JARO 13: 269–280 (2012)
DOI: 10.1007/s10162-011-0307-y

JARO
Journal of the Association for Research in Otolaryngology

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*

Received: 19 August 2011; Accepted: 18 November 2011; Online publication: 13 December 2011

ABSTRACT

How speech is separated perceptually from 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 reject to optimize recognition. Competitors were derived using formant-frequency contours extracted from extended passages spoken by the same talker and processed to alter the rate of formant-frequency variation, such that rate scale factors relative to the 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 at least 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, speech rate, formant-frequency variation, 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. Consequently, 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.

Correspondence to: Brian Roberts · Psychology, School of Life and Health Sciences · Aston University · Birmingham, B4 7ET, UK. Telephone: +44-121-2043887; fax: +44-121-2044090; email: b.roberts@aston.ac.uk

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 typically focused on well-established grouping cues, such as differences in F0 frequency (e.g., Bird and Darwin 1998; Summers et al. 2010). 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 grouping role of formant variation over time (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.

What aspects of formant-frequency variation might be important in the context of across-formant grouping? One dynamic property of speech that merits consideration is the *rate* of formant-frequency variation. Speech rate varies considerably; according to estimates of standard rates of speech by Tauroza and Allison (1990), the rate for slow speech is typically below 3.17 syllables/s and for fast speech it is above 5.33 syllables/s. As discussed in more detail below, changes in the rate of speech are commonly accompanied by changes in the rate of formant-frequency variation (e.g., Weismer and Berry 2003). Hence, in principle, differences in the rate of formant-frequency variation between talkers might provide a basis for the appropriate grouping of formants. For simplicity, consider the case in which a set of formants constituting target speech is accompanied by one extraneous formant. A hypothesis based on grouping by similarity predicts that the impact of the extraneous formant on intelligibility will be rate-tuned, such that maximum interference will occur when the rate of formant-frequency variation for the extraneous formant is most like that for the target formants. An alternative hypothesis is that faster variations are more disruptive, such that interference is proportional to the rate of formant-frequency variation in the extraneous formant. The study reported here was designed to evaluate these hypotheses.

The factors governing perceptual organization are generally revealed only when there is competition. Therefore, the experiment described below 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). The crux of the F2C paradigm is the dichotic presentation of two candidates for F2, such that intelligibility would be enhanced by the phonetic integration of one version (the true F2) with the first and third formants but impaired by the integration of the other (F2C). Hence, the listener must accept the true F2 and reject F2C to optimize recognition of the utterance. 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. F2 and F2C are presented to opposite ears to ensure that the effects of the competitor on intelligibility cannot be attributed simply to energetic masking of F2 by F2C, which would be likely to occur if the complete set of formants was presented monaurally or diotically.

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 suggest that the frequency contours of formants, but not their amplitude contours, are critical for across-formant grouping.

The aim of the experiment reported here was to elucidate the role of rate of formant-frequency variation in across-formant grouping. Synthetic three-formant analogues of natural sentence-length speech were used, in which each formant was generated using a pulse-excited second-order resonator. Target sentences were presented using a standard normalized rate of frequency variation and accompanied by competitor formants 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 5,000 Hz/s, extrapolating from the data of Tjaden and Weismer (1998)).

To explore further the grouping role of formant amplitude contours, 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. Note 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.

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. Thirty-nine 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 by rearranging items 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 vocal tract closures or unvoiced frication. A set of keywords was designated for each sentence. There is no generally agreed definition of what constitutes a keyword, and so the choice is somewhat arbitrary, but most 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 automatic estimates of the three formant frequencies 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 competitors 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). Here, we used two-formant competitors (F2C+F3C), with the aim of increasing the impact of the competitor on intelligibility; 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). A schematic illustrating the dichotic stimulus configuration is shown in Figure 1.

