The perceptual organization of sine-wave speech under competitive conditions

Brian Roberts and Robert J. Summers

Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK

Peter J. Bailey

Department of Psychology, University of York, Heslington, York YO10 5DD, UK

Running Title: Auditory grouping and sine-wave speech

PACS numbers: 43.66.Mk, 43.71.Es, 43.71.Rt, 43.66.Lj

Publication details: *Journal of the Acoustical Society of America, 2010, 128, 804-817.*

Address for correspondence:

Brian Roberts, Ph.D.

Psychology,

School of Life and Health Sciences,

Aston University,

Birmingham B4 7ET,

United Kingdom

Email: b.roberts@aston.ac.uk

ABSTRACT

Speech comprises dynamic and heterogeneous acoustic elements, yet it is heard as a single perceptual stream even when accompanied by other sounds. The relative contributions of grouping "primitives" and of speech-specific grouping factors to the perceptual coherence of speech are unclear, and the acoustical correlates of the latter remain unspecified. The parametric manipulations possible with simplified speech signals, such as sine-wave analogues, make them attractive stimuli to explore these issues. Given that the factors governing perceptual organization are generally revealed only where competition operates, the second-formant competitor (F2C) paradigm was used, in which the listener must resist competition to optimize recognition [Remez *et al.*, Psychol. Rev. **101**, 129-156 (1994)]. Three-formant (F1+F2+F3) sine-wave analogues were derived from natural sentences and presented dichotically (one ear = F1+F2C+F3; opposite ear = F2). Different versions of F2C were derived from F2 using separate manipulations of its amplitude and frequency contours. F2Cs with time-varying frequency contours were highly effective competitors, regardless of their amplitude characteristics. In contrast, F2Cs with constant frequency contours were completely ineffective. Competitor efficacy was not due to energetic masking of F3 by F2C. These findings indicate that modulation of the frequency, but not the amplitude, contour is critical for across-formant grouping.

I. INTRODUCTION

A. Background

Speech is composed of rapidly changing and diverse acoustic elements, yet it is typically heard as a single perceptual stream. In contrast, an arbitrary sequence of diverse acoustic elements repeated in rapid succession tends to undergo stream segregation, such that accurate judgments about relative order across streams become difficult or impossible (e.g., Warren et al., 1969). Furthermore, the perceptual coherence of speech remains strong even when it is accompanied by other sounds, including other speech. Despite considerable research on the acoustic cues enabling listeners to recover separate perceptual descriptions of the sources contributing to a sound mixture, a process known as auditory scene analysis (Bregman, 1990), only limited progress has been made in answering two important questions: What factors are responsible for the powerful perceptual coherence of speech? How do these factors contribute to the ability of listeners to separate the speech of concurrent talkers?

Research in auditory scene analysis has usually focused on relatively simple sounds, and has identified several general principles for grouping sound elements. These "primitives" include common spatial location, similarity, common fate, good continuation, and harmonic relations (e.g., Bregman, 1990; Darwin and Carlyon, 1995). There is clear evidence that these primitives influence the perceptual organization of speech. For example, good continuation cues provided by the formant transitions between phonetic segments (Cole and Scott, 1973; Dorman et al., 1975) or by the pitch contour (Darwin and Bethell-Fox, 1977) reduce the tendency for stream segregation in a repeating sequence of consonant-vowel (CV) syllables or vowels. When there is more than one talker, differences in fundamental frequency (F0) between the target and interfering speech improve intelligibility (e.g., Brokx and Nooteboom, 1982). This improvement does not depend simply on peripheral factors, but also on across-formant grouping by common F0 (Bird and Darwin, 1998). Also, differences in onset time or

in F0 between the formants of synthetic CV syllables influence their grouping (Darwin, 1981; Gardner et al., 1989).

Computational approaches to the separation of target speech from other non-stationary sound sources, including other speech, have had some success (Brown and Cooke, 1994; Barker and Cooke, 1999; Wang and Brown, 1999; for a review, see Cooke and Ellis, 2001). However, these approaches have been based on primitive grouping principles, which alone seem ill-equipped to provide a full account of the perceptual organization of speech. In particular, the speech produced by a single talker often seems to violate grouping principles like similarity and good continuation, owing to aspects of production such as closures (the articulation of plosives and affricates) and rapid changes in source excitation from periodic to noisy (e.g., voicing vs. frication). Nonetheless, it is worth noting that even natural speech is not immune to the enhanced tendency for stream segregation associated with rapid stimulus repetition (Warren et al., 1969). In particular, there is evidence that verbal transformations associated with the extended repetition of a spoken word (Warren, 1961) arise in part from the segregation and re-grouping of phonetic segments (Pitt and Shoaf, 2002).

Given that speech constitutes a highly familiar class of stimuli, it is likely that speech-specific factors, deriving from constraints on the dynamics of vocal tract articulators and their coordination, also contribute to its perceptual organization. That both primitive and speech-specific factors contribute to the perceptual organization of speech is widely accepted (e.g., Bregman, 1990; Davis and Johnsrude, 2007; Darwin, 2008), but different researchers vary greatly in the relative weights attributed to them (cf., e.g., Barker and Cooke, 1999; Remez, 2003, 2005). Three issues that have contributed to this disagreement are worthy of particular note. First, beyond the role of harmonic relations, there has been little systematic investigation of the contribution of primitives to the perceptual organization of speech. Second, not much is known about the grouping of non-speech broadband sounds with

complex dynamic properties. Third, despite the likely contribution of speech-specific factors to grouping, their acoustical correlates have not been well characterized in that context.

B. Sine-wave speech and the second-formant competitor (F2C) paradigm

Speech is highly redundant and can remain intelligible even after substantial distortions or simplifications of the signal. The parametric manipulations possible with simplified speech signals make them attractive experimental stimuli with which to explore the perceptual organization of speech. Here we have used sine-wave analogues of natural speech, created by adding together pure tones that follow the frequency and amplitude contours of the lower formants (Bailey et al., 1977; Remez et al., 1981). Sine-wave analogues of speech are perceptually bistable in that they can be heard either as non-speech – a collection of modulated tones – or as speech. Indeed, Cutting (1973a, b), who to our knowledge was the first person to use syllable-length sine-wave analogues in perceptual experiments, described these stimuli firmly as non-speech. Only after Roderick McInnes McGuire generated the first sentence-length sine-wave analogues at the Haskins Laboratories was their ability to support speech perception recognized and exploited for the first time (Bailey et al., 1977).

Several studies have used simplified speech signals, such as sine-wave analogues, but most have involved presenting single stimuli in quiet. The more realistic case of speech occurring simultaneously with other sounds, including other speech, has been explored extensively with natural utterances, but rarely using parametric manipulations of simplified speech. This is important, because the factors governing perceptual organization are generally revealed only where competition operates. A notable exception is the study of Barker and Cooke (1999), who found that listeners were generally unsuccessful in separating mixtures of two sine-wave speech messages, despite the reasonable intelligibility of these utterances when heard in isolation. This suggests that the sine-wave transformation largely robs these

stimuli of the acoustic cues required to retrieve them from a mixture. Interestingly, Barker and Cooke (1999) also found that listeners tended to be more successful at identifying target sine-wave speech in the mixture when the original speech contained more closures. Closures are associated with greater cross-formant synchrony in amplitude contour variation.

We used the second-formant competitor (F2C) paradigm in the current study, which involves a highly controlled form of competition. In the standard dichotic configuration, the first and third formants (F1 and F3) of the stimulus are presented in one ear and the second formant (F2) is presented in the other ear. The key aspect of the paradigm is the presentation of an alternative possibility for the second formant (the F2 competitor) in the same ear as F1 and F3; note that these spatial cues favor the fusion of F1 and F3 with F2C rather than with the true F2 (e.g., Culling and Summerfield, 1995). The extent to which the competitor disrupts the appropriate grouping of F1 and F3 with F2 is measured by the reduction in the intelligibility of the utterance. The F2C paradigm was first described by Remez et al. (1994), but to our knowledge it has been used only once since (Remez, 1996; see also Remez, 2001) and its methodological advantages have never been considered in any detail. One advantage is that the addition of F2C typically has less drastic effects on the intelligibility of sine-wave speech than does mixing two complete utterances together (cf. Barker and Cooke, 1999). The poor intelligibility typical of such mixtures probably reflects both increased peripheral masking of the individual formants and tracking errors likely to arise from sinusoidal formant analogues with crossing frequency contours. The latter issue is considered further in the General Discussion, but note that the dichotic F2C paradigm makes it straightforward to avoid the crossing of formants. Hence, it is less subject to floor effects and offers a more versatile tool for use with sine-wave speech. The other advantages of this paradigm arise from the use of changes in intelligibility as the index of changes in grouping.

