Running head: JND FOR TEMPO IN SPEECH

On the just noticeable difference for tempo in speech

Hugo Quené

Utrecht institute of Linguistics OTS, Utrecht University

PHON06-19R2

12 September 2006

address for correspondence:

Trans 10, 3512 JK Utrecht, The Netherlands

hugo.quene@let.uu.nl

Abstract

Speakers vary their speech tempo (speaking rate), and such variations in tempo are quite noticeable. But what is the just noticeable difference for tempo in speech? The present study aims at providing a realistic and robust estimate, by using multiple speech tokens from multiple speakers. The JND is assessed in two (2IAX and 2IFC) comparison experiments, yielding an estimated JND for speech tempo of about 5%. A control experiment suggests that this finding is not due to acoustic artefacts of the tempo-transformation method used. Tempo variations within speakers typically exceed this JND, which makes such variations relevant in speech communication.

Keywords: tempo; speaking rate; just-noticeable difference; difference limen; discrimination;

On the just noticeable difference for tempo in speech

Introduction

Human speech is produced by moving the vocal organs, specifically the articulators. These movements result in an articulated speech signal, in which phonetic events occur at particular moments in time. The rate at which these speech events occur constitutes the tempo or speed or rate of speech. Many textbooks in phonetics state that speakers vary their speaking rate, in anticipation of the time listeners will need to process their words. Hence, important or unpredictable portions are spoken at a relatively slower rate (e.g., Zwaardemaker & Eijkman, 1928; Nooteboom & Cohen, 1984). This tendency follows from the adaptation principle or hyperspeech-and-hypospeech principle (H&H, Lindblom, 1989): speakers adjust phonetic properties of their speech to ensure an optimal balance between economy of articulatory energy, and perceptual clarity for the listener. After all, speakers speak in order to be understood. This phonetic principle underlies the usual rhetoric advice to public speakers, to slow down when important information is conveyed (e.g., Humes, 2002).

According to the same principle, listeners may interpret speech tempo as an indicator of the importance of what is being said. This communicative use of changes in speech tempo is only possible, of course, if these tempo changes exceed the perceptual threshold. But what exactly is the difference limen (DL) or just noticeable difference (JND) for speech tempo or speaking rate? If a speaker changes tempo, how large does the tempo change have to be in order to become noticeable, and hence relevant in speech communication? The aim of the present study is to answer these questions.

Previous research on JNDs for tempo has concentrated on perception of music, and most of these studies investigated JNDs for gradual and continuous changes in tempo, i.e. gradual accelerations and decelerations, rather than tempo itself. Ellis (1991) presented listeners with a 6-bar, 24-beat musical fragment at various base tempi. After a stable period (of random duration), the tempo of the fragment started to drift gradually (up or down, with +2% or –2% on each subsequent beat, to extremes of either +16% or –10%). Using the staircase adjustement method, JNDs of 5.1% to 13.9% change in tempo were found, with thresholds depending on the direction of drift and on the base tempo.

Madison (2004, Experiment 1) presented listeners with accelerating and decelerating click sequences, in which the tempo gradually drifted up or down from an initial tempo of either 100 or 120 beats per minute (IOI or inter-onset interval of 0.6 or 0.5 s). Sequence length varied between 2 and 9 clicks; stimuli were presented using the PEST adjustement method. Typical results showed a JND of about 4% for 5-click sequences. The JND decreases exponentially with increasing sequence length.

In a study of discriminability of tempo (and not of change in tempo), Drake and Botte (1993, Experiment 3) presented listeners with two 5-tone sequences to compare (2IFC paradigm). Using the staircase adjustement method, they found JNDs for this type of stimuli to be 6% to 10% for nonmusical listeners, and 3% to 8% for musicians, with the lowest JND at a base rate of 100 beats per minute (IOI of 0.6 s). In an ERP study having a similar design, Pfeuty, Ragot, and Pouthas

(2003) reported that a 4% change in inter-onset interval in a 7-tone sequence yields a discriminability value d' of 1.52, which suggests that the JND is smaller than 4% in their experiment.

Levitin and Cook (1996) cite an interesting unpublished study by Perron (1994) involving computer sequencers or drum machines, as used in popular music. Although such machines turned out to have an average tempo deviation of 3.5%, most listeners do not notice these deviations (and neither do professional drummers). This suggests that the JND for musical tempo is at least 3.5%.

In summary, these studies show that the JND for musical tempo is quite variable, depending on various stimulus properties, and on the measurement method. Most studies report an estimated lower limit of discriminability at about 4% or 5% of the base tempo. For tempo in *speech*, however, only a few, somewhat questionable estimates of the JND have been reported.

Benguerel and D'Arcy (1986, Experiment 4) presented a few pre-selected listeners with reiterant speech stimuli /nananananana/, with exponentially decreasing or increasing duration of each subsequent syllable. Listeners judged sequences as "regular", even for decelerating sequences (with increasing duration for each subsequent syllable). However, a conventional JND value cannot be derived easily from their results, which are primarily concerned with detecting gradual *changes* in tempo. In addition, it is not clear whether and how their findings generalize to normal speech and to other listeners.

Eefting and Rietveld (1989) presented listeners with two versions of a single speech utterance (2IFC paradigm); tempo was always unchanged in one version. The reported JND of 4.4% is remarkably low, and even lower than some values reported above for musical tempo. Eefting and Rietveld argue that this may be due to listeners' adaptation to their stimuli. These consisted of various tempo-manipulated versions of (one token of) a single sentence by one speaker. The repeated presentations allow for unnatural adaptation to that single speech fragment. The lack of ecological validity in this experiment makes it questionable whether these results can be generalized to other speech stimuli in more natural contexts.

