Effects of the rate of formant-frequency variation on the grouping of formants in speech perception

Robert J. Summers[1], Peter J. Bailey[2], and Brian Roberts[1]

[1]*Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK*

[2]*Department of Psychology, University of York, Heslington, York YO10 5DD, UK*

Running Title: Grouping and rate of formant variation

Version Details: Submitted to JARO on August 19, 2011

Contact details:

(a) Email: b.roberts@aston.ac.uk

Telephone: +44 121 204 3887

Fax: +44 121 204 4090

(b) Email: r.j.summers@aston.ac.uk

(c) Email: peter.bailey@york.ac.uk

Address of corresponding author:

Brian Roberts, PhD

Psychology,

School of Life and Health Sciences,

Aston University,

Birmingham B4 7ET,

United Kingdom

1

ABSTRACT

The perceptual separation of speech from other sounds, particularly other speech, remains poorly understood. Recent research suggests that the ability of an extraneous formant to impair intelligibility depends on the modulation of its frequency, but not its amplitude, contour. This study further examined the effect of formant-frequency variation on intelligibility by manipulating the rate of formant-frequency change. Target sentences were synthetic three-formant (F1+F2+F3) analogues of natural utterances. Perceptual organization was probed by presenting stimuli dichotically (F1+F2C+F3C; F2+F3), where F2C+F3C constitute a competitor for F2 and F3 that listeners must resist to optimize recognition. Competitors were derived from formant-frequency contours extracted from extended passages spoken by the same talker and processed to vary the rate of formant-frequency variation such that rate scale factors relative to target sentences were 0, 0.25, 0.5, 1, 2, and 4 (0=constant frequencies). Competitor amplitude contours were either constant, or time-reversed and rate-adjusted in parallel with the frequency contour. Adding a competitor typically reduced intelligibility; this reduction increased with competitor rate until the rate was twice that of the target sentences. Similarity in the results for the two amplitude conditions confirmed that formant amplitude contours do not influence across-formant grouping. The findings indicate that competitor efficacy is not tuned to the rate of the target sentences; most probably it depends primarily on the overall rate of frequency variation in the competitor formants. This suggests that, when segregating the speech of concurrent talkers, differences in speech rate may not be a significant cue for across-frequency grouping of formants.

Keywords: auditory grouping, speech perception, formant frequency, informational masking

INTRODUCTION

Spoken communication is a fundamental human activity, but it is fairly uncommon in everyday life for us to hear the speech of a single talker in the absence of other background sounds. Hence, our auditory perceptual system is faced with the major challenge of grouping together those sound elements that come from the talker to whom we wish to attend and segregating them from those arising from other sources. This auditory scene analysis problem (Bregman 1990) is particularly challenging if the extraneous sound elements arise from other talkers (see, e.g., Darwin 2008). Arguably, the most important information about speech-sound identity is carried by the formants – spectral prominences associated with the resonant cavities of the vocal tract. Variation in the frequency and amplitude of a formant is an inevitable consequence of change in the size of its associated cavity as the articulators move during speech production (see, e.g., Stevens 1998). Hence, knowledge of formant frequencies and their change over time is of great benefit to listeners trying to understand a spoken message; indeed, there is evidence that this is true even in circumstances where intelligibility is often attributed primarily to temporal cues, such as for noise-vocoded speech (Roberts et al. 2011). Therefore, when more than one talker is speaking at once, choosing and grouping together the right set of formants from the mixture is critical for intelligibility.

Other than in the context of isolated vowels or syllables, relatively little research has focused specifically on across-formant grouping. Moreover, those studies that have explored across-formant grouping using sentence-length utterances have focused on well established grouping cues, such as differences in F0 frequency (e.g., Bird and Darwin 1998; Summers et al. 2010). Hence, little is known about the role of the dynamic properties of formants – particularly their frequency and amplitude contours – in the grouping and segregation of formants. To date, only a handful of studies have focused on the role of formant variation over time in grouping (Remez et al. 1994; Remez 1996, 2001; Roberts et al. 2010), and all of

them were restricted to sine-wave analogues of speech (Bailey et al. 1977; Remez et al. 1981). These studies are considered in more detail below, but in summary they suggest that it is the modulation patterns of the formant-frequency contours that are critical for across-formant grouping. It remains to be established whether the findings of these studies apply in the context of more realistic simulations of speech. The experiment reported here addresses this question and, to our knowledge for the first time, aims specifically to elucidate the role of the *rate* of formant-frequency variation in across-formant grouping.

Given that the factors governing perceptual organization are generally revealed only where competition operates, the experiment reported here used a modification of the second-formant competitor (F2C) paradigm (Remez et al. 1994; Remez 1996, 2001; Roberts et al. 2010; Summers et al. 2010) applied to synthetic-formant analogues of natural sentence-length speech. The crux of the F2C paradigm is the dichotic presentation of two possible candidates for F2, such that intelligibility would be enhanced by the phonetic integration of one version (the true F2) with the other formants but impaired by the integration of the other (F2C). Hence, the listener must resist competition to optimize recognition of the utterance (for a full discussion, see Roberts et al. 2010). The properties of F2C are typically derived from those of the true F2 – e.g., by time reversal of the original formant frequency and amplitude contours. Summers et al. (2010) first used the F2C paradigm in the context of synthetic-formant speech. There was clear evidence that F2C was an effective competitor, but the magnitude of the effect was smaller than that found in the context of sine-wave speech (Roberts et al. 2010), both in terms of its absolute and its relative impact on intelligibility. Here, we have used a two-formant competitor (F2C+F3C), with the intention of increasing the impact of the competitor on intelligibility.

Using sine-wave speech, Remez et al. (1994) first showed that an F2C generated by time-reversing F2 was an effective competitor, but that a pure tone of constant frequency and

amplitude was not. Roberts et al. (2010) used separate manipulations of the frequency and amplitude contours of competitor formants to tease apart their impact on the intelligibility of sine-wave speech. All F2Cs with time-varying frequency contours (whether time reversed or spectrally inverted with respect to F2) were highly effective competitors, regardless of their amplitude characteristics. In contrast, F2Cs with constant frequency contours were entirely ineffective competitors, irrespective of whether the amplitude contour was identical to that of the true F2 or was constant. These results indicate that the frequency contours of formants, but not their amplitude contours, are critical for across-formant grouping. Although Summers et al. (2010) have shown that an F2C with a time-reversed frequency contour can be an effective competitor for synthetic-formant speech, the primary purpose of their study was to examine the effect of F0 differences on formant grouping. Hence, they did not measure the effect of setting either the frequency or amplitude contour of F2C to be constant.

Which aspects of formant-frequency variation in a competitor make it effective? Roberts et al. (2010) suggested that this question could be addressed through the systematic manipulation of the rate of frequency excursions in competitor formants, and measurement of how this affects intelligibility. In the study reported here, synthetic three-formant analogues of target sentences (presented using a standard normalized rate of frequency variation) were accompanied by competitors (F2C+F3C) with time-reversed frequency contours, presented at a set of different relative rates ranging from zero (i.e., constant frequency) up to four times the baseline rate. The upper limit was determined by constraints arising from the synthesis method, discussed below, rather than by the range of formant-frequency variation in natural speech; note that rates of formant movement in natural speech can be fairly high (up to roughly 5000 Hz/s, extrapolating from the data of Tjaden and Weismer 1998). To explore the role of formant amplitude contours in the grouping of synthetic-formant speech, the effect of varying the rate of frequency change in the competitor was measured when the amplitude

contour of the competitor was constant and when it was time reversed. In natural speech, variation in the amount of jaw opening produces correlated changes in formant frequencies and amplitude, and so parallel adjustment of rate for the frequency and amplitude contours is a reasonable approach to take.

