l ʔam/	ipr=10.15 phones/s
Realized:	[tsʊm baɪʃpi:l ʔa̰m]	rpr=10.15 phones/s
File:	g253a004.s1h	api=1.00
Orthographic:	"Und freitags?"	dur=1.17 s
Intended:	/ʔʊnt fʁaɪta:ks/	ipr=9.38 phones/s
Realized:	[ʔʊnt fʁaɪta:ks]	rpr=9.38 phones/s
File:	g217a006.s1h	api=1.00
Orthographic:	"…beschäftigt…"	dur=1.08 s
Intended:	/bəʃɛftɪçt/	ipr=8.38 phones/s
Realized:	[bəʃɛftɪçt]	rpr=8.38 phones/s
File:	g311a007.s1h	api=1.00
Orthographic:	"…den vielleicht auf…"	dur=1.34 s
Intended:	/de:n fɪlaɪçt ʔaʊf/	ipr=8.94 phones/s
Realized:	[de:n fɪlaɪçt ʔa̰ʊf]	rpr=8.94 phones/s
File:	g085a004.s1h	api=1.00
Orthographic:	"…Dienstag nehmen?"	dur=1.07 s
Intended:	/di:nsta:k ne:mən/	ipr=11.22 phones/s
Realized:	[di:nsta:k ne:mən]	rpr=11.22 phones/s

[1] As others, we shall use the terms "clear" and "sloppy" as an opposition to describe the speaker's care of articulation, although the first term refers to perception—its articulatory counterpart could be "precise" or even "overprecise" if the possibility of spelling pronunciations is included. No negative connotations are intended with the word "sloppy," even if in everyday use it may imply that the speaker is *too* careless of his/her articulation.

[2] For the five comparisons between the categories FC and FS, for instance, listener HR responded twice with "first stimulus faster," once with "second stimulus faster," and twice with "equal" (response code 2-1-2), so that $c = (1+2)/(2+2+1) = 0.6$. The same number of "first stimulus faster" responses, but with more "equal" responses, as in response code 2-0-3, would result in $c = 0.75$, i.e., a higher certainty, because there are no "second stimulus faster" responses which go against expectation. If the subject always (five times for each comparison between two stimulus categories) responds with "first stimulus faster" (response code 5-0-0), this results in $c = 0.86$; if (s)he consistently chooses "second stimulus faster (response code 0-5-0), $c = 0.14$. Both response codes 0-0-5 and 2-2-1 result in $c = 0.5$ (uncertainty).

[3] The number of intended and realized syllables per second as well as the number of orthographic words per second were also derived. The correlation between syllable and phone rates varied between $r = 0.79$ and $r = 0.88$ ($p < 0.001$). The correlation between the orthographic word rate and the other four measures was between $r = 0.49$ and $r = 0.62$ ($p < 0.001$).

Altman, G. (**1992**). *Experimenter. A Toolkit for Multi-Modal Psycholinguistic Experimentation on the Apple Macintosh* (Laboratory of Experimental Psychology, University of Sussex).

Brinckmann, C., and Trouvain, J. (**2002**). "The role of duration models and symbolic representation for timing in speech synthesis," Int. J. Speech Technolo. **6**, 21–31.

Bromberger, S., and Halle, M. (**2000**). "The ontology of phonology (revised)," in *Phonological Knowledge: Conceptual and Empirical Issues*, edited by N. Burton-Roberts, P. Carr, and G. Docherty (Oxford U. P., Oxford).

Butcher, A. (**1981**). "Phonetic correlates of perceived tempo in reading and spontaneous speech," *Work in Progress*, University of Reading, Vol. 3, pp. 105–117.

Collings, S. N. (**1997**). *Fundamentals of Statistical Inference, Unit 9. Decision Theory and Bayesian Inference*, course M341, pp. 7–45 (Open U. P., Walton Hall, Milton Keynes).