Nooteboom and Eefting (1994) presented listeners with 6 resynthesized sentences, which were concatenated from 3–5 inter-pause phrases, in various tempo conditions. Phrase durations were (a) original: as produced in sentence context, with natural tempo variation between phrases, (b) uniform: as produced in a carrier sentence, without tempo variation between phrases [average syllable durations per phrase deviated from the original phrases by ratios between –20% and +20%], (c) variable: the uniform version (b) of a phrase was compressed or expanded to yield the same overall duration of the natural version (a) of that phrase [the concatenated sentence has natural tempo variation between phrases], and (d) unnatural, viz. with unnatural phrase durations. Intonation was identical in all resynthesized versions of a sentence. Comparisons among non-identical conditions (in a 2IAX paradigm) yielded d' values of 3.37 on average ($s = 0.89$); all d' values were greater than 1. This indicates that the tempo manipulations by about 20% were well above the JND for speech tempo, i.e., that JND is below 20%. However, a more exact estimate of the JND cannot be derived from this study.

In summary, the few studies of just-noticeable differences for tempo in speech do not provide us with a valid and robust estimate about how large tempo changes need to be in order to be relevant for communicative purposes. Benguerel and D'Arcy (1986) provide a somewhat unconventional investigation of the discriminability of local decelerations, rather than of speech tempo itself. The study by Eefting and Rietveld (1989) used a proven method, using a longer speech fragment, but only one such stimulus fragment was presented repeatedly, which hampers the validity of that study. Finally, the study by Nooteboom and Eefting (1994) used a valid method as well (with only 6 stimulus sentences), but the tempo variations were too large to provide a valid JND. Hence, the aim of the present study is to provide a valid and robust estimate of this JND for speech tempo.

The present study will report three experiments. The classical method to determine a JND is to present two stimuli shortly after each other, and to ask listeners to compare these stimuli. Two such pairwise discrimination experiments are reported, using 2IAX and 2IFC paradigms, respectively. These experiments use the same set of stimuli, which consist of 20 speech fragments spoken by 4 female speakers. Hence, the stimuli vary not only in tempo (within each pair under comparison), but also in their linguistic content and in speaker voice (between pairs). Presumably, these variations improve on the ecological validity and the robustness of the resulting JNDs. Tempo was manipulated by means of the PSOLA technique for time manipulation (Moulines & Charpentier, 1990; Moulines & Verhelst, 1995) which allows us to compress or expand the time base with very few changes in pitch and in spectral information. Nevertheless, one could argue that this tempo-manipulation procedure may have introduced artificial cues for tempo discrimination. A third control experiment is also reported, therefore, which investigates and rejects this alternative explanation.

Experiment 1: 2IAX

Method

Stimulus materials consisted of 20 speech fragments, excerpted from longer text passages that resembled short news items, consumer reports, personal anecdotes, etc. Four female speakers spoke five passages each, at a normal rate. Their productions were recorded on DAT, and re-digitized (16 kHz, 16 bits). Each passage contained about 10 s of speech in two or three sentences. Each passage was pruned to a fragment that does not contain major pauses (the resulting fragment usually corresponds to a major phrase). Means and standard deviations over the 20 fragments were as follows: duration 3.035 s (0.553), length 8 words (2), length 13.5 syllables (3.4), tempo 159.9 words per minute (37.4), average syllable duration 239 ms (77).

These fragments were then accelerated to relative durations of 0.80, 0.85, 0.87, 0.89, 0.91, 0.93, and 0.95 relative to the original duration, and decelerated to relative durations of 1.05, 1.07, 1.09, 1.11, 1.13, 1.15 and 1.20, yielding 7×2 manipulated versions plus 1 unmanipulated version for each fragment. Temporal compression or expansion was uniform throughout the fragment.

Listeners were 24 native Dutch-speaking students of a PABO college in Utrecht, who had no self-reported hearing or speech defects. They heard two versions of the same passage, with a 600 ms interval between the versions. One version was always the original or reference version, the other was one of the 14 manipulated versions or the original version. Each pair was presented in two orders, with the reference version as either the first or last member of the presentation pair. Listeners' task was to indicate whether the two versions were the same or different (as in Nooteboom & Eefting, 1994). The order of the 640 pairs (20 fragments × 8 versions × 2 orders × 2 directions, with the unmanipulated version occurring in both orders) was randomized anew for each listener. Listeners were tested individually in a quiet room. They indicated their same-or-different response by pressing one of two keys, of which the "different" key was always under their dominant hand. Total time of each session was approximately 1.5 hours, including 3 short pauses at regular intervals.

Data from three listeners were discarded: one because of her high miss rates, one because of her extreme bias towards "different" responses (yielding d' values smaller than 1 for the whole stimulus continuum), and the third because of a mild but noticeable form of stuttering in her speech. Data for one manipulated version were also discarded, leaving 19 stimulus fragments for this manipulation, because the corresponding audio file had been defective during the experiment.

Results and discussion

In a 2IAX (same–different) design, the JND is sometimes defined as the difference in tempo or duration at which half of the responses are "different". The present experiment contains a response bias, however, which renders this procedure inappropriate. If both members of a pair are the same, then no "different" responses are expected. Because all versions were compared with an unmanipulated reference version, this situation only occurred with both versions unmanipulated. However, listeners incorrectly judged these pairs as "different" in 15.1% of their responses; this false-alarm rate deviates significantly from zero (with standard error of 3.7% between individual listeners' average hit rates, $p < .001$). This possible bias may stem from the 2IAX design, which refers to a criterion internal to the listener (a criterion of equality or difference), which makes this type of experiment susceptible to response bias.