At this point, it should be acknowledged that our approach to manipulating the rate of change in the frequency and amplitude contours of competitor formants – in essence, slowing them down or speeding them up – is intentionally simple; we have not attempted to simulate the complex changes in co-articulation and segment reduction associated with changes in natural speech rate. The limitations of our approach are considered further in the Discussion, but it should be noted that changes in the rate of formant frequency change are common concomitants of changes in speech rate (e.g., Pitermann 2000; Wouters and Macon 2002; Weismer and Berry 2003). For example, the data reported by Weismer and Berry (2003) show that, for each of their six talkers, increases in speech rate of CVC syllables resulted in increases in the rate of second formant transitions – mean F2 rate of change (estimated from their Figure 5) increased from 4.9 Hz/ms for the slowest speech rate to 12.4 Hz/ms for the fastest, a fast/slow ratio of 2.5. Although it is true that not all studies have found that changes in speech rate are associated with changes in rate of formant frequency change (Gay 1978; van Son and Pols 1992), the weight of evidence suggests that rate of formant frequency change is a systematic source of variance in speech at different rates, and one that has both perceptual significance for humans (e.g., Divenyi 2009) and relevance to speech recognition by machines (Meyer et al. 2011).

METHODS

Listeners

Volunteers were first tested using a screening audiometer (Interacoustics AS208) to ensure that their audiometric thresholds at 0.5, 1, 2, and 4 kHz did not exceed 20 dB hearing

level. All volunteers who passed the audiometric screening took part in a training session designed to improve the intelligibility of our synthetic-formant speech analogues (see Procedure); all but three of these listeners completed the training successfully and took part in the experiment. 39 listeners (12 males) successfully completed the experiment (mean age = 22.3 years, range = 18.3 – 46.4, SD = 5.3 years). To our knowledge, none of the listeners had heard any of the sentences used in the main part of the experiment in any previous study or assessment of their speech perception. All listeners were native speakers of English and gave informed consent. The research was approved by the Aston University Ethics Committee.

Stimuli and conditions

The stimuli for the experiment were derived from recordings of 78 sentences spoken by a British male talker of "Received Pronunciation" English. The text for the sentences used was provided by Patel and Morse (personal communication) and consisted of variants derived from the original Bamford-Kowal-Bench (BKB) sentence lists (Bench et al. 1979). To enhance the intelligibility of the synthetic analogues, the sentences used were semantically simple and selected to contain 25% or fewer phonemes involving closures or unvoiced frication. A set of keywords was designated for each sentence; these are listed in the Appendix. There is no generally agreed definition of what constitutes a keyword and so the choice is somewhat arbitrary; most designated keywords were content words. The stimuli for the training session were spoken by a different talker; they were derived from 40 sentences taken from commercially available recordings of the IEEE sentence lists (IEEE 1969), but were also selected to contain ≤25% phonemes involving closures or unvoiced frication.

For each sentence, the pitch contour and the frequency contours of the first three formants were estimated from the waveform automatically every 1 ms by Praat (Boersma and Weenink 2008) using a 25-ms-long Gaussian window. In practice, the third-formant contour often corresponded to the fricative formant rather than F3 during phonetic segments with

frication. Gross errors in estimates of formant frequency were hand-corrected using a graphics tablet; amplitude contours corresponding to the corrected formant frequencies were extracted automatically from the spectrograms for each sentence.

Synthetic-formant analogues of each sentence were created using these frequency and amplitude contours to control three parallel second-order resonators whose outputs were summed. The excitation source for the resonators was a periodic train of simple excitation pulses modeled on the glottal waveform, which Rosenberg (1971) has shown to be capable of producing good-quality synthetic speech. The 3-dB bandwidths of the resonators corresponding to F1, F2, and F3 were set to constant values of 50, 70, and 90 Hz, respectively. In the main experiment, the excitation source was monotonous (F0 = 140 Hz) and the speech analogues were presented in a dichotic configuration (left ear = F1; right ear = F2+F3; cf. Rand 1974). Note that this configuration has an advantage over that used by Remez et al. (1994), in that the competitor formants (F2C+F3C) can be added to the left-ear input without risk of appreciable energetic masking of any of the true formants (Roberts et al. 2010; Summers et al. 2010). A schematic stimulus illustrating this configuration is shown in Fig. 1. Diotic presentation was used in the training session, for which each stimulus was generated using the pitch contour extracted from the original recording.

We computed the number of syllables per second for each target sentence in the main experiment by counting the total number of syllables for that sentence and dividing the count by the duration of the sentence. The mean number of syllables/s for the set of target sentences was 3.71; for each sentence, the duration of the set of formant contours was rescaled such that all synthetic analogues were normalized at this mean rate. The properties of the formant contours used for the competitors were defined relative to the normalized rate. Two-formant competitors (F2C+F3C) with a wide range of rate properties were derived from recordings of almost continuously voiced sentences (Binns and Culling 2007; Bird and Darwin 1998;

Stubbs and Summerfield 1990), spoken by the same talker. These sentences were spoken consecutively in clusters of three to six without break, to give 24 long and continuous samples from which competitors with rates considerably higher than baseline could be generated without the need to splice formant tracks together. The frequency and amplitude contours of F2 and F3 were first extracted from these samples, as described above for the target sentences, and then time reversed. At baseline, these time-reversed formant tracks were normalized to the same mean rate as for the target sentences (3.71 syllables/s); note that the mean rate for the sentences with almost continuous voicing, as spoken, was quite similar (3.57 syllables/s).

We used two-formant competitors (F2C+F3C) with the aim of enhancing competitor efficacy; the efficacy of single-formant competitors can be quite limited in the context of synthetic-formant speech compared with sine-wave speech (cf. Roberts et al. 2010; Summers et al. 2010). For each sentence in the main experiment, competitors were generated at various rates relative to baseline. Changes in rate for the two formant-frequency contours were always made in parallel; the rates used relative to baseline were 0 (= constant frequency), 0.25, 0.5, 1, 2, and 4. The upper limit was chosen on the basis of pilot work; the limitations of parallel synthesis using second-order resonators began to introduce audible changes in timbre if the competitor rate was increased beyond 4, presumably because of artifacts associated with rapid changes in resonator center frequency.

For relative rates from 0.25 to 4, competitor parameters were created by: i) selecting at random one of the 24 available pairs of extracted and time-reversed formant tracks; ii) selecting at random a start point for splicing out a segment of the pair of formant tracks of the required length; iii) splicing out a segment corresponding to the required duration, as determined by the duration of the target sentence and the desired competitor rate relative to baseline; iv) rescaling the duration of the spliced segment to match that of the target sentence,

9

and hence to obtain the desired relative rate. For a relative rate of 0, the same steps were performed as for a rate of 0.25, but each of the two frequency contours was set to a constant value at the geometric mean frequency for the appropriate formant in the spliced segment. For convenience, rate relative to baseline is hereafter referred to as *competitor rate*.