We computed the number of syllables/s 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 analogues were normalized at this mean rate. The

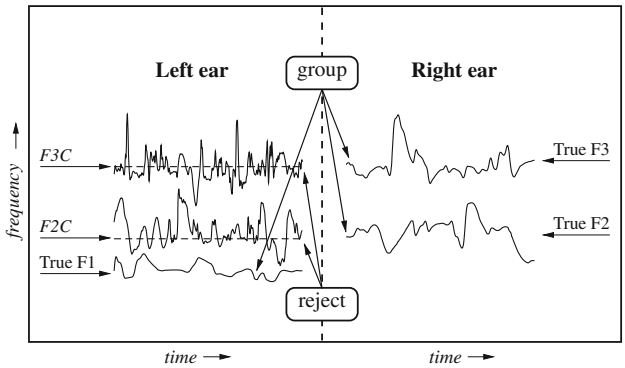

FIG. 1. Stimuli—schematic illustration of the dichotic configuration used in the experiment. The left ear receives F1 of the target sentence; the right ear receives F2 and F3. The competitor formants (F2C+F3C) are presented in the same ear as F1. The rate of frequency variation in the competitor formants can be controlled relative to that for the target sentence. Illustrated here are formant-frequency contours for the cases where the competitor rate is ×4 (*solid lines in bold*) and ×0 (*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.

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).

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 (but see "Results" for further comment on and re-evaluation of this choice). The limitations of parallel synthesis using second-order resonators introduced clear changes in timbre if the rate was increased beyond four times baseline, presumably because of artifacts associated with rapid changes in resonator center frequency. These timbre differences are undesirable because they provide additional cues for segregation, potentially making it easier for listeners to exclude the competitor when perceptually grouping the formants.

For relative rates from 0.25 to 4, competitor parameters were created by: (1) selecting at random one of the 24 available pairs of extracted and time-reversed formant tracks; (2) selecting at random a start point for splicing out a segment of the pair of formant tracks of the required length; (3) splicing out a segment corresponding to the required duration, as determined by the duration of the target sentence and the desired rate relative to baseline; (4) rescaling the duration of the spliced segment to match that of the target sentence, thereby obtaining 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, the ratio specifying the rate of variation in the pair of competitor formants relative to baseline is hereafter referred to as *competitor rate*.

In the reversed-amplitude conditions, 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 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. Stimuli were selected such that the center frequency of F2C was always ≥80 Hz from that of the true F1. 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.

Spectrograms of an example set of competitors, synthesized at different relative rates, are illustrated for the reversed- and constant-amplitude conditions in Figures 2 and 3, respectively. The examples are aligned across rate such that they share common frequency contours at 1 s on the abscissa. The effect of the rate manipulation on the pattern of frequency variation in the competitor formants is clearly appar-