Hence, the JND was defined here by means of d' values, because these are based on the percentages of hits as well as false alarms. Since d' cannot be calculated from hit rates and false alarm rates of 0 and 1, these values were first recoded to .01 and .99 (for 4 and 12 cases, respectively; cf. Macmillan, Kaplan, & Creelman, 1977, p.465). Separate d' values were then obtained for each version for each of the 21 remaining listeners (pooled over 20 fragments). These d' values were adjusted for effects of 2IAX designs, according to the procedure recommended by Macmillan and Creelman (1991, p.145). First, an adjusted percentage correct $p(c)$ was calculated from the unadjusted d'. The adjusted d' was then calculated from this adjusted $p(c)$ alone, without the false alarm rate. The resulting adjusted d' values are plotted in Figure 1. By definition, the JND equals the difference in tempo, or relative duration, at which $d' = 1$. Using linear interpolation, average JNDs in this 2IAX experiment were estimated to be –3% for accelerating tempo (SE 0.4%), and +5% for decelerating tempo (SE 0.6%).

Insert Figure 1 about here

The results show that d' values are higher for accelerated speech than for decelerated speech. This was confirmed in a two-way repeated measures ANOVA of d' values, with direction-of-change (accelerated–decelerated) and amount-of-change (.05, .07, .09, .11, .13, .15, .20) as fixed factors. The main effect of direction-of-change was indeed significant $[F(1, 15) = 50.7; p < .001]$, as was the effect of amount-of-change [$F(6, 10) = 115.6; p < .001$]; the interaction effect was not significant [$F(6, 10) = 1.7; n.s.$]. The higher d' values yield a smaller JND for accelerated speech than for decelerated speech. This was confirmed in a one-way repeated measures ANOVA of listeners' individual JND values, with direction-of-change (accelerated vs. decelerated) as a fixed factor, yielding a significant effect [$F(1, 20) = 8.9; p = .007$]. The effect size is about .55, which is regarded as medium (Cohen, 1988).

In decelerated speech, the phonetic events by which listeners measure speech tempo, such as vowel onsets (Allen, 1972), are further apart: inter-onset intervals (IOIs) are longer. In music perception, larger JNDs have been reported for discrimination of tempi with longer IOIs (e.g. Drake & Botte, 1993). Hence, the poorer performance for decelerated speech may be explained by the longer IOIs in these decelerated speech fragments. Pouliot and Grondin (2005) reported a similar difference in sensitivity for accelerations and decelerations in musical tempo, which they explain as a bias due to the listener's internal timekeeper: if the tempo deviates more from that of the internal timekeeper, then that tempo is easier to discriminate. As an additional explanation, listeners in the present experiment may have had more difficulty in maintaining good memory traces of decelerated speech fragments (as compared to accelerated fragments), because the decelerated fragments were longer in overall duration than either the unchanged or accelerated fragments.

A general problem with 2IAX experiments is that they force a listener to refer to a subjective criterion of equality or sameness, or to a subjective reference tempo. As we saw above, this may have biased listeners' responses. A 2IFC experiment does not suffer from this disadvantage, since the two versions of a pair are only compared with each other (first–second is faster), and not with a subjective criterion or reference. In order to obtain more estimates of the JND, the present 2IAX experiment was also run as a 2IFC experiment.

Experiment 2: 2IFC

Method

Stimulus materials in this 2IFC experiment were identical to those used in Experiment 1.

Listeners were 22 native Dutch-speaking students of a PABO college in Utrecht, who had no self-reported hearing or speech defects. None of them had participated in the 2IAX experiment reported above. The presentation procedure was identical to Experiment 1. In this 2IFC experiment, however, listeners' task was to indicate whether the first or the second fragment was the faster one (as in Den Os, 1985; Drake & Botte, 1993). They indicated their response by pressing one of two keys, of which the "second fragment faster" key was always under their dominant hand. Total time of each session was approximately 1.5 hours, including 3 short pauses at regular intervals. Data for one defective manipulated version were again discarded.

Results and discussion

Listeners' responses were recoded to percentages of "faster" judgements for each version for each listener (pooled over 20 fragments); these judgements were regarded as hits. Since d' cannot be calculated from hit rates of 1, these values were first recoded to .99 (29 cases) (cf. Macmillan et al., 1977, p.465). False alarm rates were again obtained from stimuli consisting of two unmanipulated versions.

Separate d' values were then calculated for each version for each listener, based on both hit rate and false alarm rate. These d' values were adjusted for effects of 2IFC designs, according to the procedure recommended by Macmillan et al. (1977, p.456), viz. multiplying the unadjusted d' by a correction factor of $\sqrt{2}$. The resulting adjusted d' values are plotted in Figure 2. Individual d' values in Figure 2 are limited by individual listeners' false alarm rates. For example, one listener yielded a false alarm rate of .329; hence his or her maximum $d'_{max} = \sqrt{2}\,|z(.99) - z(.329)| = 3.915$, as can be seen in the top left corner.

Insert Figure 2 about here

Again using linear interpolation, average JNDs in this 2IFC experiment were estimated to be –5% for accelerated tempo (SE 0.5%), and +6% for decelerated tempo (SE 1.0%).