In the reversed-amplitude conditions (see below), competitor formants were generated using time-reversed amplitude contours scaled to the desired competitor rate, in parallel with the time-reversed frequency contours. The constant-amplitude conditions differed only in that the amplitude contour for each competitor formant was set to a constant value corresponding to the root mean square power of the time-reversed contour for that formant. Note that there is no distinction between reversed and constant amplitude when the competitor rate is set to zero; hence only one condition is needed in this case. Competitors were generated by parallel synthesis using the same 3-dB bandwidths as for F2 and F3 in the target sentences (i.e., 70 Hz and 90 Hz, respectively). The excitation source and F0 used (140 Hz) were also the same as for the target sentences; note that the waveform of the excitation source for the competitor formants was not time reversed, unlike their frequency and amplitude contours. An example set of competitors, synthesized at different relative rates, is illustrated for the reversed- and constant-amplitude conditions in Fig. 2 and Fig. 3, respectively. Stimuli were selected such that the center frequency of F2C was always ≥80 Hz from the center frequency of the true F1, so formant tracks did not cross, nor did adjacent formants approach close enough to cause audible interactions. For each condition, a new set of competitors was prepared for each sentence, as a precaution against effects peculiar to particular excised segments.

There were 13 conditions in the experiment (see Table 1). One condition (C1) was a control for which a competitor (F2C+F3C) was present but the true F2 and F3 were absent; the competitor rate was set to 1 and the competitor amplitude contour was time reversed. We judged this single control condition to be sufficient, because there is no reason to expect an

increase in intelligibility were the competitor rate to be changed from 1 or the amplitude contour set to constant. Eleven conditions (C2-C12) were experimental, for which the stimuli contained the true F2 and F3 plus a competitor on one of the six rates specified above; the competitor amplitude contour was either constant or time reversed. The final condition (C13) was the dichotic reference case, for which the true F2 and F3 were present but the competitor was absent. For each listener, the sentences were divided equally across conditions (i.e., six per condition) using an allocation that was counterbalanced by rotation across each set of 13 listeners tested. Before the main experiment, each listener was presented with all 40 training sentences; no competitor formants were included in these stimuli.

All speech analogues were synthesized using MITSYN (Henke 2005) at a sample rate of 22.05 kHz and with 10-ms raised-cosine onset and offset ramps. They were played at 16-bit resolution over Sennheiser HD 480-13II earphones via a sound card, programmable attenuators (Tucker-Davis Technologies PA5), and a headphone buffer (TDT HB7). Output levels were calibrated using a sound-level meter (Brüel and Kjaer, type 2209) coupled to the earphones by an artificial ear (type 4153). Stimuli were presented at a reference level (long term average) of 75 dB SPL; this describes the case when the left ear receives F1 (the most intense formant). F1 was presented at the reference level in all conditions; hence there was some variation in the level ($\approx$2 dB) and loudness of the stimuli across conditions, depending on the presence or absence of F2, F3, and the competitor (F2C+F3C). In the training session, which used diotic stimuli presented at each utterance's natural rate, both the original recordings (44.1 kHz sample rate) and the speech analogues were presented at 72 dB SPL.

Procedure

Stimuli were always presented such that F1 was heard in the left ear and F2+F3 were heard in the right. In previous studies, we have demonstrated that there are no appreciable ear-dominance effects for sentence-length utterances in the context of the closely related

dichotic F2C paradigm (Roberts et al. 2010; Summers et al. 2010). Therefore, we did not counterbalance for ear of presentation in the current experiment.

During testing, listeners were seated in front of a computer screen and a keyboard in a sound-attenuating chamber (Industrial Acoustics 1201A). The study consisted of a training session followed by the main experiment and typically took about an hour and a half to complete; listeners were free to take a break whenever they wished. In both phases of the study, stimuli were presented in a new quasi-random order for each listener.

There were 40 trials in total for the training session. On each of the first ten trials, participants heard the synthetic version (degraded, D) and the original recording (clear, C) of a given sentence in the order DCDCD; no response was required but participants were asked to listen to these sequences carefully. On each of the remaining 30 trials, participants first heard the synthetic version of a given sentence, which they were asked to transcribe. They were allowed to listen to the stimulus up to a maximum of six times before typing in their transcription. After each transcription was entered, feedback to the listener was provided by playing the original recording followed by a repeat of the synthetic version. Davis et al. (2005) found this DCD strategy to be an efficient way of enhancing the perceptual learning of speech-like stimuli with unusual surface structures.

We set a mean criterion of ≥50% keywords correct across the training trials for inclusion in the main experiment. As for the training, participants were able to listen to each stimulus up to six times without time limit before typing in their transcription. However, they did not receive feedback of any kind on their responses in the main experiment.

Data analysis

For each listener, the intelligibility of each sentence was quantified in terms of the percentage of keywords identified correctly; homonyms were accepted. The stimuli for each condition comprised six sentences. Given the variable number of keywords per sentence (2–

4), the mean score for each listener in each condition was computed as the percentage of keywords reported correctly giving equal weight to all the keywords used (always 18 or 19 per set of six sentences). Following the procedure of Roberts et al. (2010), we classified responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly (see Foster et al. 1993).

RESULTS

Figure 4 shows the mean percentage scores (and inter-subject standard errors) across conditions in terms of keywords identified correctly. The results for the two types of competitor amplitude contour are shown by separate curves, one for the constant-amplitude case (filled circles, solid line) and one for the reversed-amplitude case (open circles, dashed line). The results for the control condition (C1) and the dichotic-reference condition (C13) are shown using asterisk symbols on the left- and right-hand sides of the figure, respectively. Note that intelligibility was near-floor in C1, for which the true F2 and F3 were absent, and high in C13, for which competitor formants were absent. A one-factor within-subjects analysis of variance (ANOVA) for the whole dataset showed a highly significant effect of condition on intelligibility [$F_{(12,456)}=74.385$, $p<0.001$]. Paired-samples comparisons (two-tailed) were computed using the restricted least-significant-difference test (Keppel, 1991). With the sole exception of the case where competitor rate was 0.25 x baseline and the competitor amplitude contour was time reversed ($p=0.063$), the addition of a two-formant competitor (F2C+F3C) always reduced recognition performance significantly with respect to the dichotic reference condition (range: $p=0.014 – p<0.001$).

To assess the effects of competitor rate and amplitude contour, a two-factor ANOVA restricted to the eleven experimental conditions (C2-C12)[1] was performed. This showed that the main effect of competitor rate was highly significant [$F_{(5,190)}=5.131$, $p<0.001$], but that neither the main effect of competitor amplitude contour [$F_{(1,38)}=0.030$, $p=0.863$] nor the

interaction of the two factors [$F_{(5,190)}= 0.263$, $p=0.933$] was significant. Compared with the dichotic reference case, the reduction in keyword score associated with a competitor rate of zero (7.4 percentage points) more than doubled when the competitor rate was increased to twice baseline (16.4 percentage points, when collapsed across amplitude contour). Overall, the results indicate that the effect of the competitor on intelligibility is not sharply tuned to the rate of formant-frequency variation in the competitor, relative to that for the target speech. Whether this relationship is better described as broadly tuned around a competitor rate of two or as one in which efficacy rises gradually and progressively with increasing competitor rate depends on the interpretation of the results for a competitor rate of four times baseline. This issue is considered in the Discussion, in relation to synthesis constraints, but note that the rise in mean keyword score when competitor rate was increased from two to four was significant for the reversed-amplitude case ($p=0.007$) but not for the constant-amplitude case ($p=0.331$).