Listeners typically find speech identification relatively straightforward, and such tasks avoid the problems inherent in exploring changes in perceptual organization using judgments of the number of sources heard (see, e.g., Cutting, 1976; Darwin, 1981). This is because duplex perception may occur, in which one acoustic element contributes simultaneously to two distinct percepts. Although first described in the context of speech stimuli (e.g., Rand, 1974; Liberman et al., 1981; Mann and Liberman, 1983), duplex perception has also been observed for musical chords (e.g., Pastore et al., 1983; Collins, 1985) and even for environmental sounds (e.g., Fowler and Rosenblum, 1990). Bregman (1990) has argued that, owing to the transparent nature of sound, duplex perception can arise whenever there are both conflicting acoustic cues and strong competition between different perceptual organizations. Given the perceptual bistability of sine-wave speech, the addition of a competitor formant may influence the associated speech or non-speech percepts, or both of them. Whatever this complex pattern of effects might be, measuring changes in intelligibility specifically targets the impact of the competitor formant on the *phonetic coherence* of the stimulus. Furthermore, no assumptions are necessary about whether increasing competition from F2C leads to the progressive exclusion of the true F2 from the phonetic percept (a displacement model) or to the progressive dilution of the contribution of the true F2 (an information-summation model). Both models predict essentially the same pattern of results; whichever applies, F2C must be rejected from the phonetic percept for optimum intelligibility.

C. Exploring the acoustic correlates of speech-like variation

Remez et al. (1994) showed, using sine-wave analogues of speech, that an F2C created by time-reversing F2 was a highly effective competitor; it reduced the mean syllabic intelligibility of the test materials by more than half. In contrast, a pure tone with a constant frequency and amplitude, set to the mean frequency of F2 and roughly equal to the amplitude of F2 at the syllabic nuclei, was not. They concluded that listeners were unable to exclude the

time-reversed F2 from the phonetic percept because it displayed "speech-like variation" (cf. Remez and Rubin, 1990), but were easily able to exclude the steady tone because it did not.

This finding of Remez et al. (1994) was replicated in a subsequent study that included a wider range of competitors, but which has been described only in outline (Remez, 1996, 2001). In the so-called "dithering" condition, the constant-frequency pure-tone competitor was replaced with a sequence of 200-ms pure-tone bursts, alternating between frequencies 10% above and below the mean F2 frequency. This competitor, which was described as non-stationary but also non-phonetic, remained ineffectual. The other two conditions were in effect a transition between the time-reversed F2C and the constant-frequency F2C, in which the frequency (and presumably also the amplitude) variation around the mean F2 frequency was scaled down by one-third or two-thirds of its original value. This "formant squash" manipulation applied to the time-reversed formant track reduced but did not eliminate its efficacy as a competitor. Remez (1996, 2001) interpreted this outcome in terms of the compression leading to a reduction in the competitor's speech-like properties.

Although it is plausible that the impact of the competitor formant on the intelligibility of sine-wave speech depends on speech-specific grouping factors rather than on primitives (Remez et al., 1994; Remez, 1996, 2001), thus far there has been a failure to characterize acoustically precisely what constitutes the speech-like variation of a competitor that causes it to disrupt the appropriate across-formant grouping. What we can say is that it is the modulation patterns of the formant contours that are likely to be critical for across-formant grouping. Remez and Rubin (1990) explored the relative contributions of variations in the frequency and amplitude contours of formants to the intelligibility of sine-wave speech, and proposed that frequency variation is more important than amplitude variation for maintaining intelligibility. However, to date neither the critical aspects of frequency and amplitude variation for across-formant grouping nor their relative contributions have been investigated.

We have used separate manipulations of the frequency and amplitude contours of competitor formants to tease apart their relative impacts on the intelligibility of sine-wave speech.

II. EXPERIMENT 1

Previous research has generated F2 competitors with plausibly speech-like frequency variation only by means of time-reversing the frequency contour of the true F2, and this manipulation has always been paired with time-reversal of the corresponding amplitude contour (Remez et al., 1994; Remez, 1996, 2001). In addition to independent control of the frequency and amplitude contours of time-reversed competitors in the current experiment, we also included a plausibly speech-like competitor generated by inverting the frequency contour of the true F2 (i.e., reflecting the contour about a mean value). This kind of manipulation was originally devised by Blesser (1972) and continues to be used in contexts where unintelligible stimuli with speech-like variation are required (e.g., Scott et al., 2000). However, the entire spectrum rather than a single formant has been inverted in previous studies, and to our knowledge inverted speech has not been used in the context of studies of auditory grouping. Unlike a competitor with a time-reversed frequency contour, one with an inverted frequency contour retains the correlations that the true F2 has with the motions of F1 and F3, but with a reversed polarity. It is unclear a priori whether this difference between time-reversed and inverted frequency contours will influence competitor efficacy.

A. Method

1. Listeners

Volunteers were tested initially 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 sine-wave speech (see Procedure). About

60% of them completed the training successfully and took part in the main experiment. A few of these listeners were subsequently replaced owing to below criterion performance in a follow-up test of intelligibility (see below). Overall, about half of the volunteers who passed the audiometric screening were included in the final dataset. This reflected considerable variation across listeners in the propensity to hear sine-wave analogues as intelligible speech, even after training; such variation has been observed in previous studies (e.g., Remez and Rubin, 1990; Remez et al., 1994). 24 listeners (8 males) successfully completed the experiment (mean age = 21.8 years, range = 18.5 – 35.6). All listeners were native speakers of British English.

2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of 72 sentences comprising almost continuously voiced speech as spoken by a British male talker of "Received Pronunciation" English. These sentences were taken from three sources (Binns and Culling, 2007; Bird and Darwin, 1998; Stubbs and Summerfield, 1990). The choice of speech with almost continuous voicing was intended to minimize grouping cues that might otherwise arise from the across-formant synchronies associated with closures (cf. Bird and Darwin, 1998; Barker and Cooke, 1999). The stimuli for the training session were derived from 40 sentences selected from commercially available recordings of the IEEE sentence lists (IEEE, 1969). 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; most designated keywords were content words.

For each sentence, the frequency contours of the first three formants were estimated from the waveform automatically every 1 ms by Praat (Boersma & Weenink, 2008) using a 25-ms-long Gaussian window. In practice, the third contour 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. These contours were used to generate sine-wave analogues of the sentences using three time-varying sinusoids (Bailey et al., 1977; Remez et al., 1981). Following Remez et al. (1994) and Remez (1996, 2001), in the main experiment the sine-wave analogues were presented in a dichotic configuration (left ear = F1+F3; right ear = F2). An example stimulus, which includes an F2 competitor (F2C), is illustrated in Fig. 1.

For each sentence in the main experiment, a set of F2 competitors was created by various manipulations of the frequency and amplitude contours of F2 (*f, a*). The frequency contour of F2C could be time reversed (R), inverted about the spectral centroid of the formant (I), or constant at its spectral centroid (C). A log frequency scale was used to compute the spectral centroid of the time series comprising the paired amplitude and frequency values for the F2 track; this scale was also used when inverting the frequency contour. The amplitude contour could be time reversed (R), time forward (i.e., normal, N), or constant at a value that preserved the RMS power (C). Stimuli were selected such that the F2C frequency was always ≥40 Hz from the F1 and F3 frequencies at any moment in time. Hence, there were no crossovers of formant tracks or approaches close enough to generate salient beats. The set of contours used to construct the variants of F2C is illustrated for an example sentence in Fig. 2. To keep the experiment within acceptable bounds, not every possible combination of the available frequency and amplitude contours was used.

There were 12 conditions in the main experiment; Table I summarizes which formants were presented to each ear and the properties of the added F2C (when present). Five of the conditions (C1-C5) were controls, for which the true F2 was absent. The stimuli for C1 comprised F1 and F3 only; the stimuli for C2-C5 also contained an F2C, which was chosen to be representative of the F2Cs used in the experimental conditions. Six of the conditions (C6-

C11) were experimental, for which the stimuli contained the true F2 and an F2C with one of the six pre-selected combinations of frequency and amplitude contours. Note that these conditions included the time-reversed frequency contour combined with each of the three amplitude contours tested (time reversed, normal, and constant). The final condition (C12) was the reference case, for which no F2C was present. 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 12 listeners tested. The follow-up session comprised all 72 sentences presented diotically without competitors.

All sine-wave analogues were synthesized using MITSYN (Henke, 2005) at a sampling rate of 22.05 kHz and with 10-ms raised-cosine onset and offset ramps. The stimuli 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 of 75 dB SPL; this describes the case when the left ear received F1 (the most intense formant) and F3. For a given sentence, F1 and F3 were presented at the reference level in all conditions; hence there was some variation in the overall level and loudness of the stimuli across conditions depending on the presence or absence of F2 and F2C. All sentences were presented at 72 dB SPL in the diotic follow-up (F1+F2+F3). In the training session, both the original recordings (44.1 kHz sampling rate) and the sine-wave analogues were presented diotically at 72 dB SPL.

3. Procedure

Listeners were seated in front of a computer screen and a keyboard in a sound-attenuating booth. There were three phases to the study – training, the main experiment, and the diotic follow-up. Listeners were free to take a break whenever they wished; typically the

study took about two hours to complete and consisted of either one or two testing sessions (first = training and main, second = diotic follow-up after brief refresher training). Stimuli were presented in quasi-random order in all phases of the study.