The results show that d' values are again higher for accelerated speech than for decelerated speech. This was confirmed in a two-way repeated measures ANOVA of d' values, with direction-of-change and amount-of-change as fixed factors. The main effect of direction-of-change was indeed significant [$F(1, 21) = 5.9; p = .025$], as was the effect of amount-of-change [$F(6, 16) = 93.5; p < .001$]. Their interaction effect was also significant [$F(6, 16) = 3.2; p = .029$]; the difference between accelerated and decelerated speech increases with larger amounts of change. As before, the higher d' values yield a somewhat smaller JND for accelerated speech than for decelerated speech. Due to larger variability between listeners in the present 2IFC experiment, however, this difference in individual listeners' JND values was not significant. The effect of direction-of-change in a one-way repeated measures ANOVA of listeners' individual JND values yielded $F(1, 21) = 1.3; p = .263$. The effect size is about .36, between small and medium (Cohen, 1988).

In conclusion, the latter two experiments combined yield an estimated just-noticeable difference in speech tempo of about 5% overall (–4% for accelerations and +6% for decelerations). One could argue, however, that listeners in the above experiments did not listen for differences in speech tempo, but that they based their responses on acoustic artefacts of the tempo manipulations. The manipulated versions sounded highly natural, but artefacts are nevertheless possible. One such artefact could be that if a [ba] fragment is slowed down, for example, the resulting formant transitions may be heard as [wa] (Miller, O'Rourke, & Volaitis, 1997, for American English). Another possible artefact is that the vocalic (V) and consonantal (C) portions in the manipulated speech are heard to be unnatural, because the manipulation was done uniformly over V and C portions, whereas speakers vary V portions more than C portions (Nooteboom, 1972).

In order to test these hypotheses, a control experiment was added, in which listeners scaled a stimulus fragment according to its perceived naturalness (cf. Honing, 2006, Experiment 2). If the manipulated stimuli contain phonetic artefacts like those described above, then the manipulations will be easy to identify, and result in lower naturalness ratings. If the PSOLA time manipulation leaves no such artefacts, however, then listeners' naturalness ratings will be roughly equal across the manipulated versions. Hence, the ratings in a simple scaling experiment will tell us about the discriminability of the PSOLA time manipulations in the stimulus materials.

Experiment 3: Scaling

Method

Stimulus materials in this scaling experiment were selected from those used in Experiments 1 and 2. All 20 fragments were included. If all $7+1+7$ relative durations were presented, however, then listeners might be able to identify the original, just by zooming in to the center of the tempo range. This possible strategy was prevented by limiting the number of presentations: Each text passage was only presented at relative durations of 0.87, 0.93, 1.00, 1.07 and 1.13. These values sample the range of relative durations, while keeping down the number of presentations (from 15 to 5 for each passage). While performing their scaling task, listeners might also develop an "absolute ear" for the 5 selected tempi, and learn to zoom in onto this cue during the scaling experiment. This possible strategy was countered by including the 4 practice passages used in Experiments 2 and 3, at relative durations of .91, 1.00 and 1.09. The $(5 \times 20) + (3 \times 4)$ stimuli were presented once, in random order.

Listeners were 24 native Dutch-speaking students at Utrecht University. None of them had participated in the other experiments reported above. Listeners were tested individually in a sound-proof cabin. Their task was to rate the perceived naturalness of each stimulus on a 7-point scale, by pressing a button on the computer screen. There was no time limit, but they were instructed to respond quickly. The experiment was self-paced. Total time of each session was about 15 minutes.

Results and discussion

For each stimulus item, its average rating of naturalness is plotted in Figure 3. These ratings were analyzed by means of repeated measures ANOVAs, by listeners and by stimuli, respectively (Clark, 1973). Relative duration was included as a within-listener and within-stimulus factor; the four speakers were included as a within-listener, between-stimulus factor.

Insert Figure 3 about here

These results clearly indicate that stimuli were rated differently across the manipulated versions. Indeed, the main effect of relative duration turned out to be significant [$F_1(4, 20) = 16.9, p < .001$; $F_2(4, 13) = 111, p < .001$; $minF'(4, 17) = 14.7, p < .001$]. The main effect of speaker was not significant [$F_1(3, 21) = 3.8, p = .025$; $F_2(3, 16) = 6.7, p = .004$; $minF'(3, 31) = 2.4, p = .084$], nor was the interaction between relative duration and speaker [$F_1(12, 12) = 1.1, n.s.$; $F_2(12, 45) = 1.7, n.s.$]. Posthoc tests showed that ratings were significantly different for each relative duration or manipulated version of the stimuli, with the exception of those at 0.87 and 0.93, which yielded similar ratings (nominal $\alpha = .05$, with Bonferroni adjustement).

Although naturalness ratings differ among original and manipulated versions, the results do not suggest that the manipulated versions contain artificial cues for tempo detection. First, highest ratings were observed not for the original but for the accelerated versions, in which the relative durations had indeed been manipulated; this is incompatible with an artificial-cues explanation. Second, listeners in this control experiment did not report such artefacts during debriefing. Instead, their remarks suggest an alternative explanation for their high ratings of accelerated versions, viz. the relatively slow baseline tempo in the original recordings. The average syllable duration (henceforth ASD) in the 20 original versions was 0.239 s (standard deviation 0.077; see Experiment 1 above). This ASD was compared with a large sample ($N = 27516$ inter-pausal chunks, after excluding chunks of 1 or 2 syllables), spoken by 80 speakers of Dutch, for which an average syllable duration of 0.194 s was observed (for details see Quené, 2005). The ASDs in the 20 original fragments in the present study were significantly longer than the large-sample reference [$t(19) = 2.7, p = .015$]. Hence, the baseline tempo in the 20 original stimuli is indeed slower than normal, by about 25%.