DISCUSSION

Sentence intelligibility is typically reduced when the target speech is accompanied by a two-formant competitor (F2C+F3C), created using rate-adjusted versions of the time-reversed frequency contours of F2 and F3 extracted from different passages spoken by the same talker. The impact of the competitor on keyword recognition cannot be explained in terms of energetic masking, because the F1 of the target sentence was lower in frequency and more intense than the pair of competitor formants presented in the same ear (cf. Roberts et al. 2010; Summers et al. 2010). Overall, the results indicate that competitors tend to become progressively more effective for rates up to at least twice baseline; there is no indication of sharp tuning of the rate function, as one might have expected if differences in speech rate are used as a cue for the appropriate grouping and segregation of formants.

At this point, an issue that merits consideration is the extent to which stimulus artifacts may have influenced the results obtained for the fastest competitor rate tested. As

noted in the Methods, the upper limit of competitor rate was set to four times baseline because the synthesis method introduced audible changes in timbre at higher competitor rates. This effect was more evident when the competitor amplitude contour was time reversed than when it was constant, presumably because both frequency and amplitude contours were changing rapidly in the former case. Given this, and the suggestion in the data that the change in direction of the rate function at four times baseline was greater for the reversed- than for the constant-amplitude condition, we would suggest that stimulus artifacts were beginning to influence the results for the fastest competitor rate tested. Specifically, the timbre differences introduced by such artifacts may have provided additional segregation cues that made it easier for listeners to exclude the competitor when perceptually grouping the formants. Hence, until more reliable data are available, we consider it prudent to set aside the results for the fastest competitor rate condition, and to conclude simply that there is an approximately linear relationship between the rate of variation in competitor formant frequency and the impact of a competitor on intelligibility, at least for rates up to twice baseline.

Acoustic consequences of changes in speech rate

Despite the body of evidence that changes in the rate of formant frequency variation are common concomitants of changes in speech rate (Pitermann 2000; Wouters and Macon 2002; Weismer and Berry 2003), our rate manipulation can be criticized on the grounds that it fails to simulate many of the changes associated with changes in natural speech rate. Increasing or decreasing the rate at which speech is produced – which talkers do frequently in normal discourse for a variety of linguistic and paralinguistic reasons – has complex articulatory (and hence acoustic) consequences in addition to the manifest effect of decreasing or increasing overall utterance duration. Increases in speech rate typically involve, among other things: (i) increased co-articulation, reduction, and assimilation (Gay 1981; Agwuele et al. 2008); (ii) reductions in segment duration that are dependent on stress

15

placement and phoneme class (Lehiste 1970; Byrd and Tan 1996); and (iii) increases in articulator velocity, as well as reorganization of speech motor control strategies (Adams et al. 1993). Moreover, talkers differ in the articulatory strategies that they use to effect changes in speech rate (Tjaden and Weismer 1998; Weismer and Berry 2003). As a result, the acoustic changes associated with changes in rate of speech are non-linear and talker dependent, and difficult to model in detail.

Nonetheless, we argue that our approach to rate manipulation is a reasonable one. Fortunately for those interested in the perceptual effects of speech rate manipulations, it turns out that adequately convincing simulations of speech at different rates can be created by fairly gross, quasi-linear scaling of overall utterance duration and rates of formant frequency change, without the need to model the detailed consequences of rate-dependent articulatory strategies – e.g., by processing natural speech waveforms using synchronized overlap and add algorithms (Roucus and Wilgus 1985), or for synthetic speech by scaling the time-base of the synthesizer. Estimates of the amount of this time compression or expansion that can be applied without compromising intelligibility vary, but for sentence-length utterances intelligibility is typically not substantially reduced for rates between about 50%-200% of a normal speech rate (Korabic et al. 1978; Beasley et al. 1980).

Informational masking of speech

Any effect of the competitor formants on intelligibility in the dichotic procedure we have used could have several possible origins. However, we can rule out an explanation in terms of energetic masking, because F1 of the target sentence was lower in frequency and more intense than the pair of competitor formants presented in the same ear (cf. Roberts et al. 2010; Summers et al. 2010) and F2 and F3 of the target sentence were presented in the other ear (cf. Rand 1974). Therefore, the observed effects of F2C+F3C may be characterized as

involving primarily informational masking (Pollack 1975), acknowledging that applying this label does little to elucidate the processes by which target sentence intelligibility is affected.

Roberts et al. (2010) suggested that competitor formants reduce intelligibility because listeners are unable to reject them completely from the auditory perceptual organization of the sentence, which for optimum intelligibility should include only F1+F2+F3; as a result, elements of the competitor formants are integrated with the other formants, thus changing the phonetic specification of the target sentence and leading to word recognition errors. This has some similarity to the interpretation of an early example of informational masking involving the influence of non-speech formant patterns in one ear on the perceptual interpretation of a speech signal in the other ear, reported by Porter and Whittaker (1980). Identification of a synthetic CV syllable presented to the left ear was systematically influenced by the direction of isolated second formant transitions presented to the right ear. These authors suggested that this was the result of a pre-phonetic process that operates on a central, salience-dependent combination of spectro-temporal auditory cues from both ears.

Most experiments on informational masking of or by speech have focused on the challenge faced when listening to a talker in the presence of one or more concurrent talkers (Kidd et al. 2008). It was demonstrated early on that when responding to the speech of one of two talkers presented dichotically, the contralateral masking talker has little effect (Cherry 1953). This contrasts with the demonstration by Brungart and Simpson (2002) that a contralateral speech masker *does* affect responding to the speech of one of *two* competing talkers presented to the target ear. A similar effect was found when the contralateral masker was time-reversed speech, but not when it was steady noise. Thus, in a sufficiently complex dichotic listening task involving perceptual segregation, as for our competitor procedure, listeners were unable to exclude entirely a spectro-temporally varying contralateral masker – an effect which Kidd et al. (2003) have shown is not specific to speech.

The amount of informational masking observed when a target talker's speech is presented together with a masking talker depends on how similar the two talkers are, but the dependence of informational masking on the signal-to-masker ratio is relatively less than it is for energetic masking. Information masking is maximal when target and masking speech come from the same talker; masking is reduced if the masker is spoken by a different talker, particularly if target and masking talkers differ in gender, and some talkers are more resistant to masking (or more effective maskers) than others (Brungart 2001). It appears that talkers know about speech-on-speech masking constraints, and when in the presence of other talkers they actively monitor the masking potential of competing background speech and adjust their speech patterns accordingly to reduce the risk of being masked (Cooke and Lu 2010).

Effects of mixing speech of different rates together

Of particular relevance for our study is the question of whether the amount of speech-on-speech masking depends on the relative rates of target and masking speech. In particular, is there any evidence that faster interferers are more effective than slower ones at impairing the intelligibility of target speech? Relatively little research has addressed this issue, but we are aware of two germane studies whose findings are broadly consistent with ours. Gordon-Salant and Fitzgibbons (2004) found evidence of rate dependency in the effect of background babble on speech recognition. Specifically, the intelligibility of 50% time-compressed speech (i.e., rate = twice baseline) was more degraded when it was accompanied by 50% time-compressed babble than when it was accompanied by slower rates of babble (25% time-compressed or normal uncompressed babble, which didn't differ in their effect).[2]

Chen et al. (2008) reported two preliminary experiments exploring the effect of manipulating the rate of masking speech on recognition of nonsense sentences spoken by a target talker at a natural rate. When target and masker were co-presented, a procedure presumably involving both energetic and informational masking, masking tended to be

greater for faster than for slower maskers, but showed a local maximum when the rates of target and masker speech were the same. When the precedence effect was used to create a perceived spatial separation between target and masker (Freyman et al. 2001), so that any unmasking could be assumed to reflect reduced informational masking, the results indicated that unmasking increased as masker speech rate was increased over the range 50%-150% of the target rate. This pattern implies that the impact of informational masking on target-speech recognition was maximal for the fastest masker speech rate.