On each of the 40 trials in the training session, participants were able to listen to one of the sine-wave training stimuli up to a maximum of six times before typing in their transcription of the sentence. After each transcription was entered, feedback to the listener was provided by playing the original recording followed by a repeat of the sine-wave stimulus. Davis et al. (2005) found this "degraded-clear-degraded" presentation strategy to be an efficient way of enhancing the perceptual learning of speech analogues with unusual surface structures. We set a mean criterion of ≥50% keywords correct across the second half of the training trials for a listener to be included in the main experiment.

In the main experiment, each listener only heard any particular sentence during one trial in the experiment. As in the training, participants were able to listen to each stimulus up to six times before typing in their transcription, and the time available to respond was not limited. However, in the main experiment listeners did not receive feedback of any kind on their responses. Owing to the rotation of sentence allocations across conditions, the total number of listeners needed to produce a balanced dataset for the experiment was a multiple of twelve. Afterwards, using the same procedure, the listeners heard all 72 sentences used in the main experiment again but under diotic presentation and without competitors. Listeners were replaced in the main experiment if their performance in the diotic follow-up did not meet the mean criterion of ≥50% keywords correct.

4. Data analysis

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

6), 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 22 or 23 per set of six sentences). We classified responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly, and loose scoring, in which only the stem of the keyword must be identified correctly (see Foster et al., 1993).

The results for tight scoring are presented here, but in practice it made little difference which type of scoring was used (scores were about 2–3 percentage points lower on average for tight scoring). For the dichotic and diotic reference conditions, the responses were also converted automatically into phonetic representations using eSpeak (Duddington, 2008) for comparison with stored phonetic representations of the original sentences. We computed the mean percentage of phonetic segments that were correctly identified across all words in the sentences using HResults, part of the HTK software (Young et al., 2006). HResults uses a string alignment algorithm to find an optimal match between two strings by inserting or removing tokens from one of them. In effect, tight scoring and phonetic scoring represent the lower and upper limits of the intelligibility measures that can be computed for the test sentences used. We also conducted a simple analysis of the number of repeat listens (0–5) typically used for different conditions and by different listeners.

B. Results

Figure 3 shows the mean percentage scores (and intersubject standard errors) across conditions in terms of keywords identified correctly. Within-subjects analysis of variance (ANOVA) showed a highly significant effect of condition on intelligibility for the whole dataset [$F(11,253)=30.867$, $p<0.001$][1] and also when restricted to the six experimental conditions [$F(5,115)=20.019$, $p<0.001$]. The white, grey, and black bars indicate the results for the control, experimental, and dichotic reference conditions, respectively. The control conditions indicate that intelligibility was near floor for F1+F3 alone, and when any of the

F2Cs tested was added to F1+F3 in the absence of the true F2. Hence, these F2Cs did not in themselves support intelligibility. Paired-samples comparisons (two-tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967). The scores for the four control conditions that included a competitor were significantly different from those for their counterparts in the experimental conditions (f_R, a_R – C2 vs. C6; f_R, a_N – C3 vs. C7; f_I, a_N – C4 vs. C9; f_C, a_C – C5 vs. C11; $p \leq 0.001$, in all cases). The scores for all five control conditions did not differ from one another.

All F2Cs with time-varying frequency contours were highly effective competitors, regardless of their amplitude characteristics (see Fig. 3, first four grey bars). On average, these competitors caused performance to fall by 17.5 percentage points from that for the dichotic reference condition. If the impact on intelligibility of adding F2C is defined in relation to performance for the dichotic reference condition (corresponding to 0% competitor efficacy) and for the pooled F1+F2C+F3 control conditions (corresponding to 100% competitor efficacy), then the mean efficacy of these competitors was 67.5%. In contrast, F2Cs with constant frequency contours were completely ineffective competitors, irrespective of whether the amplitude contour was identical to that of the true F2 or was constant (see Fig. 3, last two grey bars). Pairwise comparisons indicated that the scores for the four experimental conditions for which the frequency contour of F2C was time-varying were significantly different from those for the two where it was constant, and also from the dichotic reference score ($p<0.001$, in all cases). The scores for the two experimental conditions where the frequency contour of F2C was constant did not differ from one another or from the dichotic reference score.

The mean diotic follow-up score for the 72 sentences used in this experiment was 62.4%. This was considerably higher than the mean dichotic reference score of 29.8%, albeit with the caveat that listeners had been exposed to degraded versions of these sentences during

the main experiment. Nonetheless, the apparent performance cost of dichotic presentation for sine-wave analogues of speech (32.6 percentage points) is rather greater than for otherwise comparable synthetic-formant analogues of speech (Summers et al., 2009). The corresponding mean phonetic scores for diotic and dichotic performance without competitors were 73.8% and 46.0%, respectively. Hence, the dichotic performance for sine-wave speech observed here is broadly comparable with that found by Remez et al. (1994) and Remez (1996, 2001), who reported syllable scores in the range 40%-45%.

There was inevitable variability in intelligibility across sentences, and for sentences at the low end of the range, intelligibility scores may have been susceptible to floor effects. Therefore, the data from the main experiment were explored further using a median split by rank order of the mean diotic follow-up scores for all 72 sentences. ANOVA was not applied to the partitioned data, because the median split resulted in variable patterns of contribution by individual listeners across conditions. Figure 4 shows the mean percentage scores separately for the upper and lower rank orders across conditions (upward and downward pointing triangles, respectively). Owing to the nature of the rank ordering, these scores were weighted equally per sentence rather than per keyword. As expected, differences between conditions were relatively small for the less intelligible sentences, plausibly as a result of floor effects. However, the more intelligible half of the sentences showed the impact on performance of the various F2Cs tested even more clearly than did the complete dataset. In particular, a comparison of the f_R, a_N and f_I, a_N experimental conditions indicated that F2C was a more effective competitor when its frequency contour was inverted than when it was time-reversed [Mann-Whitney $U(36,36) = 480.0$, $p<0.05$]. This finding is considered further in the General Discussion. The average fall in performance arising from the addition of a competitor with a time-varying frequency contour was 26.8 percentage points for the more

intelligible sentences (cf. 17.5 percentage points for all sentences), but note that the efficacy of 68.3% for these competitors was very similar to that for all sentences (67.5%).

There was some variation in the use of repeat listens across conditions; means ranged from 3.24 to 3.85 when pooled across listeners (overall mean = 3.63). There was a significant negative correlation across conditions between these values and the mean scores for keyword identification (Pearson's $r(10) = -0.81$, $p=0.01$), indicating a greater tendency to listen again to the less intelligible stimuli. However, the variation in the use of repeat listens between listeners was much greater; means ranged from 0.93 to 5.00 when pooled across conditions.

C. Discussion

F2 competitors typically reduced intelligibility, plausibly by providing an alternative to F2 in the perceptual organization of the sentences (Remez et al., 1994; Remez, 1996, 2001). Competitor efficacy was critically dependent on the time-varying properties of the frequency contour, regardless of the amplitude contour. This indicates that modulation of the frequency contour, but not of the amplitude contour, is the critical factor governing across-formant grouping. One caveat regarding the generality of this conclusion concerns the use of stimuli derived from almost continuously voiced speech. While it is true for our stimuli that there was typically considerable variation in the amplitude contour of each formant during the course of each sentence, the absence of closures did reduce the extent of this variation. Therefore, it remains possible that modulation of the amplitude contour would have more impact on across-formant grouping for speech that includes closures. This possibility was evaluated in experiment 2.

Another issue that merits consideration concerns the choice of the most appropriate constant-frequency counterparts to the dynamic-frequency competitors. In experiment 1, the constant frequency used for these competitors was set to equal the spectral centroid of the true F2 on a log frequency scale. Owing to the spectral tilt of our stimuli, the constant

frequency would typically have been about 100 Hz higher had it instead been set to equal the geometric mean frequency of the true F2. Such a rise might conceivably affect the efficacy of a constant-frequency competitor by reducing the extent of upward spread of masking from the more intense F1 to F2C. This issue was addressed in experiment 2.

III. EXPERIMENT 2

The generality of the findings from experiment 1 was explored using sine-wave analogues derived from speech including closures and unvoiced fricatives. There were fewer conditions than for experiment 1, but the experiment retained the key comparisons between competitors with constant or time-varying frequency and amplitude contours. Competitors with inverted frequency contours were not included in this experiment.

A. Method

The sine-wave analogues used here were derived from speech involving closures and unvoiced fricatives as spoken by the same talker. The sentences used were selected from the BKB sentence lists (Bench et al., 1979), and so were semantically simpler and more predictable than the almost continuously voiced sentences used in experiment 1. Nine conditions were used (3 controls, 5 experimental, 1 reference); those conditions from experiment 1 that were not included here are indicated by asterisks in Table I. All stimuli were generated as described for experiment 1, except that the frequency of F2C when constant was set to match the geometric mean frequency of the true F2, rather than its spectral centroid. Owing to the spectral tilt of natural speech, typically about -6 dB/oct above 400 Hz for a male talker, the geometric mean frequency of F2 is on average about 100 Hz higher than the spectral centroid for our stimuli. This should tend to reduce any upward spread of masking from the more intense F1 to F2C in the constant-frequency condition.