For accelerated versions at .93 and .87, the resulting ASDs were not significantly different from the above reference value [$t(19) = 1.8, p = 0.084$ and $t(19) = 1.0, n.s.$, respectively]. The natural or normal tempo of these accelerated versions matches with the high naturalness ratings observed here. For decelerated versions, the slow-baseline original versions were slowed down even further during manipulation, to unnaturally slow rates, which yielded progressively lower ratings for stimuli at the original and even slower tempi.

Third, if listeners' tempo discriminations were based on audible artefacts, then one would predict a negative correlation between the naturalness ratings (Experiment 3) and discrimination d' values (Experiment 2) of each item. Less natural stimuli, containing stronger artificial cues, would be easier to discriminate. Scattergrams of these outcomes did not show any systematic relation. Computed correlations were low, positive-signed, and not significant, for all relative durations.

This absence of negative correlation is not compatible with the artificial-cues explanation.

Finally, one could persist that the unnatural tempo in the decelerated versions constituted an audible artefact in itself, which may have inflated listeners' tempo discrimination. But the discrimination results contradict this view. Resulting tempi are less natural in decelerated stimuli than in accelerated ones (providing stronger artefacts in the decelerated versions); nevertheless, listeners' discrimination is worse for decelerated versions than for accelerated ones. In summary, these findings indicate that listeners' discrimination of tempo-manipulated speech stimuli was not affected by acoustic-phonetic artefacts of the manipulation procedure to construct these stimuli.

General Discussion

The aim of this study was to quantify the just-noticeable difference (JND) for tempo in speech. The comparison experiments 1 and 2 provide an estimated JND of about 5% of the base tempo of a speech utterance. Tempo variations exceeding this difference limen are likely to be noticeable, and relevant in speech communication.

Ideally, all pairs of stimulus versions would have to be presented in a comparison study, so that each manipulated tempo was compared with all other tempi. This would have yielded thousands of presentations, however, with each session lasting many days. For practical purposes, the comparison experiments used a single baseline tempo, viz. the original version of the speech stimuli, with which the manipulated versions were compared. This baseline tempo turned out to be slower than normal speech tempo (as discussed above), which might threaten the validity of the JND estimates. Are JNDs similar for normal-tempo speech, as compared to the slow speech in the present study? Discriminability results for musical tempo show that JNDs for faster tempi are smaller (better) than for slower tempi (Drake & Botte, 1993). Let us assume that this also applies to speech tempo. Then the present JNDs are in fact overestimations, and the true JND for normal-tempo speech is probably smaller (better) than the present estimate.

Tempo variations within speakers tend to exceed the 5% JND. Several studies indicate that within-speaker variations are well above the JND theshold, and hence that these variations are potentially relevant for speech communication. For example, Nooteboom and Eefting (1994) reported on a highly trained professional speaker, whose ASD ranged between 118 and 289 ms per phrase (with mean 176, sd 35); this amounts to a supraliminal change of (or between) –33% and +64%. Chafe (2002) reported on a spontaneous conversation, in which one speaker accelerated by 33% to convey her high emotional involvement (her ASD decreased from 150 to 100 ms).

Such tempo variation is indeed communicatively relevant, as exemplified by two studies. First, Biemans (2000, Chapter 9) investigated the effects of speech tempo on perceived gender of several speakers of Dutch. Tempo was manipulated up or down, by 0.7 or 1.4 syll/s, again by means of PSOLA. Listeners' task was to rate each manipulated version on two 7-point scales (in separate blocks), for the attributes 'masculine' and 'feminine'. Her results showed a weak but significant effect of tempo. Significantly different ratings were observed when the manipulated tempo differed by 1.4 syll/s or more, with faster manipulated tempi associated with perceived masculinity. Thus changes of about 20% in ASD (baseline tempo was 6.1

syll/s) yield a significant difference in perceived gender of the speaker.

Second, Megehee, Dobie, and Grant (2003) investigated the effects of speaking rate on young adults' responses to a spoken advertisement. Relative to the original version at normal rate (100%), the accelerated version (time-compressed by –15%, to 85% of the original duration) yielded a more positive attitude to the speaker ("trustworthy, secure, favorable", etc.), and a higher number of affective responses to an open-ended question about the advertised product. The decelerated version (time-expanded by +15%, to 115%) yielded a more positive attitude to the message, and a higher number of "cognitive" responses to the open-ended question.

These considerable changes in tempo do not pose great problems for a listener. Even if speech is compressed to 65% of its original duration (by –35%), it is still reported to be "perfectly intelligible" (Janse, 2004, p.160). And if speech is time-compressed even further, to 35% of its original duration (by –65%), which amounts to an unnaturally fast speaking rate, even then intelligibility does not fall below 53% correct identifications for real words (Janse, Nooteboom, & Quené, 2003).

Considerably smaller tempo variations, however, may also be noticeable under the appropriate circumstances. Eefting and Nooteboom (1993) reported that one professional speaker produced a 4% change in tempo, due to newness of the relevant phrase, after controlling for sentence accent. This amount of change equals the JND observed by Eefting and Rietveld (1989), for the same professional speaker. As mentioned in the Introduction, the small JND observed may have been due to listeners' adaptation to the single stimulus utterance used in their experiment.

Such adaptation was not possible, however, in the present comparison experiments, which have better ecological validity than previous studies into discriminability of speech tempo. The number of different stimulus fragments is higher, stimuli are from multiple speakers, stimuli vary in number of phrases and in pause distribution. These various types of natural variation probably prevented listeners from adapting to minute differences in speech tempo. Nevertheless, the previously reported JND value of 4% (Eefting & Rietveld, 1989) is similar to the estimates obtained in the present study. This suggests that the former estimate may indeed have been valid (and only mildly affected by adaptation) after all. However, the better ecological validity in the present study strengthens the robustness of the observed JND value of about 5%.