Conclusions

Our results indicate that competitor efficacy is dependent on competitor rate; the impact of a pair of competitor formants on the intelligibility of target sentences increases gradually and progressively as the rate of change of frequency variation in the competitor formants increases, at least for rates up to twice baseline. Although it is not possible to prove beyond doubt that the trend for competitor efficacy to decline at higher rates arose from stimulus artifacts, it is certainly the case that the effect of rate of frequency variation is not sharply tuned. In contrast, rate of change of amplitude variation in the competitor formants had no discernible effect on intelligibility. This has been shown previously for sine-wave speech (Roberts et al. 2010), but the current findings are important because synthetic-formant speech provides a more complete simulation of natural speech than is provided by sine-wave analogues. Our findings are compatible with those of speech-on-speech masking studies, particularly when the effects of energetic masking have been controlled so that the impact on intelligibility of the interferer can be attributed to informational masking (Chen et al. 2008). Overall, our results suggest that differences in speech rate may not be a significant cue for across-frequency grouping of formants when segregating the speech of concurrent talkers.

ACKNOWLEDGMENTS

FOOTNOTES

(1) There is no distinction between reversed and constant amplitude when the competitor rate is set to zero; hence the data for C2 served as the zero-rate case for both amplitude-contour conditions in the two-factor ANOVA.

(2) Gordon-Salant and Fitzgibbons (2004) interpreted this result as preliminary evidence in support of the notion that rate-mismatched babble has less impact than rate-matched babble on the intelligibility of target speech. However, they did not include conditions for which the babble rate was greater than the speech rate. Hence, we argue that their findings are better interpreted as evidence that faster interferers have a greater impact on intelligibility.

APPENDIX

Inventory of the target sentences used in the main part of the experiment; the designated keywords are shown in bold. All items are from the unpublished sentence lists of Patel and Morse (personal communication), which were derived by rearranging items from the original BKB sentence lists (Bench et al., 1979).

1. A **man's turning** the **sign**.
2. The **baby** was **lovely**.
3. The **children** are **running home**.
4. The **farmer feeds animals**.
5. The **milkman comes early**.
6. **They're buying some onions**.
7. The **angry man argued**.
8. The **chair** is **in** the **corner**.
9. The **dinner** was **ready**.
10. **They like yellow pears**.
11. **They're buying bananas**.
12. **They're running near** the **park**.
13. A **mouse ran near** the **door**.
14. She's **moving home**.
15. The **bench** is **near** the **tree**.
16. The **mirror hangs on** the **wall**.
17. The **mother went out**.
18. The **room** was **very clean**.
19. An **old man** was at **home**.
20. The **ambulance** was **noisy**.
21. The **oven door** was **warm**.
22. The **toys** are **on** the **table**.
23. **They called** the **fire engine**.
24. **They're eating dinner**.
25. An **old woman ties** the **string**.
26. The **family** are **laughing**.
27. The **fire engine** is **red**.
28. The **five men** are **dancing**.
29. The **room's very hot**.
30. The **sign hangs above** the **door**.

31. The **eggs** are **in** the **cupboard**.
32. The **farmer** was **funny**.
33. The **flowers grow in** the **ground**.
34. The **man cleaned** the **room**.
35. The **towel** was **very wet**.
36. The **tree fell on** the **lorry**.
37. He's **cleaning** the **van**.
38. The **apples fell on** the **floor**.
39. The **fruit lies on** the **table**.
40. The **girl** has **green eyes**.
41. The **lorry** was **quite long**.
42. The **mother clears** the **room**.
43. **Father drove** the **lorry**.
44. He **drinks milk in** the **morning**.
45. The **cleaner washed** the **windows**.
46. The **marmalade** was **sweet**.
47. The **mud** was **brown**.
48. The **old car's very noisy**.
49. The **bedroom door** was **open**.
50. The **bulls** are **running quickly**.
51. The **man uses** a **spoon**.
52. The **old glove** has a **small hole**.
53. The **strawberries** were **nice**.
54. **They're cycling down** the **hill**.
55. The **broom's near** the **cupboard**.
56. The **cat lies on** the **ground**.
57. The **flowers** are **in water**.
58. The **girl's running on** the **road**.
59. The **gloves** are **green**.
60. The **scissors** are **on** the **chair**.
61. Her **long hair** is **brown**.
62. **Mother's cutting** the **apple**.
63. The **drawer** was **open**.
64. The **thin dog ran along**.
65. The **three boys** are **running**.
66. The **washing machine's full**.
67. **Mother ate** the **oranges**.
68. The **boy knows** the **way**.

69. The **fireman** was **hungry**.
70. The **new house** was **lovely**.
71. The **school had nine rooms**.
72. **They made** a **snowman**.
73. The **baby wears small shoes**.
74. The **cheese pie** was **warm**.
75. The **cleaner** was **angry**.
76. The **clown** was **very funny**.
77. The **police went along**.
78. The **snow lay on** the **window**.

REFERENCES

Adams SG, Weismer G, Kent RD. Speaking rate and speech movement velocity profiles. J. Speech Hear. Res. 36:41-54, 1993.

Agwuele A, Sussman HM, Lindblom B. The effect of speaking rate on consonant vowel coarticulation. Phonetica 65:194-209, 2008.

Bailey PJ, Summerfield Q, Dorman M. On the identification of sine-wave analogues of certain speech sounds. Haskins Lab. Status Rep. Speech Res. SR-51/52:1-25, 1977.

Beasley DS, Bratt GW, Rintelmann WF. Intelligibility of time-compressed sentential stimuli. J. Speech Hear. Res. 23:722-731, 1980.

Bench J, Kowal A, Bamford J. The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. Brit. J. Audiol. 13:108-112, 1979.

Binns C, Culling JF. The role of fundamental frequency contours in the perception of speech against interfering speech. J. Acoust. Soc. Am. 122:1765-1776, 2007.

Bird J, Darwin CJ. Effects of a difference in fundamental frequency in separating two sentences. In: Palmer AR, Rees A, Summerfield AQ, Meddis R (eds) Psychophysical and Physiological Advances in Hearing. Whurr, London, pp. 263-269, 1998.

Boersma P, Weenink D. Praat: doing phonetics by computer, software package, version 5.0.18, 2008. Retrieved from http://www.praat.org/ (Last viewed 7/29/2011).

Bregman AS. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, MA, 1990.

Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. 109:1101-1109, 2001.

Brungart DS, Simpson BD. Within-ear and across-ear interference in a cocktail-party listening task. J. Acoust. Soc. Am. 112:2985-2995, 2002.

Byrd D, Tan CC. Saying consonant clusters quickly. J. Phonet. 24:263-282, 1996.

Chen J, Wu XH, Zou XF, Zhang ZP, Xu LJ, Wang MY, Li L, Chi HS. Effect of speech rate on speech-on-speech masking. J. Acoust. Soc. Am. 123:3713, 2008.

Cherry EC. Some experiments on the recognition of speech, with one and with two ears. J. Acoust. Soc. Am. 25:975-979, 1953.

Cooke M, Lu Y. Spectral and temporal changes to speech produced in the presence of energetic and informational maskers. J. Acoust. Soc. Am. 128: 2059–2069, 2010.