Crystal, T. H., and House, A. S. (**1990**). "Articulation rate and the duration of syllables and stress groups in connected speech," J. Acoust. Soc. Am. **88**, 101–112.

Dankovičová, J. (**1999**). "Articulation rate variation within the intonation phrase in Czech and English," *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS)*, San Francisco, Vol. 1, pp. 269–272.

Den Os, E. (**1985**). "Perception of speech rate, a scaling experiment," Prog. Rep. Inst. Phonetics Utrecht (PRIPU) **10**, 35–43.

Duden (**2000**). *Duden: Aussprachewörterbuch* (Mannheim, Dudenverlag).

Fougeron, C., and Jun, S.-A. (**1998**). "Rate effect on French intonation: Prosodic organization and phonetic realization," J. Phonetics **26**, 45–69.

Funk and Wagnalls (**1973**). *Standard Dictionary of the English Language, International Edition* (Funk & Wagnalls, New York).

Gay, T. (**1981**). "Mechanisms in the control of speech rate," Phonetica **38**, 148–158.

Goldman-Eisler, F. (**1968**). *Psycholinguistics* (Academic Press, London).

Hawkins, S. (**2003**). "Roles and representations of systematic fine phonetic detail in speech understanding," J. Phonetics **31**, 373–405.

IPDS (**1994**). *The Kiel Corpus of Read Speech, CD-ROM #1* (Institut für Phonetik und digitale Sprachverarbeitung, Kiel), www.ipds.uni-kiel.de/forschung/kielcorpus.en.html, last viewed: 10.06.2004.

IPDS (**1995–1997**). *The Kiel Corpus of Spontaneous Speech, CD-ROM #2–4* (Institut für Phonetik und digitale Sprachverarbeitung, Kiel), www.ipds.uni-kiel.de/forschung/kielcorpus.en.html, last viewed: 10.06.2004.

Kohler, K. J. (**1986**). "Parameters of speech rate perception in German words and sentences: Duration, F0 movement and F0 level," Lang Speech **29**, 115–139.

Kohler, K. J. (**1994**). "Glottal stops and glottalization in German," Phonetica **51**, 38–51.

Kohler, K. J. (**2001**). "Articulatory dynamics of vowels and consonants in speech communication," J. Int. Phonetic Assoc. **31**(1), 1–16.

Kohler, K. J., and Rodgers, J. (**2001**). "Schwa deletion in German read and spontaneous speech," Arb. Inst. Phonetik Dig. Signalverarbeitung Universität Kiel (AIPUK) **35**, 97–123.

Koreman, J. (**2003**). "The perception of articulation rate," *Proc. 15th Int. Congress of Phonetic Sciences (ICPhS)*, Barcelona.

Lass, N. J. (**1970**). "The significance of intra- and intersentence pause times in perceptual judgments of oral reading rate," J. Speech Hear. Res. **13**, 777–784.

Lindblom, B. (**1963**). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. **35**, 1773–1781.

Lindblom, B. (**1990**). "Explaining phonetic variation: A sketch of the H&H theory," in *Speech Production and Speech Modelling*, edited by W. J. Hardcastle and A. Marchal (Kluwer Academic Publishers, Dordrecht).

Pfitzinger, H. R. (**1999**). "Local speech rate perception in German speech," *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS)*, San Francisco, Vol. 2, pp. 893–896.

Rietveld, T., and Gussenhoven, C. (**1987**). "Perceived speech rate and intonation," J. Phonetics **15**, 273–285.

Trouvain, J., and Grice, M. (**1999**). "The effect of tempo on prosodic structure," *Proc. 14th Int. Congress of Phonetic Sciences (ICPhS)*, San Francisco, Vol. 2, pp. 1067–1070.

Wesener, T. (**1999**). "The phonetics of function words in German spontaneous speech," Arb. Inst. Phonetik Dig. Signalverarbeitung Universität Kiel (AIPUK) **34**, 327–377.

Zellner, B. (**1998**). "Temporal structures for fast and slow speech rate," *Proc. 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, pp. 143–146.