One interesting finding is that present JNDs for speech tempo are approximately equal to those reported for musical tempo (between 4% and 8%). On the one hand, this apparent similarity suggests that tempo discriminability may be similar for speech and for music. If so, then expressive timing variations could play a similar role in listening to speech and in listening to music. Speakers vary the timing of an utterance, depending on its linguistic context (Nooteboom & Eefting, 1994). Similarly, pianists vary the timing of a melody, depending on its musical context (Timmers, Ashley, Desain, & Heijink, 2000). Hence, expressive timing may convey both the position of that utterance or melody in the larger context (phrase structure), and its emphasis or accent, in music as well as in speech. If tempo discriminability is indeed similar, then the perceptual use of such timing variations could also be similar in speech and in music. On the other hand, the observed similarity may have been accidental, as different tasks and techniques were used in musical tempo studies (cf. Introduction) and in this study. Further research is necessary, in which comparable stimuli and similar techniques are used in speech and in music. The present similarity in tempo discriminability would have to hold up under more meticulous experimentation before any conclusions are warranted about similar tempo perception in speech and in music.

In conclusion, this study suggests that the just noticeable difference for tempo in speech is about 5%. Tempo variations within speakers typically exceed this JND, which adds to the importance of these tempo variations in speech communication.

References

Allen, G. D. (1972). The location of rhythmic stress beats in English: An experimental study. Parts I and II. Language and Speech, 15, 72–100 and 179–195.

Benguerel, A.-P., & D'Arcy, J. (1986). Time-warping and the perception of rhythm in speech. Journal of Phonetics, 14(2), 231–246.

Biemans, M. (2000). Gender variation in voice quality. PhD thesis, Katholieke Universiteit Nijmegen.

Chafe, W. (2002). Prosody and emotion in a sample of real speech. In P. H. Fries, M. Cummings, D. Lockwood, & W. Spruiell (Eds.), Relations and functions within and around language (pp. 277–315). London: Continuum.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior, 12, 335–359.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Den Os, E. A. (1985). Perception of speech rate of Dutch and Italian utterances. Phonetica, 42, 124–134.

Drake, C., & Botte, M. C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. Perception & Psychophysics, 54(3), 277–286.

Eefting, W., & Nooteboom, S. (1993). Accentuation, information value and word duration: effects on speech production, naturalness and sentence processing.

In V. Van Heuven & L. C. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality text-to-speech generation* (pp. 225–240). Berlin: Mouton de Gruyter.

Eefting, W., & Rietveld, A. (1989). Just noticeable differences of articulation rate at sentence level. *Speech Communication*, *8*, 355–361.

Ellis, M. C. (1991). Thresholds for detecting tempo change. *Psychology of Music*, *19*(1), 164–169.

Honing, H. (2006). Evidence for tempo-specific timing in music using a web-based experimental setup. *J. Experimental Psychology: Human Perception and Performance*, *32*(3), 780–786.

Humes, J. C. (2002). *Speak Like Churchill, Stand Like Lincoln: 21 Powerful secrets of history's greatest speakers.* New York: Three Rivers Press.

Janse, E. (2004). Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. *Speech Communication*, *42*(2), 155–173.

Janse, E., Nooteboom, S., & Quené, H. (2003). Word-level intelligibility of time-compressed speech: prosodic and segmental factors. *Speech Communication*, *41*(2-3), 287–301.

Levitin, D. J., & Cook, P. R. (1996). Memory for musical tempo: Additional evidence that auditory memory is absolute. *Perception & Psychophysics*, *58*(6), 927–935.

Lindblom, B. (1989). Explaining phonetic variation: A sketch of the H&H

theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge: Cambridge University Press.

Macmillan, N. A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, *84*(5), 452–471.

Madison, G. (2004). Detection of linear temporal drift in sound sequences: empirical data and modelling principles. *Acta Psychologica*, *117*(1), 95–118.

Megehee, C. M., Dobie, K., & Grant, J. (2003). Time versus pause manipulation in communications directed to the young adult population: Does it matter? *Journal of Advertising Research*, *43*(3), 281–292.

Miller, J. L., O'Rourke, T. B., & Volaitis, L. E. (1997). Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, *54*(3-4), 121–137.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Moulines, E., & Verhelst, W. (1995). Time-domain and frequency-domain techniques for prosodic modification of speech. In W. Kleijn & K. Paliwal (Eds.), *Speech coding and synthesis* (pp. 519–555). Amsterdam: Elsevier.

Nooteboom, S. (1972). *Production and perception of vowel duration: A study of durational properties of vowels in Dutch.* PhD thesis, Rijksuniversiteit Utrecht.

Nooteboom, S., & Cohen, A. (1984). *Spreken en verstaan: Een nieuwe inleiding tot de experimentele fonetiek* (2nd ed.). Assen: Van Gorcum.

Nooteboom, S., & Eefting, W. (1994). Evidence for the adaptive nature of speech on the phrase level and below. *Phonetica*, *51*(1–3), 92–98.

Perron, M. (1994). *Checking tempo stability of midi sequencers.* (Paper presented at the 97th Convention of the Audio Engineering Society, San Francisco)

Pfeuty, M., Ragot, R., & Pouthas, V. (2003). Processes involved in tempo perception: A CNV analysis. *Psychophysiology*, *40*(1), 69–76.

Pouliot, M., & Grondin, S. (2005). A response-time approach for estimating sensitivity to auditory tempo changes. *Music Perception*, *22*(3), 389–399.

Quené, H. (2005). Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In *Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech), 4–8 September* (pp. 2457–2460). Lisbon, Portugal.