Darwin CJ. Listening to speech in the presence of other sounds. In: Moore BCJ, Tyler LK, Marslen-Wilson W (eds) The Perception of Speech: From Sound to Meaning. Special Issue, Phil. Trans. R. Soc. B. 363:1011-1021, 2008.

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C. Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. Gen. 134:222-241, 2005.

Divenyi P. Perception of complete and incomplete formant transitions in vowels. J. Acoust. Soc. Am. 126:1427-1439, 2009.

Foster JR, Summerfield AQ, Marshall DH, Palmer L, Ball V, Rosen S. Lip-reading the BKB sentence lists: Corrections for list and practice effects. Brit. J. Audiol. 27:233-246, 1993.

Freyman RL, Balakrishnan U, Helfer KS. Spatial release from informational masking in speech recognition. J. Acoust. Soc. Am. 109:2112-2122, 2001.

Gay T. Effect of speaking rate on vowel formant movements. J. Acoust. Soc. Am. 63:223-230, 1978.

Gay T. Mechanism in the control of speech rate. Phonetica 38:148-158, 1981.

Gordon-Salant S, Fitzgibbons PJ. Effects of stimulus and noise rate variability on speech perception by younger and older adults. J. Acoust. Soc. Am. 115:1808-1817, 2004.

Henke WL. MITSYN: A coherent family of high-level languages for time signal processing, software package. Belmont, MA, 2005, e-mail: mitsyn@earthlink.net. Retrieved from http://home.earthlink.net/~mitsyn (Last viewed 7/29/2011).

Institute of Electrical and Electronics Engineers (IEEE). IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. AU-17:225-246, 1969.

Keppel G. Design and Analysis: A Researcher's Handbook. Prentice-Hall, Upper Saddle River, NJ, 1991.

Kidd G, Mason CR, Arbogast TL, Brungart DS, Simpson BD. Informational masking caused by contralateral stimulation. J. Acoust. Soc. Am. 113:1594-1603, 2003.

Kidd G, Mason CR, Richards VM, Gallun FJ, Durlach NI. Informational masking. In: Yost WA, Fay RR (eds) Auditory Perception of Sound Sources (Springer Handbook of Auditory Research, Vol. 29). Springer, Berlin, pp. 143-189, 2008.

Korabic EW, Freeman BA, Church GT. Intelligibility of time-expanded speech with normally hearing and elderly subjects. Audiology 17:159-164, 1978.

Lehiste I. Suprasegmentals. MIT Press, Cambridge, MA, 1970.

Meyer BT, Brand T, Kollmeier B. Effect of speech-intrinsic variations on human and automatic recognition of spoken phonemes. J. Acoust. Soc. Am. 129:388-403, 2011.

Pitermann M. Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. J. Acoust. Soc. Am. 107:3425-3437, 2000.

Pollack I. Auditory informational masking. J. Acoust. Soc. Am. 57:S5, 1975.

Porter RJ, Whittaker RG. Dichotic and monotic masking of CV's by CV second formants with different transition starting values. J. Acoust. Soc. Am. 67:1772-1780, 1980.

Rand TC. Dichotic release from masking for speech. J. Acoust. Soc. Am. 55:678-680, 1974.

Remez RE. Perceptual organization of speech in one and several modalities: Common functions, common resources. In: ICSLP-1996, Philadelphia, PA, pp. 1660-1663, 1996.

Remez RE. The interplay of phonology and perception considered from the perspective of perceptual organization. In: Hume E, Johnson K (eds) The Role of Speech Perception in Phonology. Academic Press, San Diego, pp. 27-52, 2001.

Remez RE, Rubin PE, Berns SM, Pardo JS, Lang JM. On the perceptual organization of speech. Psychol. Rev. 101:129-156, 1994.

Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science 212:947-950, 1981.

Roberts B, Summers RJ, Bailey PJ. The perceptual organization of sine-wave speech under competitive conditions. J. Acoust. Soc. Am. 128:804-817, 2010.

Roberts B, Summers RJ, Bailey PJ. The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes. Proc. Roy. Soc. Lond. B 278:1595-1600, 2011.

Rosenberg AE. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am. 49:583-590, 1971.

Roucus S, Wilgus AM. High quality time-scale modification for speech. IEEE Proceedings on Acoustics, Speech and Signal Processing, pp. 493-496, 1985.

Stevens KN. Acoustic Phonetics. MIT Press, Cambridge, MA, 1998.

Stubbs RJ, Summerfield Q. Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners. J. Acoust. Soc. Am. 87:359-372, 1990.

Summers RJ, Bailey PJ, Roberts B. Effects of differences in fundamental frequency on across-formant grouping in speech perception. J. Acoust. Soc. Am. 128:3667-3677, 2010.

Tjaden K, Weismer G. Speaking-rate-induced variability in F2 trajectories. J. Speech Lang. Hear. Res. 41:976-989, 1998.

van Son RJJH, Pols LCW. Formant movements of Dutch vowels in a text read at normal and fast rate. J. Acoust. Soc. Am. 92:121-127, 1992.

Weismer G, Berry J. Effects of speaking rate on second formant trajectories of selected vocalic nuclei. J. Acoust. Soc. Am. 113:3362-3378, 2003.

Wouters J, Macon MW. Effects of prosodic factors on spectral dynamics. I. Analysis. J. Acoust. Soc. Am. 111:417-427, 2002.

TABLE 1

Stimulus properties for the conditions used in the main session of the experiment.

Condition	Stimulus configuration (left ear; right ear)	Amplitude contour of competitor	Competitor rate (relative to target speech)
1	(F1+F2C+F3C; −)	R	1
2	(F1+F2C+F3C; F2+F3)	C	0
3	(F1+F2C+F3C; F2+F3)	C	0.25
4	(F1+F2C+F3C; F2+F3)	C	0.5
5	(F1+F2C+F3C; F2+F3)	C	1
6	(F1+F2C+F3C; F2+F3)	C	2
7	(F1+F2C+F3C; F2+F3)	C	4
8	(F1+F2C+F3C; F2+F3)	R	0.25
9	(F1+F2C+F3C; F2+F3)	R	0.5
10	(F1+F2C+F3C; F2+F3)	R	1
11	(F1+F2C+F3C; F2+F3)	R	2
12	(F1+F2C+F3C; F2+F3)	R	4
13	(F1; F2+F3)	− −	− −