All three phases of the experiment (training, main, diotic follow-up) were run in the same way as their counterparts in experiment 1; the data were also scored in the same way. There were 54 sentences in total (9 conditions x 6 sentences per condition). Every sentence contained either three or four keywords; each sentence group contained 19 keywords. A multiple of nine listeners was required to produce a balanced dataset, owing to the rotation of sentence allocations across conditions. 18 listeners (10 males) successfully completed the experiment (mean age = 24.4 years, range = 18.3 – 43.9); two of these listeners also took part in experiment 1.

B. Results

Figure 5 shows the mean percentage scores (and intersubject standard errors) across conditions. Within-subjects ANOVA showed a highly significant effect of condition on intelligibility for the whole dataset [$F(8,136)=13.240$, $p<0.001$][1] and also when restricted to the five experimental conditions [$F(4,68)=6.840$, $p<0.001$]. The control conditions (white bars) indicate that intelligibility was low for F1+F3 alone, and near floor when any of the F2Cs tested was added to F1+F3 in the absence of the true F2. Hence, these F2Cs did not in themselves support intelligibility. Pairwise comparisons indicated that the scores for the two control conditions that included a competitor were significantly different from those for their counterparts in the experimental conditions (f_R, a_R – C2 vs. C6, $p=0.011$; f_R, a_N – C3 vs. C7, $p=0.005$). The scores for the three control conditions did not differ from one another.

As for experiment 1, all F2Cs with time-reversed frequency contours were highly effective competitors, regardless of their amplitude characteristics (see Fig. 5, first three grey bars). On average, these competitors caused performance to fall by 16.0 percentage points from that for the dichotic reference condition, which is very similar to the fall observed in experiment 1 for sine-wave analogues of almost continuously voiced speech. In terms of competitor efficacy, the impact of F2Cs with time-varying frequency contours observed here

was 56.3%, which is substantial but about 10 percentage points lower than for experiment 1. In contrast, F2Cs with constant frequency contours were once again completely ineffective competitors, irrespective of whether the amplitude contour was identical to that of the true F2 or was constant (see Fig. 5, last two grey bars). Pairwise comparisons indicated that the scores for the three experimental conditions where the frequency contour of F2C was time-varying were significantly different from those for the two where it was constant, and also from the dichotic reference score (range: $p=0.016$ to 0.001). The scores for the two experimental conditions where the frequency contour of F2C was constant did not differ from one another or from the dichotic reference score.

The mean diotic follow-up score for the 54 sentences used was 61.8%, which was considerably higher than the mean dichotic reference score of 35.6%. Although the apparent cost of dichotic presentation (26.2 percentage points) for sine-wave analogues of speech including closures and unvoiced fricatives was greater than for synthetic-formant analogues of speech (Summers et al., 2009), it was somewhat less than the cost for sine-wave analogues of speech with almost continuous voicing observed in experiment 1 (32.6 percentage points). Note that the diotic scores for experiments 1 and 2 were very similar, and so the observed difference in cost reflects primarily the higher dichotic scores for experiment 2. Indeed, overall performance across conditions in experiment 2 was 7.3 percentage points better than for the corresponding nine conditions in experiment 1. This probably reflects semantic differences between the sentences used, rather than the change from analogues of speech with almost continuous voicing to those of speech including closures and unvoiced fricatives. The corresponding mean phonetic scores for diotic and dichotic performance without competitors were 73.7% and 53.4%, respectively. Hence, the dichotic performance for sine-wave speech observed here was again broadly comparable with that reported by Remez et al. (1994) and

Remez (1996, 2001). The pattern of use of repeat listens across conditions was broadly similar to that observed for experiment 1; the overall mean was 3.70.

C. Discussion

Experiment 2 demonstrated essentially the same pattern of F2C efficacy as found in experiment 1, but using sine-wave analogues derived from speech involving closures and unvoiced fricatives. In particular, the main finding that the grouping of formants is governed by modulation of their frequency contours, but not their amplitude contours, was replicated in a context where there was greater variation in the amplitude contours of each formant than occurs for stimuli derived from almost continuously voiced speech.

The constant-frequency competitors used here differed from their counterparts in experiment 1 in that they were set to match the geometric mean frequency of the true F2 rather than its log spectral centroid. Clearly, these constant-frequency competitors were as ineffective as their counterparts in experiment 1, regardless of their amplitude contours, but also despite the average rise in frequency of about 100 Hz resulting from the change in synthesis strategy. This finding indicates that it is the lack of frequency change in these competitors, rather than their precise frequency in relation to the true F2, that is critical for their lack of efficacy. This is perhaps unsurprising given previous research that suggests a good deal of tolerance in the perceptual estimation of the frequency of F2 (e.g., Flanagan, 1955; Mermelstein, 1978). In particular, the results of Mermelstein (1978) indicate a mean difference limen for F2 in a dynamic context of about 185 Hz, which is about twice as large as the mean difference between the spectral centroid and geometric mean frequency of F2.

The issue of energetic masking has already been raised in the context of upward spread of masking from the more intense F1 to F2C. However, of greater concern for the interpretation of our findings is whether or not competitor efficacy might be related primarily to energetic masking of the less intense F3 by F2C, rather than to genuine competition

between F2C and the true F2 for integration with F1 and F3. The apparent robustness of sentence intelligibility in the presence of constant-frequency F2Cs implies that partial masking of F3 by F2C is unlikely to be a major factor governing competitor efficacy, but this interpretation does not take into account the much closer approaches of the frequency contours for F2C and F3 that often arise when time-varying competitors are used.

Several studies have shown that the relative levels of formants can be adjusted over a wide range without greatly affecting intelligibility (e.g., Ainsworth and Millar, 1972; Klatt, 1982a, b). This suggests that it might be possible to attenuate F2C to a level where energetic masking of F3 becomes negligible without substantially reducing its impact on intelligibility. However, the attenuation of F2C would inevitably increase its energetic masking by F1, and previous research demonstrating a broad tolerance to changes in relative formant level has not been conducted under competitive conditions. As an alternative approach to controlling for the effects of masking, for experiment 3 the F2 competitor paradigm was modified, based on the procedure described by Rand (1974). He observed that synthetic CV syllables which were highly intelligible under diotic presentation remained so under dichotic presentation of the form (F1; F2+F3). For the current purpose, the advantage of the form (F1±F2C; F2+F3) over (F1±F2C+F3; F2), where ± indicates the presence or absence of the competitor, is that the energetic masking of F3 is entirely independent of whether or not F2C is present. The two dichotic forms are hereafter referred to as the Rand and standard configurations, respectively. Controlling for masking effects of F2C on F1 is unnecessary, because F1 is both more intense and lower in frequency than F2C.

In certain circumstances, the perception of speech sounds shows a right-ear advantage (REA) in most listeners, such that they are reported more accurately than when presented to the left ear (e.g., Bryden, 1963; Shankweiler and Studdert-Kennedy, 1967). Two conditions have been found to promote the REA for speech – dichotic competition (e.g., Berlin et al.,

1973) and the use of stimuli with dynamic spectro-temporal properties (e.g., Shankweiler and Studdert-Kennedy, 1967; Darwin, 1971). The perceptual processes underlying the REA for speech are not well understood, but have often been conceptualized as a correlate of the left-hemisphere lateralization of speech processes (e.g., Kimura, 1961). Given that our experiments entailed dichotic presentation of sine-wave analogues of time-varying formants, it is possible that the relative efficacy of the formant competitors may have been influenced by differences in ear dominance. In particular, the F2C and the true F2 have always been presented to left and right ears, respectively. However, the few studies that allow an assessment of ear advantages for the processing of specific formants suggest that the particular assignment of formants to ears that we have used may not have affected our results. For example, Rand (1974) found no differences in consonant identification for synthetic CV syllables when presented in the configurations (F1; F2+F3) or (F2+F3; F1). Similarly, Cutting (1974) did not find an ear advantage in a dichotic temporal-order judgment task when the stimuli were sine-wave analogues of CV syllables (heard in this case as non-speech sounds). Given the equivocal nature of the background literature, it is unclear whether or not ear dominance effects should be expected in the F2 competitor paradigm. When this paradigm was first introduced, Remez et al. (1994) counterbalanced the assignment of formant analogues to ears, but they did not report whether or not any ear dominance effects were apparent.

IV. EXPERIMENT 3

The purpose of this experiment was to assess the extent to which the pattern of results observed in experiments 1 and 2 may have been influenced by effects of energetic masking or ear dominance. The extent to which the effect of an F2 competitor was attributable to masking was explored by including conditions in which F3 was presented to the opposite ear

to that receiving F2C. Possible effects of ear dominance were evaluated by including additional controls in which the assignment of formants to ears was counterbalanced.