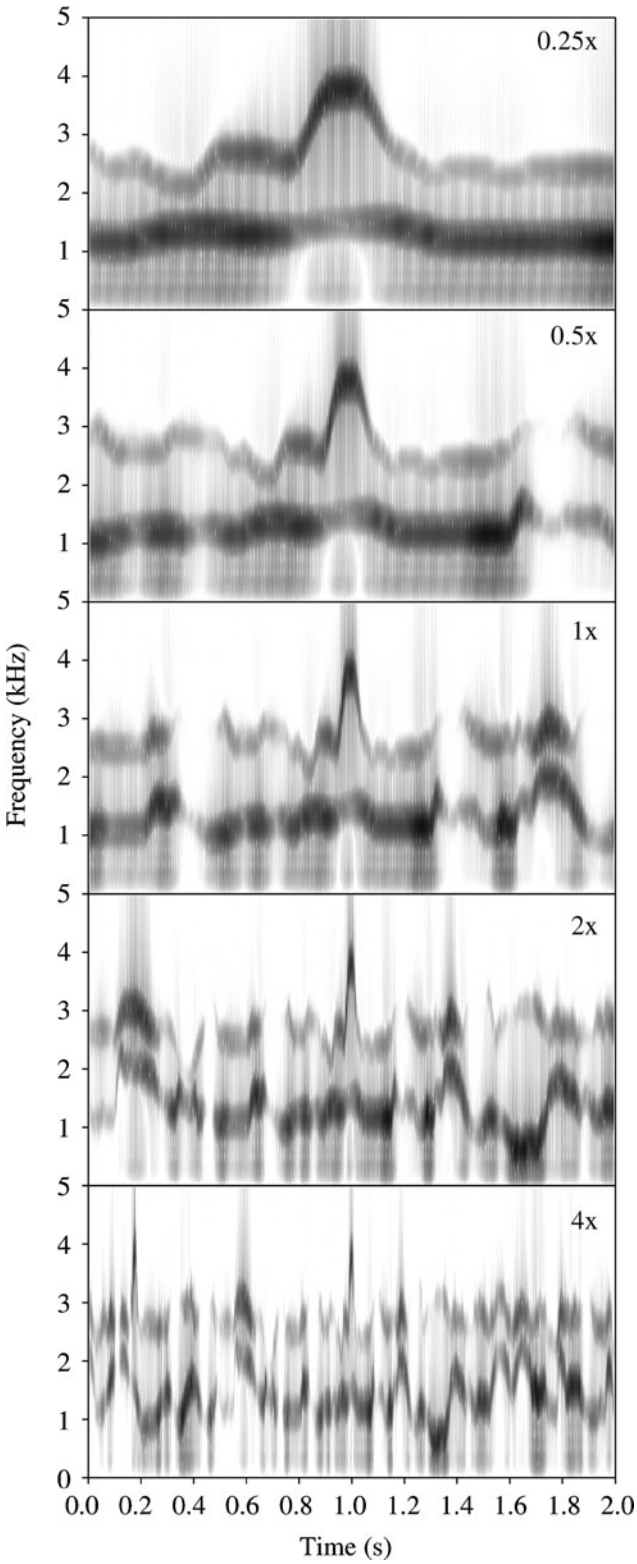

FIG. 2. Stimuli—spectrograms of an example set of competitors (F2C+F3C) used in the reversed-amplitude conditions, synthesized at different rates relative to that for the target sentences. The zero-rate case is not illustrated here.