The amplitude contour of the competitor (F2C+F3C), when present, is either time reversed (R) or constant (C). Competitor rate refers to the rate of variation in formant frequency relative to that for F2 and F3 in the target sentences. Relative rate also applies to the amplitude contour of the competitor, on occasions when it is time reversed (R).
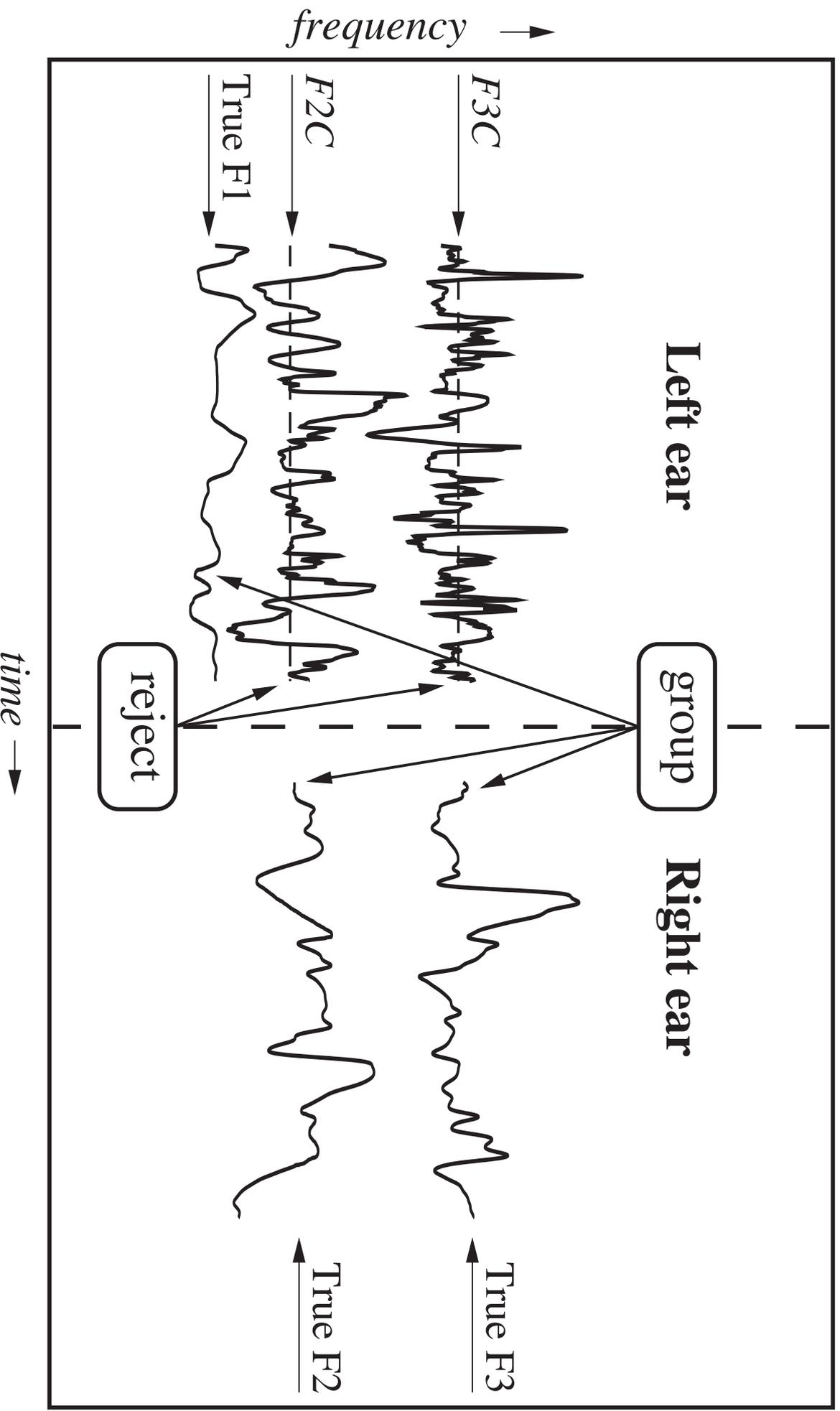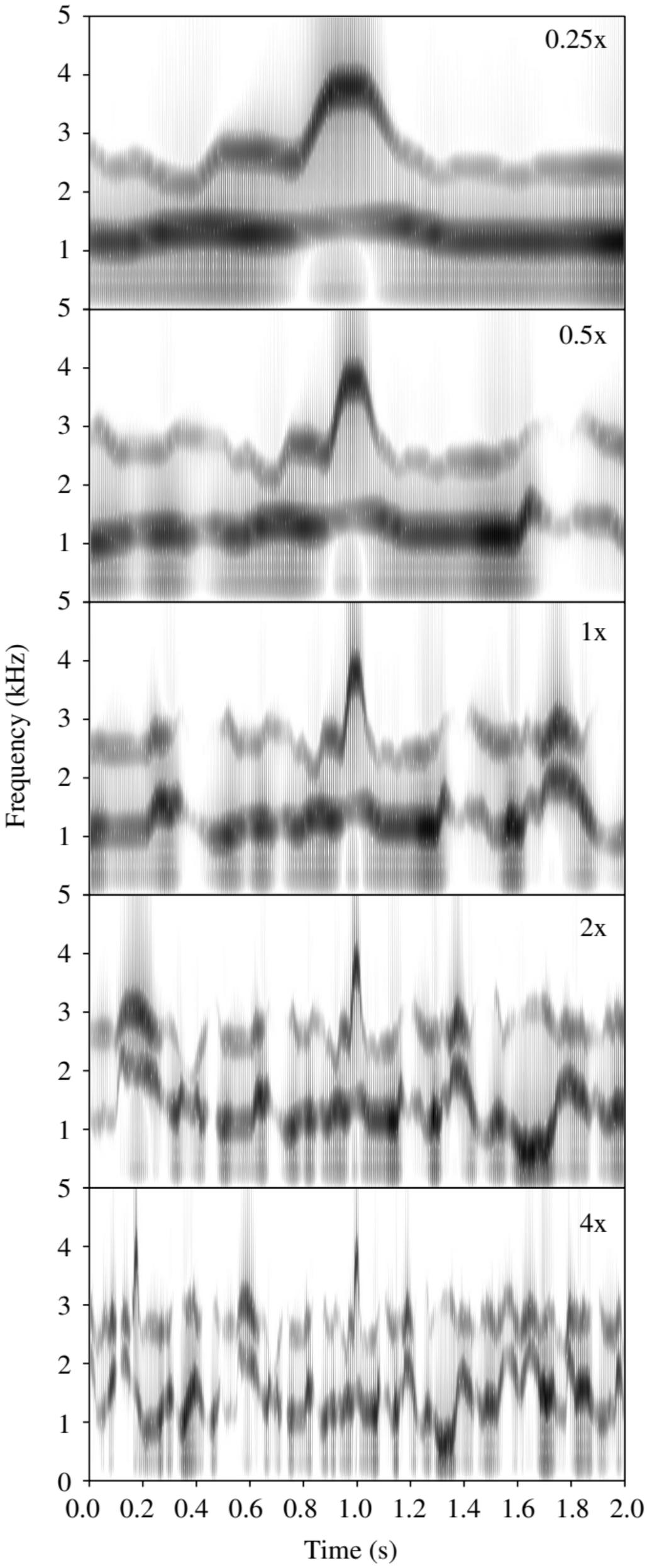
FIGURE CAPTIONS