A. Method

As for experiment 2, the sine-wave analogues were derived from speech involving closures and unvoiced fricatives, spoken by the same talker. However, the standard dichotic configuration (F1±F2C+F3; F2) was supplemented by one based on Rand's (1974) study, in which F3 was presented in the opposite ear to F1 (i.e., F1±F2C; F2+F3). The Rand configuration eliminated the possibility that F2C efficacy might have been influenced by masking of F3 by the more intense F2C. In experiments 1 and 2, F2C and the true F2 were always presented to the left and right ears, respectively, but here half of the listeners were tested using a reversal by ear of the stimulus configurations [i.e., standard case = (F2; F1±F2C+F3); Rand case = (F2+F3; F1±F2C)]. The sentences were taken from the BKB lists; the sine-wave analogues used were selected specifically to have relatively high intelligibility, on the basis of pre-testing with other listeners using the standard dichotic configuration.

There were four conditions in the main experiment, two with the standard dichotic configuration and two with the Rand configuration. For each configuration, the sine-wave analogues were presented either alone or in the presence of an F2C created by time reversing the frequency and amplitude contours of the true F2 (f_R, a_R). There were 16 sentences in total (4 conditions x 4 sentences per condition). Every sentence contained either three or four keywords; each sentence group contained either 12 or 13 keywords. Competitor efficacy was balanced approximately across sentence groups based on the outcome of the pre-testing. Multiples of eight listeners were required to produce a balanced dataset in this experiment. For each set of listeners, there were two complete rotations of the four sentence allocations across conditions; the ear receiving F1 (and F2C, when present) was the left ear in the first set of rotations and the right ear in the second set. 16 listeners (4 males) successfully completed

the experiment (mean age = 20.6 years, range = 18.7 – 24.1). None of the listeners had taken part in experiments 1 or 2, or to our knowledge in any previous studies of speech perception. All three phases of the experiment (training, main, diotic follow-up) were run in the same way as their counterparts in experiments 1 and 2; the data were also scored in the same way.

B. Results

Figure 6 shows the mean percentage scores (and intersubject standard errors) across conditions. For each condition, the means for the between-subjects variable – whether the ear receiving F1 was the left or the right – are indicated by leftward- and rightward-pointing triangles, respectively. A within-subjects ANOVA on the complete dataset (i.e., pooled across all 16 listeners, irrespective of the ear receiving F1) showed a highly significant effect of condition on intelligibility [$F(3,45)=10.071$, $p<0.001$]. Pairwise comparisons indicated that the effect on the scores of adding F2C was significant for both the standard [$t(15)=2.73$, $p=0.015$] and Rand configurations [$t(15)=5.74$, $p<0.001$]. The scores for the standard and Rand configurations did not differ from one another either when F2C was absent [$t(15)=0.69$, $p=0.496$] or when it was present [$t(15)=0.04$, $p=0.967$].

A competitor with time-reversed frequency and amplitude contours was highly effective in the standard context, as before, and equally effective in the Rand context. The absolute impact on performance of these competitors was very similar to that observed for the comparable BKB stimulus materials used in experiment 2 (average fall in intelligibility = 18.1 percentage points; standard and Rand = 16.6 and 19.6 percentage points, respectively). In terms of competitor efficacy, the absence of F1+F3 control conditions means that these changes in performance can only be used to compute minimum estimates of efficacy (standard = 35.5%, Rand = 39.5%). Note, however, that these values would rise to around 45% if we assume similar performance for the (missing) F1+F2C+F3 control cases as observed in experiment 2. This still implies that competitor efficacy was somewhat lower

than for experiments 1 and 2, but this may simply reflect the different criteria used to select the sentences for experiment 3.

The intelligibility cost for the standard configuration of presenting the stimuli dichotically rather than diotically (46.8% vs. 64.2%, cost = 17.4 percentage points) was rather less than for experiments 1 and 2. However, this is unsurprising given that the stimuli for experiment 3 were selected specifically on the grounds of their high dichotic intelligibility in pilot testing. Indeed, the reduced intelligibility cost resulted from a rise in the dichotic scores, not a fall in the diotic scores, compared with experiments 1 and 2. The corresponding mean phonetic scores for diotic and (standard) dichotic performance without competitors were 72.0% and 60.0%, respectively. Hence, the dichotic performance for sine-wave speech observed here was at least as good as that reported by Remez et al. (1994) and Remez (1996, 2001). For each of the four conditions, the effect of reversal by ear of the stimulus configurations was explored using an independent-samples pairwise comparison (two tailed). None of these comparisons was significant [in all cases, $t(14) \leq 0.35$, $p > 0.5$]. The pattern of use of repeat listens across conditions was broadly similar to that observed for experiments 1 and 2; the overall mean was 2.80.

C. Discussion

The results indicate that the impact on intelligibility of an F2C with time-reversed frequency and amplitude contours was very similar when the F3 was moved to the opposite ear. This outcome provides clear evidence that the effect of the extra formant was not due appreciably to energetic masking of F3 by F2C; presumably, the effect arose from genuine competition between the true F2 and F2C for integration with the other formants. Also consistent with a minimal role for energetic masking is our finding from experiments 1 and 2 that competitor efficacy did not increase when the amplitude contour of F2C was time-reversed rather than normal. Time reversal of speech changes the fluctuations of its amplitude

envelope from more damped to more ramped in character, and this change is known to increase the forward masking exerted by an interferer on target speech (Rhebergen et al., 2005). This reaffirms our interpretation of the results of experiments 1 and 2, namely that it is modulation of the frequency, but not the amplitude contour of a formant that governs its grouping with the other formants. While it is not possible to prove the null hypothesis, the absence of any discernible differences in outcome between the two sets of assignments of formants to ears suggests that any effects of ear dominance in the context of the F2C paradigm are negligible or absent.

V. GENERAL DISCUSSION

The results confirm and extend those of Remez et al. (1994) and Remez (1996, 2001). F2 competitors typically reduced intelligibility, plausibly by providing an alternative to F2 in the perceptual organization of the sentences. The impact of the F2 competitor on keyword intelligibility was just as great when F3 was moved to the opposite ear; hence the upward spread of masking from F2C to F3 is unlikely to have made a significant contribution to competitor efficacy. Rather, the effect of the competitor formant on intelligibility can be regarded as another example of informational masking (Pollack, 1975; see, e.g., Brungart, 2001; Arbogast et al., 2002). The main new finding is that modulation of the frequency contour, but not the amplitude contour, is critical for across-formant grouping. This is true regardless of whether the sine-wave analogues are derived from speech with almost continuous voicing or with closures and voiceless fricatives. Although not conclusive, the complete absence of an effect of constant-frequency competitors implies that the impact on intelligibility of an F2C is through the perceptual displacement of F2 rather than through information summation (i.e., the progressive dilution of the contribution of F2).

The clear absence of a role for modulation of the amplitude contour in across-formant grouping is rather surprising. In experiment 1, we used sine-wave analogues derived from

speech with almost continuous voicing to minimize any primitive grouping cues that might otherwise arise from the across-formant synchronies associated with closures (cf. Bird and Darwin, 1998). Indeed, as noted in the Introduction, Barker and Cooke (1999) found that listeners tended to be more successful at identifying the target in a mixture of two sine-wave speech signals when the original speech contained more closures. More generally, one might have expected coherent changes in amplitude across formants, associated with the alternation between between more open and more closed vocal-tract configurations, to provide a useful grouping cue to bind the formants together.

Carrell and Opie (1992) reported that applying coherent sinusoidal amplitude modulation (AM) at a high rate (50 or 100 Hz) to the different formant tracks of sine-wave speech improved its intelligibility. Although an account for this result based on increased perceptual fusion is plausible, it might also have arisen from band-widening of the formants, which would make the sine-wave analogues more similar to natural speech. Therefore, Carrell and Opie (1992) also explored the effect of applying different rates of AM to each sine-wave track. They found that intelligibility was lower in the mismatched than in the comodulated case and concluded that coherent AM did increase across-formant fusion. However, a recent follow-up study has supported the band-widening hypothesis by showing that the intelligibility of sine-wave speech increases even when the high-rate AM cue is conflicting (Lewis and Carrell, 2007). Indeed, such an effect of band-widening has been demonstrated recently in the context of sine-vocoded speech (Souza and Rosen, 2009). More generally, it should be noted that the AM rates used by Carrell and Opie (1992) were far higher than those characteristic of the AM spectrum of English and of other languages, which peaks around 5-6 Hz (core range of the syllable = 3-10 Hz; Greenberg and Arai, 2004). It is also worth noting that coherent AM does not appear to act as a grouping cue for binding

together harmonics – either for non-speech stimuli (Darwin et al., 1994) or for concurrent vowels (Summerfield and Culling, 1992; Moore and Alcántara, 1996).

The finding that the constant-frequency competitors had no discernible effect on the intelligibility of sine-wave speech under dichotic presentation does not necessarily imply that this would be the true for diotic presentation. First, diotic presentation would be likely to cause appreciable masking interactions between the true F2 and F2C. Second, an important aspect of separating voices is how the auditory system deals with formants whose frequency contours cross one another. The nature of this problem is highlighted by research, using non-speech sounds, showing that the crossing trajectories of continuous glides are perceived to bounce rather than cross unless the glides are distinguished by differences in timbre (Halpern, 1977, as reported in Bregman, 1990; Tougas and Bregman, 1990). The predominance of bouncing percepts for intersecting trajectories reflects the strong tendency for the auditory system to group acoustic elements by frequency proximity (Bregman, 1990). There has been little discussion of the significance of these findings for speech perception (an exception is Grossberg et al., 2004) and the only study of how crossing contours are perceived using speech-like stimuli has focused on intersecting pitch contours (Culling and Darwin, 1993). Indeed, the very low intelligibility typical of a mixture of two sine-wave speech signals (Barker and Cooke, 1999) probably reflects at least in part the cost of intersecting formant tracks in the absence of other cues for segregation, such as differences in pitch and timbre.