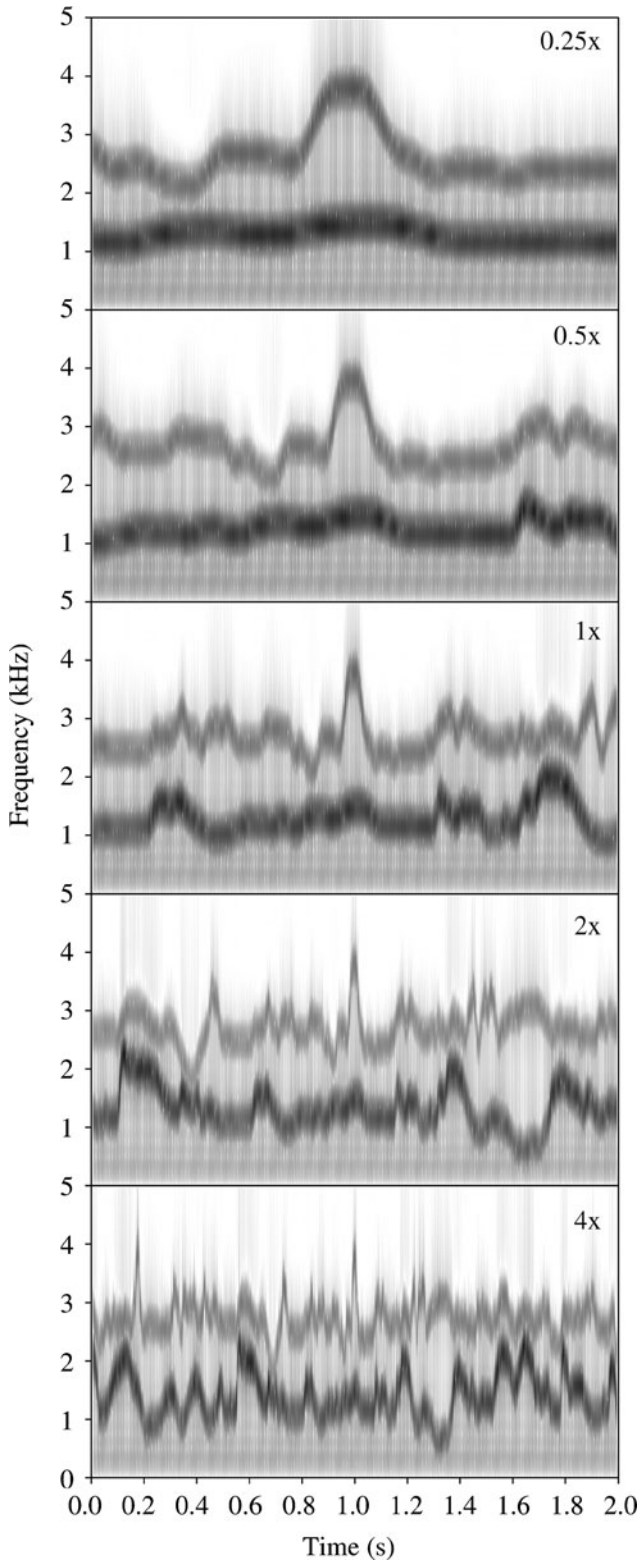

FIG. 3. Stimuli—spectrograms of an example set of competitors (F2C+F3C) used in the constant-amplitude conditions, synthesized at different rates relative to that for the target sentences. The zero-rate case is not illustrated here.

ent in both figures. The competitors were derived from almost continuously voiced speech with very few vocal tract closures, but nonetheless showed considerable variation in the amplitude contours for the reversed-amplitude conditions. This variation can be seen in the Figure 2 spectrograms, but it is better

illustrated in Figure 4, which shows the normalized amplitude contours of F2C and F3C for the ×1 rate exemplar shown in the spectrograms. The variation in the amplitude contours for the constant-amplitude conditions was nominally zero, though it should be acknowledged that there is inevitable variation in the spectral peak amplitude as formant frequencies traverse the harmonics, and also variation in formant amplitude with formant frequency arising from the use of second-order resonators with unity DC gain. These variations are too small to be apparent in the spectrograms shown in Figure 3.

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 C1 to be the only control condition necessary, because there is no reason to expect an increase in intelligibility were the competitor rate to be changed from 1 or the amplitude contour to be 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. There was no indication that F2 and F3 ever fused binaurally with F2C and F3C in these conditions. Presumably, the absence of binaural fusion reflected the different patterns of formant-frequency variation in the two ears. 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, which were generated using the natural pitch contours extracted from the original recordings. Diotic presentation was used in the training session; no competitor formants were present.

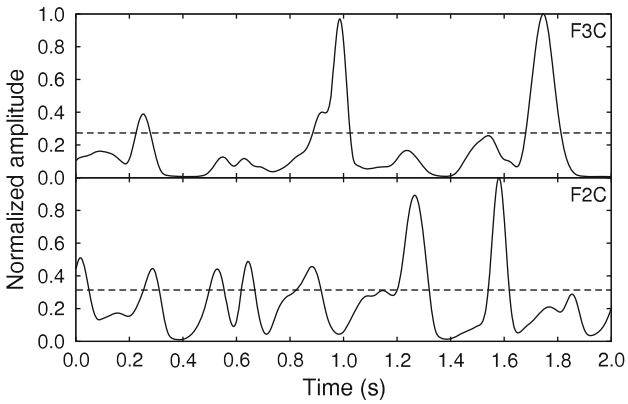

FIG. 4. Stimuli—amplitude contours corresponding to the competitor formants for the ×1 rate exemplars whose spectrograms are shown in Figures 2 and 3. The *lower* and *upper panels* show these contours for F2C and F3C, respectively; the amplitude contours for the reversed- and constant-amplitude conditions are indicated by *solid* and *dashed lines*, respectively. In each case, both contours are normalized to the maximum value for the reversed-amplitude contour.