(1) Stimuli – schematic illustration of the dichotic configuration used in the experiment. The left ear receives F1 of the target sentence; the right ear receives F2 and F3. The competitor formants (F2C+F3C) are presented in the same ear as F1. The rate of frequency variation in the competitor formants can be controlled relative to that for the target sentence. Illustrated here are formant-frequency contours for the cases where the competitor rate is x4 (solid lines in bold) and x0 (dashed lines); in the latter case, the frequency of each competitor formant is set to be constant at the geometric mean frequency of the formant track from which it was derived (see main text). Formant amplitude contours are not shown in this schematic.
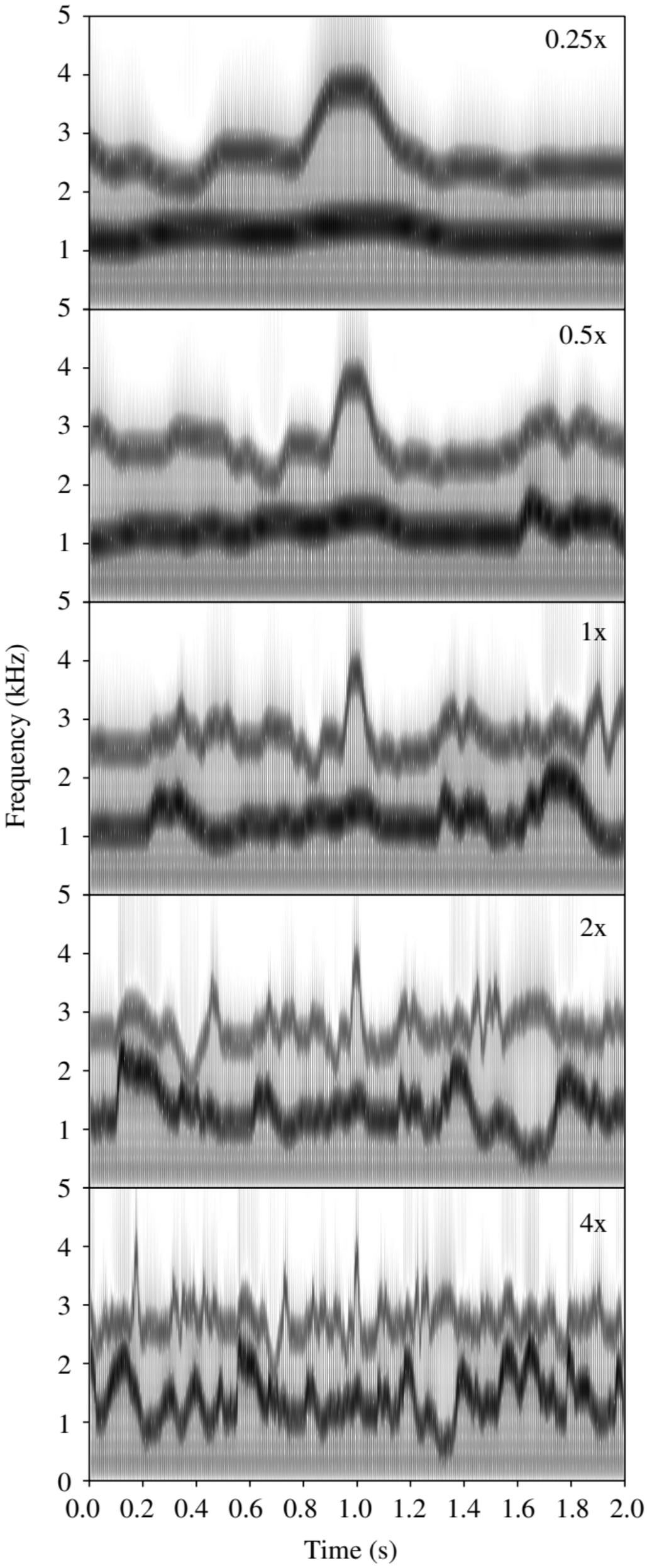
(2) Stimuli – spectrograms of an example set of competitors (F2C+F3C) used in the reversed-amplitude conditions, synthesized at different rates relative to that for the target sentences. The zero-rate case is not illustrated here.

(3) Stimuli – spectrograms of an example set of competitors (F2C+F3C) used in the constant-amplitude conditions, synthesized at different rates relative to that for the target sentences. The zero-rate case is not illustrated here.

(4) Results – influence of rate of formant-frequency variation on the effect of competitors (F2C+F3C) on the intelligibility of synthetic-formant analogues of target sentences. Mean scores and inter-subject standard errors (n=39) are shown separately for the constant-amplitude conditions (filled circles, solid line) and for the reversed-amplitude conditions (open circles, dashed line). The results for the control and dichotic-reference conditions are shown using asterisk symbols on the left- and right-hand sides of the figure, respectively. The top axis indicates which formants were presented to each ear; the bottom axis indicates the relative 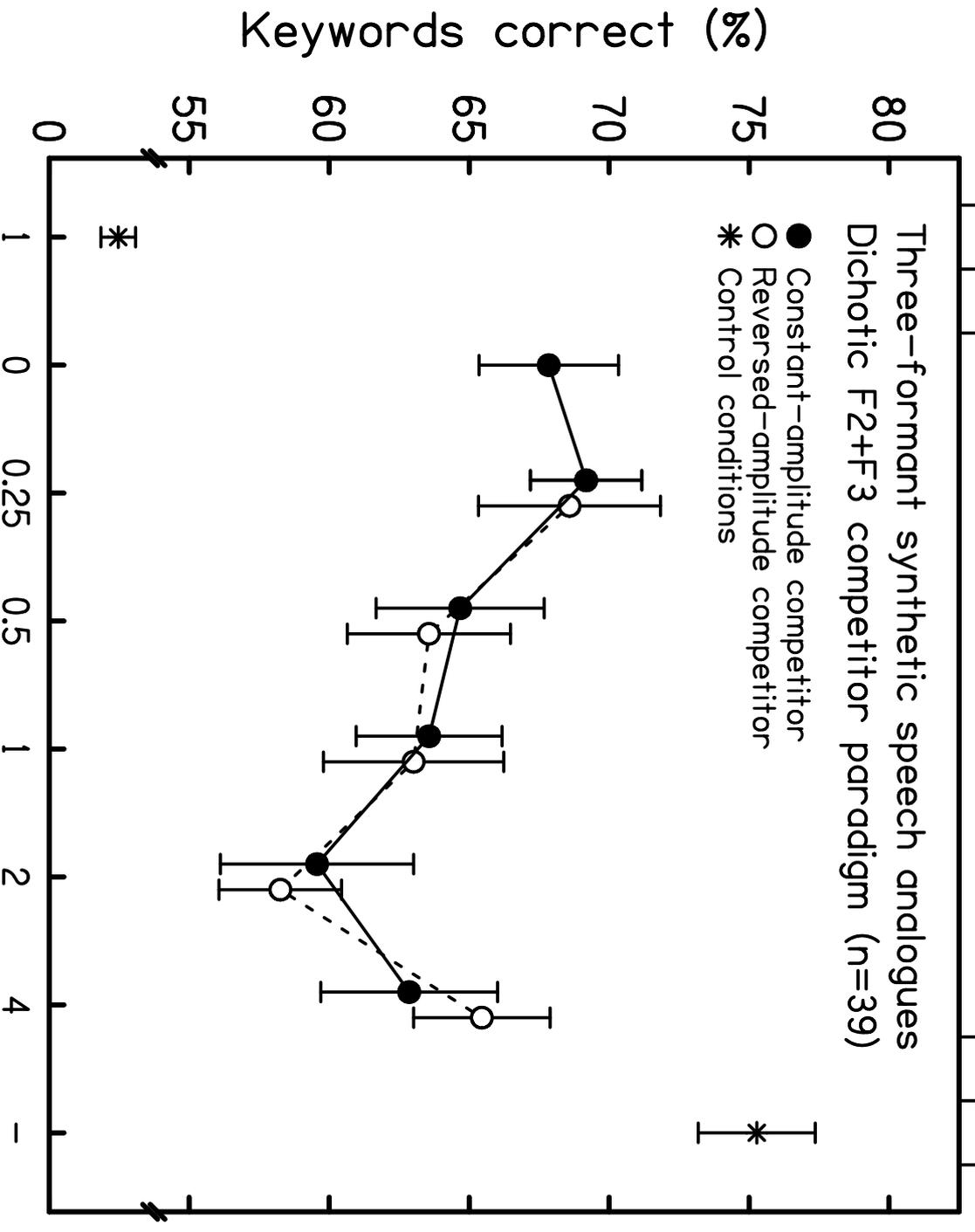rate for the competitor formants (when present).