What aspects of frequency variation make F2C an effective competitor? The complex motions of the different formants preclude the possibility of across-formant grouping based simply on coherent frequency modulation (FM). Indeed, even when present at rates broadly in the same range as those for formant motion, there is evidence that coherent FM of the components of a complex tone does not act as a grouping cue. When harmonicity effects are controlled, slow-rate coherent FM does not favor the perceptual fusion of the components of

simple non-speech stimuli (Carlyon, 1991). Also, differences in the coherence of slow-rate FM do not help to segregate concurrent vowels (Summerfield and Culling, 1992; Lyzenga and Moore, 2005). In the context of research on informational masking, one might speculate that the frequency motion of a formant captures attention better than does amplitude motion, or that frequency motion of the additional formant generates a greater number of alternative phonetic hypotheses for the original utterance. As yet, there is insufficient evidence to determine whether the impact of frequency variation on across-formant grouping arises from speech-specific properties or from the operation of a hitherto undescribed primitive which influences the perceptual organization of dynamic broadband stimuli more generally. Nonetheless, the latter suggestion gains credence from recent evidence indicating that budgerigars can also display classic speech context effects, such as rate normalization, previously attributed to specialized speech-based principles (Welch et al., 2009).

The results of the current study may begin to constrain the possibilities for what constitutes speech-like variation in the context of grouping. Time reversal of the true F2 contour is likely to cause a *haphazard* reduction in the plausibility of the articulatory motions implied by F2C, in relation to F1 and F3. Consider, for example, the change in frequency of F2 during production of a labial stop consonant before and after a front vowel: F2 rises in the syllable-initial consonant as the vocal tract moves from a closed configuration to the open configuration for the vowel, and falls symmetrically as the vocal tract closes for the syllable-final consonant (e.g. Stevens, 1998). This relationship suggests that time reversal of the F2 contour is likely on average to have more impact on the plausibility of F2C in the context of an asymmetric utterance (e.g., a CV syllable) than on a more symmetric one (e.g., a CVC syllable). In contrast, frequency inversion of the F2 contour is likely to lead to a *systematic* reduction in the plausibility of these motions in relation to the other formants. Therefore, it is interesting to note that the inverted-frequency and normal-amplitude F2C was at least as

effective as the reversed-frequency and normal-amplitude F2C. This suggests that moment-to-moment relations between the frequency contours of different formants are not important. Rather, the critical aspects may relate to similarities in the longer-term statistical properties of different formant tracks, such as in the rate and magnitude of their frequency variations.

Whatever the relative contributions of primitive and speech-specific factors, the results of the current study suggest that systematic manipulation of the rate and magnitude of frequency excursions in the F2C, and measurement of how this affects intelligibility, should help to elucidate the factors governing across-formant grouping. Future research might also explore the extent to which these findings for sine-wave speech generalize to very different surface structures, such as noise-vocoded speech (Shannon et al., 1995), and to more realistic approximations of natural speech, such as synthetic-formant analogues.

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. Our thanks go to Quentin Summerfield for enunciating the test sentences and to Adele Goman for her assistance with data collection. We are grateful to Chris Darwin, Brian Moore, and the anonymous reviewers for their comments on an earlier version of this manuscript. We are also indebted to the late Oliver Postgate for inspiring our interest in simplified analogues of speech signals.

Presentations on this research were given at the 157th Meeting of the Acoustical Society of America (Portland, Oregon, May 2009) and at the British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness (University of Southampton, September 2009).

REFERENCES

Ainsworth, W.A., and Millar, J.B. (**1972**). "The effect of relative formant amplitude on the perceived identity of synthetic vowels," Language and Speech **15,** 328-341.

Arbogast, T.L, Mason, C.R., Kidd, G. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112,** 2086–2098.

Bailey, P.J., Summerfield, Q., and 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.

Barker, J., and Cooke, M. (**1999**). "Is the sine-wave speech cocktail party worth attending?" Speech Commun. **27,** 159-174.

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

Berlin, C.I., Lowe-Bell, S.S., Cullen, J.K., Thompson, C.L., and Loovis, C.F. (**1973**). "Dichotic speech perception: An interpretation of right-ear advantage and temporal offset effects," J. Acoust. Soc. Am. **53,** 699-709.

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

Bird, J., and Darwin, C.J. (**1998**). "Effects of a difference in fundamental frequency in separating two sentences," in *Psychophysical and Physiological Advances in Hearing*, edited by A.R. Palmer, A. Rees, A.Q. Summerfield, and R. Meddis (Whurr, London), pp. 263-269.

Blesser, B. (**1972**). "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," J. Speech Hear. Res. **15,** 5-41.

Boersma, P., and Weenink, D. (**2008**). "Praat, a system for doing phonetics by computer," (Institute of Phonetic Sciences, University of Amsterdam, The Netherlands).

Bregman, A.S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

Brokx, J.P.L., and Nooteboom, S.G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phonet. **10,** 23-36.

Brown, G.J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8,** 297-336.

Brungart, D.S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109,** 1101-1109.

Bryden, M.P. (**1963**). "Ear preference in auditory perception," J. Exp. Psychol. **65,** 103-105.

Carrell, T.D., and Opie, J.M. (**1992**). "The effect of amplitude comodulation on auditory object formation in sentence perception," Percept. Psychophys. **52,** 437-445.

Carlyon, R.P. (**1991**). "Discriminating between coherent and incoherent frequency modulation of complex tones," J. Acoust. Soc. Am. **89,** 329-340.

Cole, R.A., and Scott, B. (**1973**). "Perception of temporal order in speech: The role of vowel transitions," Can. J. Psychol. **27,** 441-449.

Collins, S.C. (**1985**). "Duplex perception with musical stimuli: A further investigation," Percept. Psychophys. **38,** 172-177.

Cooke, M., and Ellis, D.P.W. (**2001**). "The auditory organization of speech and other sources in listeners and computational models," Speech Commun. **35,** 141-177.

Culling, J.F., and Darwin, C.J. (**1993**). "The role of timbre in the segregation of simultaneous voices with intersecting F0 contours," Percept. Psychophys. **54,** 303-309.

Culling, J.F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98,** 785-797.

Cutting, J.E. (**1973a**). "Perception of speech and nonspeech, with and without transitions," Haskins Lab. Status Rep. Speech Res. **SR-33,** 37-46.

Cutting, J.E. (**1973b**). "Perception of speech and nonspeech, with speech-relevant and speech-irrelevant transitions," Haskins Lab. Status Rep. Speech Res. **SR-35/36,** 55-64.

Cutting, J.E. (**1974**). "Two left-hemisphere mechanisms in speech perception," Percept. Psychophys. **16,** 601-612.

Cutting, J.E. (**1976**). "Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening," Psychol. Rev. **83,** 114-140.

Darwin, C.J. (**1971**). "Ear differences in the recall of fricatives and vowels." Q. J. Exp. Psychol. **23,** 46-62.

Darwin, C.J. (**1981**). "Perceptual grouping of speech components differing in fundamental frequency and onset-time," Q. J. Exp. Psychol. **33A,** 185-207.

Darwin, C.J. (**2008**). "Listening to speech in the presence of other sounds," in *The Perception of Speech: From Sound to Meaning*, edited by B.C.J. Moore, L.K. Tyler, and W. Marslen-Wilson (Special Issue, *Phil. Trans. R. Soc. Lond. B* **363,** 1011–1021).

Darwin, C.J., and Bethell-Fox, C.E. (**1977**). "Pitch continuity and speech source attribution," J. Exp. Psychol. Hum. Percept. Perform. **3,** 665-672.

Darwin, C.J., and Carlyon, R.P. (**1995**). "Auditory grouping," in *Hearing: Handbook of Perception and Cognition (2nd ed.),* edited by B.C.J. Moore (Academic Press, London), pp. 387-424.

Darwin, C.J., Ciocca, V., and Sandell, G.J. (**1994**). "Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic," J. Acoust. Soc. Am. **95,** 2631-2636.

Davis, M.H., and Johnsrude, I.S. (**2007**). "Hearing speech sounds: Top-down influences on the interface between audition and speech perception," Hear. Res. **229,** 132-147.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., and 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.

Dorman, M.F., Cutting, J.E., and Raphael, L.J. (**1975**). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol. Hum. Percept. Perform. **2,** 121-129.

Duddington, J. (**2008**). eSpeak 1.36, http://espeak.sourceforge.net/ (date last viewed 4/23/10).

Flanagan, J.L. (**1955**). "A difference limen for vowel formant frequency," J. Acoust. Soc.

Am. **27,** 613-617.