TABLE 1

Stimulus properties for the conditions used in the main 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)

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 sound pressure level (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 across conditions in the level (≈2 dB in the left ear) and loudness of the stimuli, 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 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. The results from all listeners who completed the main experiment were included in the data analysis.

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 (two to four), 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

The results for the conditions where the competitor rate was four times baseline (C7 and C12) have been excluded; these data are not considered reliable owing to indications that they were contaminated by the stimulus artifacts identified for faster rates during pilot work.[1] Figure 5 shows the mean percentage scores (and inter-subject standard errors) across the remaining 11 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 asterisks 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 the competitor formants were absent. A one-factor within-subjects analysis of variance (ANOVA) for all 11 conditions showed a highly significant effect of condition on intelligibility [$F(10,380) = 86.759$, $p<0.001$]. 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 times 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, $p=0.014$ to $p<0.001$).

To assess the effects of competitor rate and amplitude contour, a two-factor ANOVA restricted to the (remaining) nine experimental conditions (C2–C6, C8–C11)

[1] A more extensive retrospective evaluation of the stimuli identified some instances of audible changes in timbre when the competitors were presented at four times the baseline rate, particularly for the reversed-amplitude condition. Consistent with these changes in timbre providing additional segregation cues, performance began to improve when the competitor rate was increased from twice to four times baseline, rather than continuing to decline. Until more reliable data are available, which will almost certainly require a different method of stimulus synthesis, we consider it prudent to set aside the results for the fastest competitor-rate conditions and to restrict our conclusions to rates up to twice baseline. For the record, the mean scores for C7 (constant amplitude) and C12 (reversed amplitude) were 62.9% and 65.5%, respectively. Note that excluding these data from the ANOVAs presented here had no effect on which terms were significant and which were not.

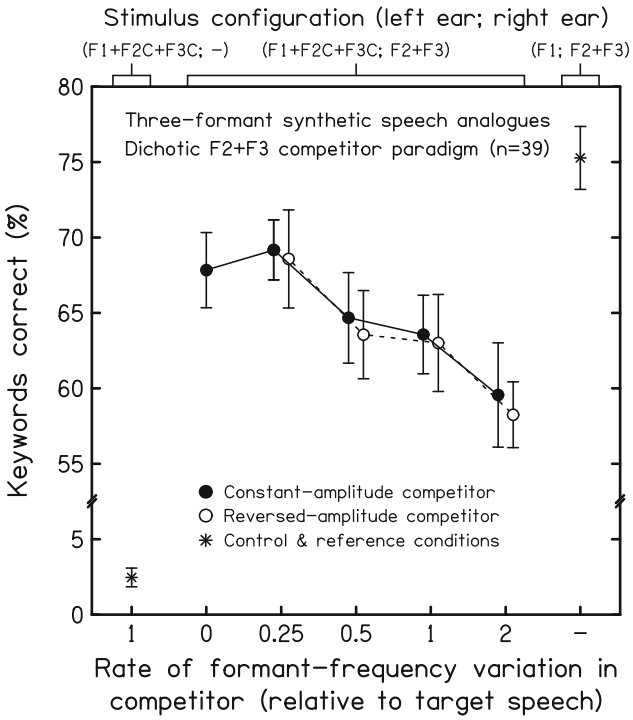

FIG. 5. Results—influence of rate of formant-frequency variation on the effect of competitors (F2C+F3C) on the intelligibility of synthetic-formant analogues of target sentences. Mean scores and inter-subject standard errors (n=39) are shown separately for the constant-amplitude (*filled circles, solid line*) and reversed-amplitude conditions (*open circles, dashed line*). The results for the ×4 rate cases are not included, owing to the likelihood that they were affected by stimulus artifacts (see text). The results for the control and dichotic-reference conditions are shown using *asterisks* 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).