Timmers, R., Ashley, R., Desain, P., & Heijink, H. (2000). The influence of musical context on tempo rubato. *Journal of New Music Research*, *29*(2), 131–158.

Zwaardemaker, H., & Eijkman, L. (1928). *Leerboek der Phonetiek: inzonderheid met betrekking tot het Standaard-Nederlandsch.* Haarlem: Erven F. Bohn.

Author Note

Hugo Quené, Utrecht institute of Linguistics OTS, Utrecht University.

Preliminary versions of this work were presented at the 147th Meeting of the Acoustical Society of America, New York, NY, 24–28 May 2004, in a Festschrift On Speech and Language: Studies for Sieb G. Nooteboom (Utrecht: Netherlands Graduate School of Linguistics LOT, 2004), at a meeting of the Nederlandse Vereniging voor Fonetische Wetenschappen, Utrecht, The Netherlands, 23 December 2004, and at the 10th Workshop on Rhythm Perception and Production, in Alden Biesen, Belgium, 2–6 July 2005. My sincere thanks are due to Grace Postma, Petra van de Ree and Esther Janse for running experiments, to Vincent van Heuven, Sieb Nooteboom, Carolyn Drake and Henkjan Honing for helpful comments and suggestions, to Theo Veenker for technical assistance, and to Bert Schouten for advice on data analysis.

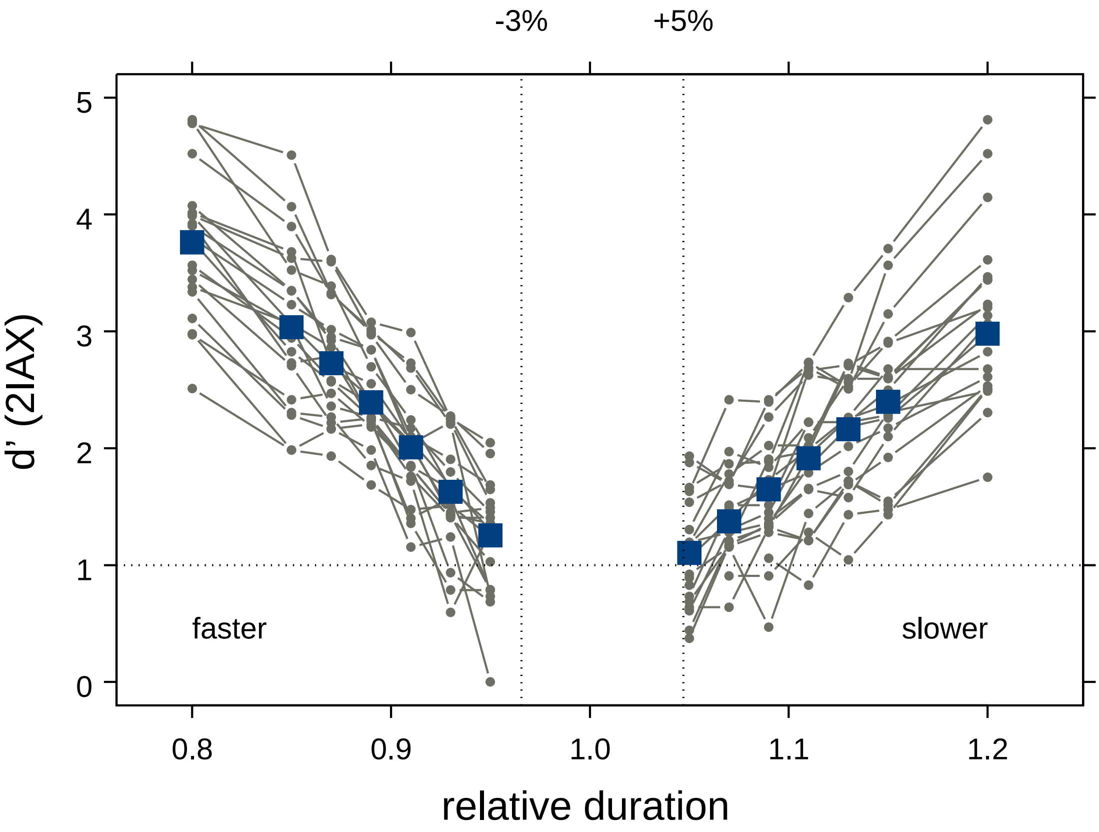
Correspondence concerning this article should be addressed to: Hugo Quené, Utrecht institute for Linguistics OTS, Trans 10, NL-3512 JK Utrecht, The Netherlands, e-mail hugo.quene@let.uu.nl.