Foster, J.R., Summerfield, A.Q., Marshall, D.H., Palmer, L., Ball, V., and Rosen, S. (**1993**). "Lip-reading the BKB sentence lists: Corrections for list and practice effects," Brit. J. Audiol. **27,** 233-246.

Fowler, C.A., and Rosenblum, L.D. (**1990**). "Duplex perception: A comparison of monosyllables and slamming doors," J. Exp. Psychol. Hum. Percept. Perform. **16,** 742-754.

Gardner, R.B., Gaskill, S.A., and Darwin, C.J. (**1989**). "Perceptual grouping of formants with static and dynamic differences in fundamental frequency," J. Acoust. Soc. Am. **85,** 1329-1337.

Greenberg, S., and Arai, T. (**2004**). "What are the essential cues for understanding spoken language?" IEICE Trans. Inf. & Syst. **E87-D,** 1059-1070.

Grossberg, S., Govindarajan, K.K., Wyse, L.L., Cohen, M.A. (**2004**). "ARTSTREAM: a neural network model of auditory scene analysis and source segregation," Neural Networks **17,** 511–536.

Henke, W.L. (**2005**). *MITSYN: A Coherent Family of High-Level Languages for Time Signal Processing*, software package (Belmont, MA); e-mail: mitsyn@earthlink.net; http://home.earthlink.net/~mitsyn (date last viewed 4/23/10).

Institute of Electrical and Electronics Engineers (IEEE) (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Transactions on Audio and Electroacoustics, **AU-17,** 225–246.

Keppel, G., and Wickens, T.D. (**2004**). *Design and Analysis: A Researcher's Handbook, 4th ed.* (Pearson Prentice Hall, NJ).

Kimura, D. (**1961**). "Cerebral dominance and the perception of verbal stimuli." Can. J. Psychol. **15,** 166-171.

Klatt, D.H. (**1982a**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," Proc. IEEE International Conference on Speech, Acoustics, and Signal Processing, 1278-1281.

Klatt, D.H. (**1982b**). "Speech processing strategies based on auditory models," in *The Representation of Speech in the Peripheral Auditory System,* edited by R. Carlson and B. Granstrom (Elsevier, Amsterdam), pp. 181-196.

Lewis, D.E., and Carrell, T.D. (**2007**). "The effect of amplitude modulation on intelligibility of time-varying sinusoidal speech in children and adults, Percept. Psychophys. **69,** 1140-1151.

Liberman, A.M., Isenberg, D., and Rakerd, B. (**1981**). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," Percept. Psychophys. **30,** 133-143.

Lyzenga, J., and Moore, B.C.J. (**2005**). "Effect of frequency-modulation coherence for inharmonic stimuli: Frequency-modulation phase discrimination and identification of artificial double vowels," J. Acoust. Soc. Am. **117,** 1314-1325.

Mann, V.A., and Liberman, A.M. (**1983**). "Some differences between phonetic and auditory modes of perception," Cognition **14,** 211-235.

Mermelstein, P. (**1978**). "Difference limens for formant frequencies of steady-state and consonant-bound vowels," J. Acoust. Soc. Am. **63,** 572-580.

Moore, B.C.J., and Alcántara, J.I. (**1996**). "Vowel identification based on amplitude modulation," J. Acoust. Soc. Am. **99,** 2332-2343.

Pastore, R.E., Schmuckler, M.A., Rosenblum, L., and Szczesiul, R. (**1983**). "Duplex perception with musical stimuli," Percept. Psychophys. **33,** 469-474.

Pitt, M.A., and Shoaf, L. (**2002**). "Linking verbal transformations to their causes," J. Exp. Psychol. Hum. Percept. Perform. **28,** 150-162.

Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **57,** S5.

Rand, T.C. (**1974**). "Dichotic release from masking for speech," J. Acoust. Soc. Am. **55,** 678-680.

Remez, R.E. (**1996**). "Perceptual organization of speech in one and several modalities: Common functions, common resources," in ICSLP-1996, 1660-1663.

Remez, R.E. (**2001**). "The interplay of phonology and perception considered from the

perspective of perceptual organization," in *The Role of Speech Perception in Phonology*, edited by E. Hume and K. Johnson (Academic Press, San Diego), pp. 27-52.

Remez, R.E. (**2003**). "Establishing and maintaining perceptual coherence: Unimodal and multimodal evidence," J. Phonet. **31,** 293–304.

Remez, R.E. (**2005**). "Perceptual organization of speech," in *Handbook of Speech Perception*, edited by D.B. Pisoni and R.E. Remez (Blackwell, Oxford), pp. 28-50.

Remez, R.E., and Rubin, P.E. (**1990**). "On the perception of speech from time-varying acoustic information: Contributions of amplitude variation," Percept. Psychophys. **48,** 313-325.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S., and Lang, J.M. (**1994**). "On the perceptual organization of speech," Psychol. Rev. **101,** 129-156.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (**1981**). "Speech perception without traditional speech cues," Science **212,** 947-950.

Rhebergen, K.S., Versfeld, N.J., and Dreschler, W.A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118,** 1274-1277.

Scott, S.K., Blank, C.C., Rosen, S., and Wise, R.J.S. (**2000**). "Identification of a pathway for intelligible speech in the left temporal lobe," Brain **123,** 2400-2406.

Shankweiler, D.P., and Studdert-Kennedy, M. (**1967).** "Identification of consonants and vowels presented to left and right ears," Q. J. Exp. Psychol., **19**, 59-63.

Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270,** 303-304.

Snedecor, G.W., and Cochran, W.G. (**1967**). *Statistical Methods, 6th ed.* (Iowa U.P., Ames, Iowa).

Souza, P., and Rosen, S. (**2009**). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," J. Acoust. Soc. Am. **126,** 792–805.

Stevens, K.N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA).

Stubbs, R.J., and Summerfield, Q. (**1990**). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listers," J. Acoust. Soc. Am. **87,** 359-372.

Summerfield, Q., and Culling, J.F. (**1992**). "Auditory segregation of competing voices: Absence of effects of FM or AM coherence," Phil. Trans. Roy. Soc. Lond. B, **336,** 357-366.

Summers, R.J., Roberts, B., and Bailey, P.J. (**2009**). "Effects of differences in fundamental frequency on cross-formant grouping in speech perception," J. Acoust. Soc. Am. **125,** 2605-2606 (A).

Tougas, Y., and Bregman, A.S. (**1990**). "Auditory streaming and the continuity illusion," Percept. Psychophys. **47,** 121-126.

Wang, D.L., and Brown, G.J. (**1999**). "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Networks **10,** 684–697.

Warren, R.M. (**1961**). "Illusory changes of distinct speech upon repetition – The verbal transformation effect," Brit. J. Psychol. **52,** 249-258.

Warren, R.M., Obusek, C.J., Farmer, R.M., and Warren, R.P. (**1969**). "Auditory sequence: Confusion of patterns other than speech and music," Science **164,** 586-587.

Welch, T.E., Sawusch, J.R., and Dent, M.L. (**2009**). "Effects of syllable-final segment duration on the identification of synthetic speech continua by birds and humans," J. Acoust. Soc. Am. **126,** 2779–2787.

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., et al. (**2006**). The HTK book [for HTK version 3.4] (Cambridge University Engineering Department).

FOOTNOTE

(1) As a precaution, given the low scores obtained for the control conditions, the ANOVA was repeated using arcsine-transformed data ($Y' = 2 \arcsin(\sqrt{Y})$, where Y is the proportion correct score; see Keppel and Wickens, 2004, p.155). The results confirmed the outcome of the original analysis.

TABLE I

Stimulus properties for the conditions used in experiment 1 (main session). The frequency and amplitude contours of F2C were derived from those of the true F2. The frequency contour (*f*) could be time reversed (R), inverted about the spectral centroid of F2 (I), or constant at its spectral centroid (C). The amplitude contour (*a*) could be time reversed (R), time forward (i.e., normal, N), or constant at a value that preserved the RMS power (C). Asterisks indicate conditions that were not included in experiment 2; all others were included. Note, however, that the frequency contour for F2C in experiment 2, when constant, matched the geometric mean frequency of F2 rather than its spectral centroid.

Condition	***Stimulus configuration (left ear; right ear)***	***Frequency (f) and amplitude (a) contours of F2C (when present)***
1	(F1+F3; –)	– –
2	(F1+F2C+F3; –)	f_R, a_R
3	(F1+F2C+F3; –)	f_R, a_N
4*	(F1+F2C+F3; –)	f_I, a_N
5*	(F1+F2C+F3; –)	f_C, a_C
6	(F1+F2C+F3; F2)	f_R, a_R
7	(F1+F2C+F3; F2)	f_R, a_N
8	(F1+F2C+F3; F2)	f_R, a_C
9*	(F1+F2C+F3; F2)	f_I, a_N
10	(F1+F2C+F3; F2)	f_C, a_N
11	(F1+F2C+F3; F2)	f_C, a_C
12	(F1+F3; F2)	– –

FIGURE CAPTIONS

FIG. 1. Stimuli for experiment 1 – standard dichotic stimulus configuration, illustrated using a sine-wave analogue of the sentence "Every man won a lemon razor." The upper panels show the amplitude envelopes of the signals presented to each ear, normalized to the maximum value in the left ear. The lower panels show the corresponding spectrograms. In this example, the second-formant competitor (F2C) was generated by time-reversing the frequency and amplitude contours of F2.