was performed.[2] This showed that the main effect of competitor rate was highly significant [$F(4,152)=6.354$, $p<0.001$], but that neither the main effect of competitor amplitude contour [$F(1,38)=0.508$, $p=0.480$] nor the interaction of the two factors [$F(4,152)=0.032$, $p=0.998$] 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 tuned to the rate of formant-frequency variation in the competitor, relative to that for the target speech. Rather, the results suggest an approximately linear relationship, such that competitor efficacy increases with the rate of formant-frequency variation in the competitor, at least for rates up to twice baseline.

[2] 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.

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. This is 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, and the higher formants of the target sentence were presented in the contralateral ear (cf. Roberts et al. 2010; Summers et al. 2010).

Optimum performance clearly requires the listener to be able to select and combine information appropriately within and across ears. Although not measured here, target sentence intelligibility in the absence of competitors was undoubtedly lower under dichotic than under diotic presentation. A considerable dichotic cost has been reported previously for keyword scores using similar sentence-length materials (10–20 percentage points; Summers et al. 2010). Moreover, the dichotic configuration used here favors the integration of F1 with the pair of competitor formants rather than with the contralateral F2 and F3. In that regard, it is interesting that the competitors had rather less impact on intelligibility than one might have expected. Nonetheless, competitor efficacy was sufficient to distinguish between the conditions tested here. 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 tuning of the rate function around the baseline, as one might have expected if differences in speech rate are used as a cue for the appropriate grouping and segregation of formants.

Acoustic consequences of changes in speech rate

Our rate manipulation was intentionally simplistic and might 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 therefore acoustic) consequences in addition to the manifest effect of decreasing or increasing overall utterance duration. Increases in speech rate typically involve, among other things: (1) increased co-articulation, reduction, and assimilation (Gay 1981; Agwuele et al. 2008); (2) reductions in segment duration that are dependent on stress placement and phoneme class (Lehiste 1970; Byrd and Tan 1996); and (3) 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. In particular, 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).

Finally, it is also worth noting 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 (Roucos 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

Given that the spectro-spatial configuration of our stimuli was designed specifically to exclude the possibility of an explanation in terms of energetic masking, the observed effects of F2C+F3C may be characterized as involving primarily informational masking (Pollack 1975; for a recent review, see Kidd et al. 2008). A useful distinction can be made between those aspects of informational masking that cause difficulty in auditory object formation and those that cause difficulty in object selection (e.g., Ihlefeld and Shinn-Cunningham 2008). Consistent with this distinction, 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 what can be considered 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. Porter and Whittaker (1980) reported that 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. They 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, listeners were unable to exclude entirely a spectro-temporally varying contralateral masker. Although the dichotic arrangement of formants used in our study was rather different from the dichotic arrangement of voices used by Brungart and Simpson (2002), our results suggest that the inability of listeners to exclude completely a spectro-temporally varying contralateral masker can occur under a variety of circumstances. Indeed, Kidd et al. (2003) have shown that this kind of effect is not specific to speech.

The dependence of informational masking on the signal-to-masker ratio is relatively less than it is for energetic masking. Rather, the amount of informational masking observed when a target talker's speech is presented together with that of a masking talker

typically depends on the similarity of the two voices. Informational masking is maximal when target and masking speech come from the same talker, but it is reduced if the masker is spoken by a different talker, particularly if target and masking talkers differ in gender. In addition, some talkers are more resistant to masking (or act as more effective maskers) than others (Brungart 2001). Our findings for the effects of rate of formant-frequency variation on informational masking contrast with those previously reported for the effects of similarity between voices. Rather than showing a tendency for informational masking to increase with masker-target rate similarity, masking increases as the rate of formant-frequency variation in the competitor is increased. The findings of a recent study may be relevant to these contrasting outcomes. 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). However, at least for the procedure used by Cooke and Lu, it is interesting to note that this adjustment did not involve systematic changes in speech rate.