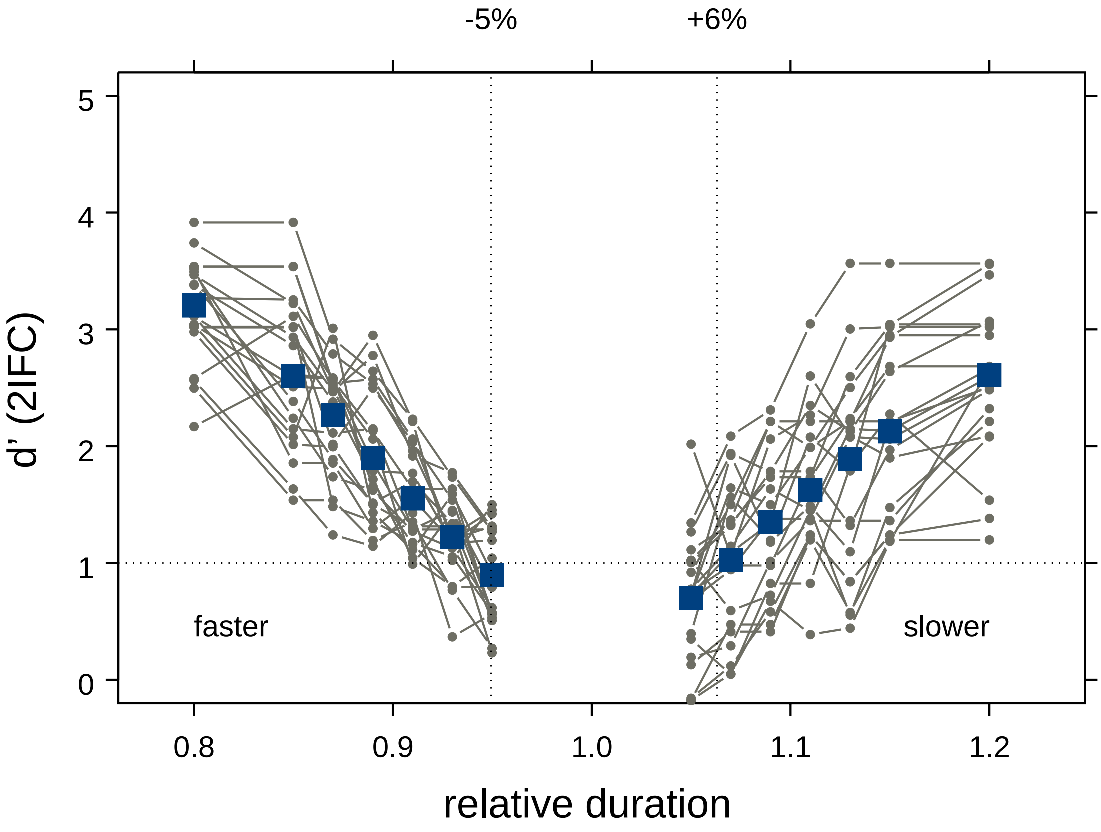
Figure Captions

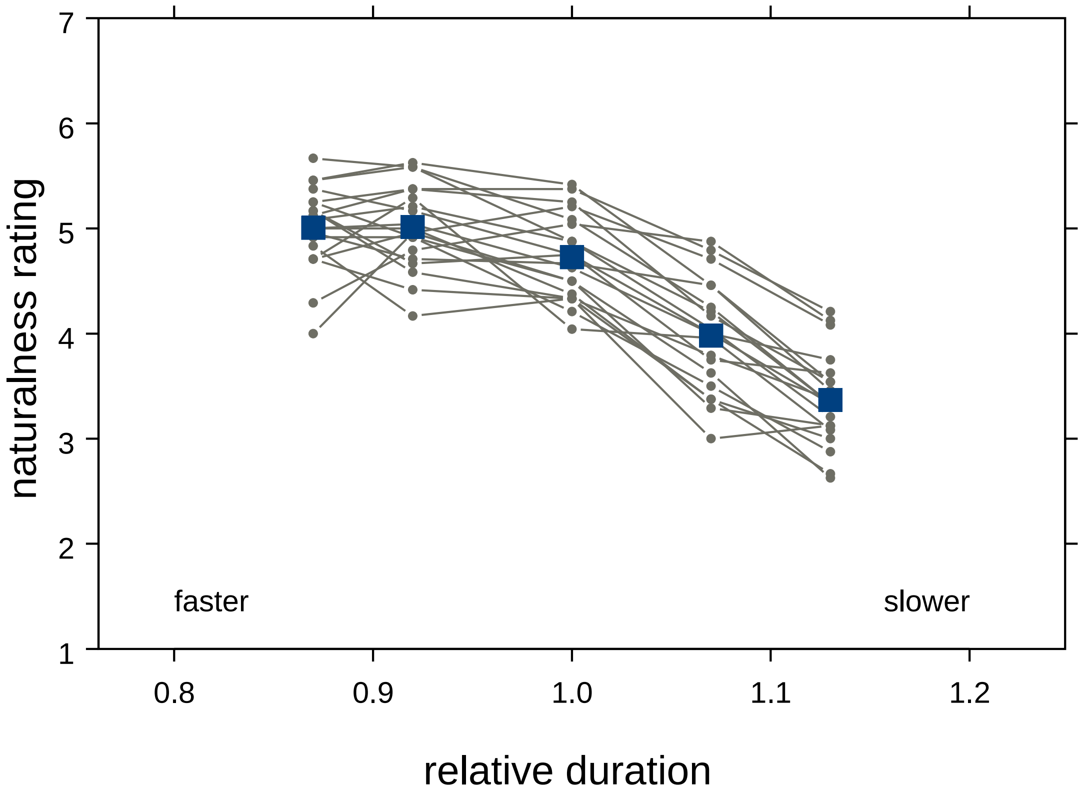
Figure 1. Individual listeners' (dots) and average (squares) values of d', as a function of relative duration or tempo, in the 2IAX experiment.

Figure 2. Individual listeners' (dots) and average (squares) values of d', as a function of relative duration or tempo, in the 2IFC experiment.

Figure 3. Individual stimulus (dots) and average (squares) naturalness ratings over 24 listeners, as a function of relative duration or tempo, in the scaling experiment.


-3%
+5%
d' (2IAX)
5
4
3
2
1
0
faster
slower
0.8
0.9
1.0
1.1
1.2
relative duration


-5%
+6%
d' (2IFC)
0
1
2
3
4
5
faster
slower
0.8
0.9
1.0
1.1
1.2
relative duration



naturalness rating
7
6
5
4
3
2
1
0.8
0.9
1.0
1.1
1.2
relative duration
faster
slower