Effects of mixing speech of different rates together

Of particular relevance to 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 did not differ in their effect).[3]

Chen et al. (2008) presented a preliminary report on two 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, suggesting that our finding is not peculiar to the dichotic configuration of target and competitor formants used here.

[3] 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. We argue that their findings are better interpreted as evidence that faster interferers have a greater impact on intelligibility.

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. This finding is consistent with the hypothesis that faster variation in the frequencies of extraneous formants is more disruptive. 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

This research was supported by Research Grant EP/F016484/1 from the Engineering and Physical Sciences Research Council (UK) to Brian Roberts and Peter Bailey. We are grateful to Quentin Summerfield for enunciating the test sentences and to Meghna Patel and Rob Morse for allowing us access to their remodeled BKB sentences. Our thanks also go to Barb Shinn-Cunningham and the reviewers for their helpful comments on an earlier version of this manuscript.

REFERENCES

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

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

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

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

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

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

Bird J, Darwin CJ (1998) 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

Boersma P, Weenink D (2008) 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 (1990) Auditory scene analysis: the perceptual organization of sound. MIT, Cambridge, MA

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

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

Byrd D, Tan CC (1996) Saying consonant clusters quickly. J Phon 24:263–282

Chen J, Wu XH, Zou XF, Zhang ZP, Xu LJ, Wang MY, Li L, Chi HS (2008) Effect of speech rate on speech-on-speech masking. J Acoust Soc Am 123:3713–3714

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

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

Darwin CJ (2008) 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

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

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

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

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

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

Gay T (1981) Mechanism in the control of speech rate. Phonetica 38:148–158

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

Henke WL (2005) 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)

Ihlefeld A, Shinn-Cunningham B (2008) Spatial release from energetic and informational masking in a selective speech identification task. J Acoust Soc Am 123:4369–4379

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

Keppel G (1991) Design and analysis: a researcher's handbook. Prentice-Hall, Upper Saddle River, NJ

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

Kidd G, Mason CR, Richards VM, Gallun FJ, Durlach NI (2008) 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

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

Lehiste I (1970) Suprasegmentals. MIT, Cambridge, MA

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

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

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

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

Rand TC (1974) Dichotic release from masking for speech. J Acoust Soc Am 55:678–680

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

Remez RE (2001) 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, San Diego, pp 27–52

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

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

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

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

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

Roucos S, Wilgus AM (1985) High quality time-scale modification for speech. IEEE Proceedings on Acoustics, Speech and Signal Processing, vol. 10, pp. 493–496

Stevens KN (1998) Acoustic phonetics. MIT, Cambridge, MA

Stubbs RJ, Summerfield Q (1990) 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

SUMMERS RJ, BAILEY PJ, ROBERTS B (2010) Effects of differences in fundamental frequency on across-formant grouping in speech perception. J Acoust Soc Am 128:3667–3677

TAUROZA S, ALLISON D (1990) Speech rates in British English. Appl Linguist 11:90–105

TJADEN K, WEISMER G (1998) Speaking-rate-induced variability in F2 trajectories. J Speech Lang Hear Res 41:976–989

VAN SON RJJH, POLS LCW (1992) Formant movements of Dutch vowels in a text read at normal and fast rate. J Acoust Soc Am 92:121–127

WEISMER G, BERRY J (2003) Effects of speaking rate on second formant trajectories of selected vocalic nuclei. J Acoust Soc Am 113:3362–3378

WOUTERS J, MACON MW (2002) Effects of prosodic factors on spectral dynamics. I. Analysis. J Acoust Soc Am 111:417–427