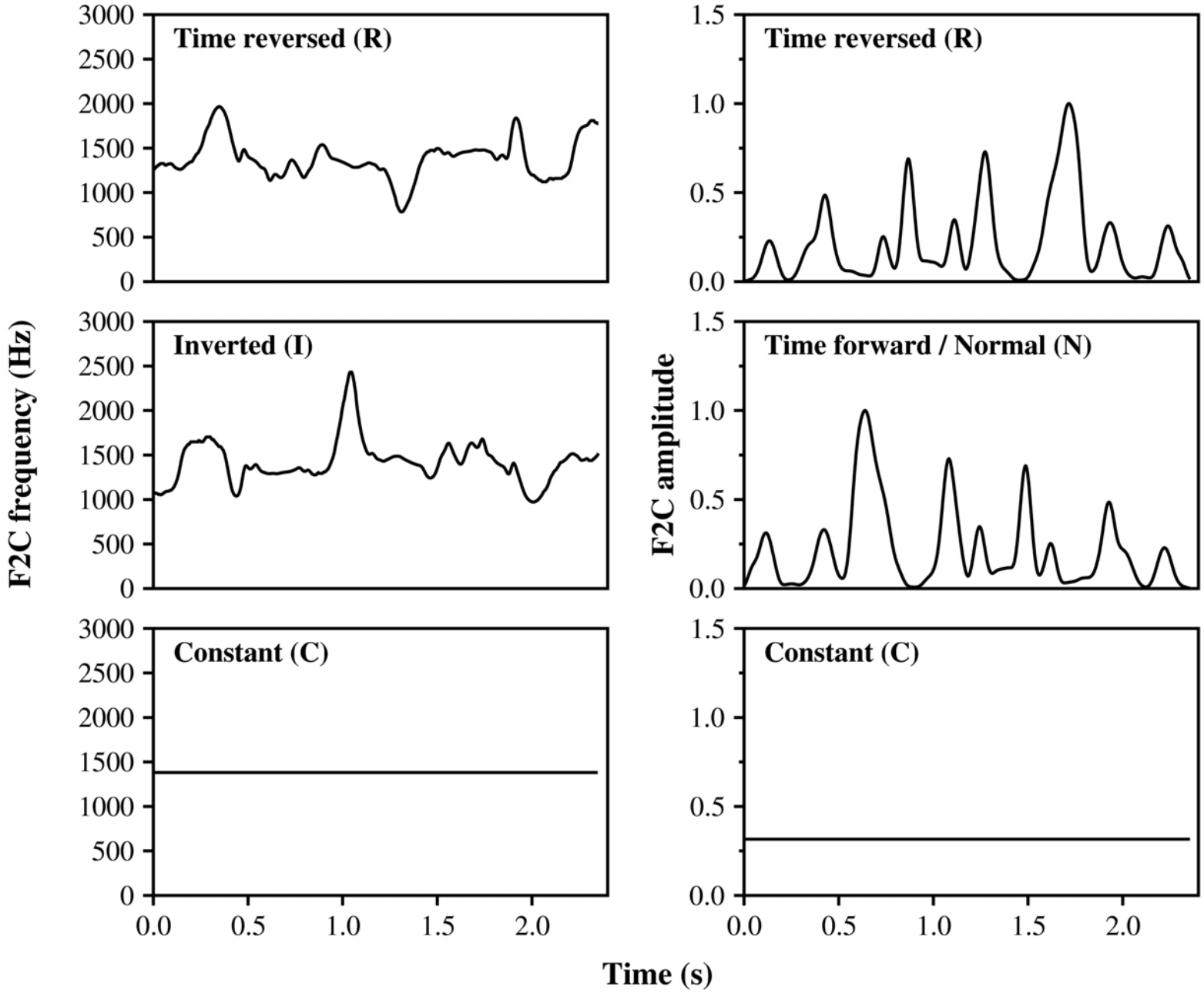
FIG. 2. Stimuli for experiment 1 – frequency and amplitude contours for the different competitors (F2Cs) added to the sine-wave analogue of the sentence "Every man won a lemon razor." The left- and right-hand panels show, respectively, the set of frequency and amplitude contours for F2C derived from F2. Amplitude contours are shown normalized to the maximum value in the original F2 contour. The frequency contour was time-reversed (R), inverted about its spectral centroid (I), or constant at its spectral centroid (C). The amplitude contour was time-reversed (R), time-forward/normal (N), or constant at a value preserving the RMS power (C).

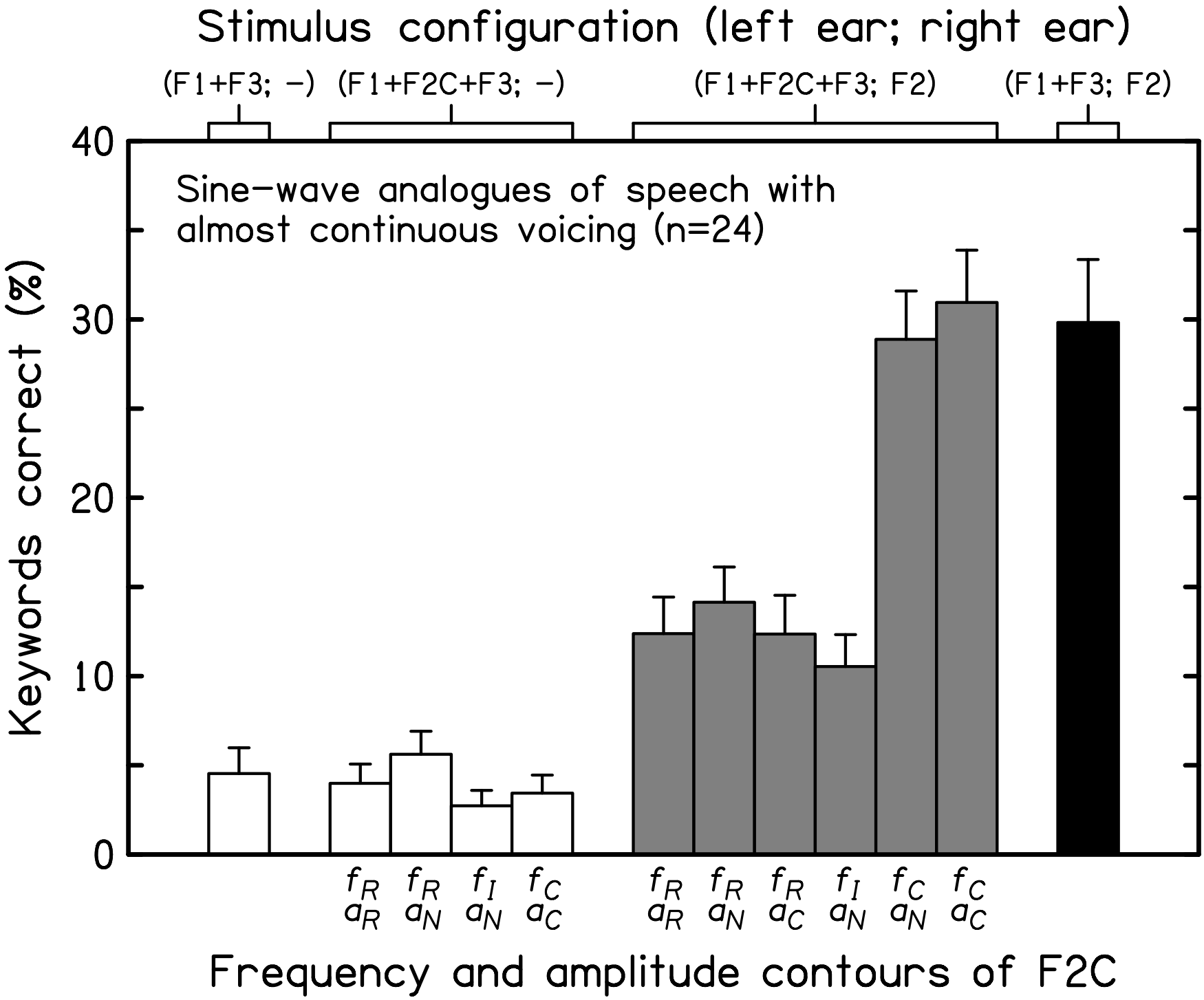
FIG. 3. Results for experiment 1 – influence of frequency and amplitude contour on the effect of competitors (F2Cs) on the intelligibility of sine-wave analogues of sentences spoken with almost continuous voicing. Mean scores and inter-subject standard errors (n=24) are shown for the control conditions (white bars), experimental conditions (grey bars), and the dichotic reference condition (black bar). The top axis indicates which formants were presented to each ear; the bottom axis indicates the frequency (*f*) and amplitude (*a*) contours of F2C (when present). The frequency contour was time reversed (R), inverted about its spectral centroid (I), or constant at its spectral centroid (C). The amplitude contour was time reversed (R), time forward (i.e., normal, N), or constant at a value preserving the RMS power (C).

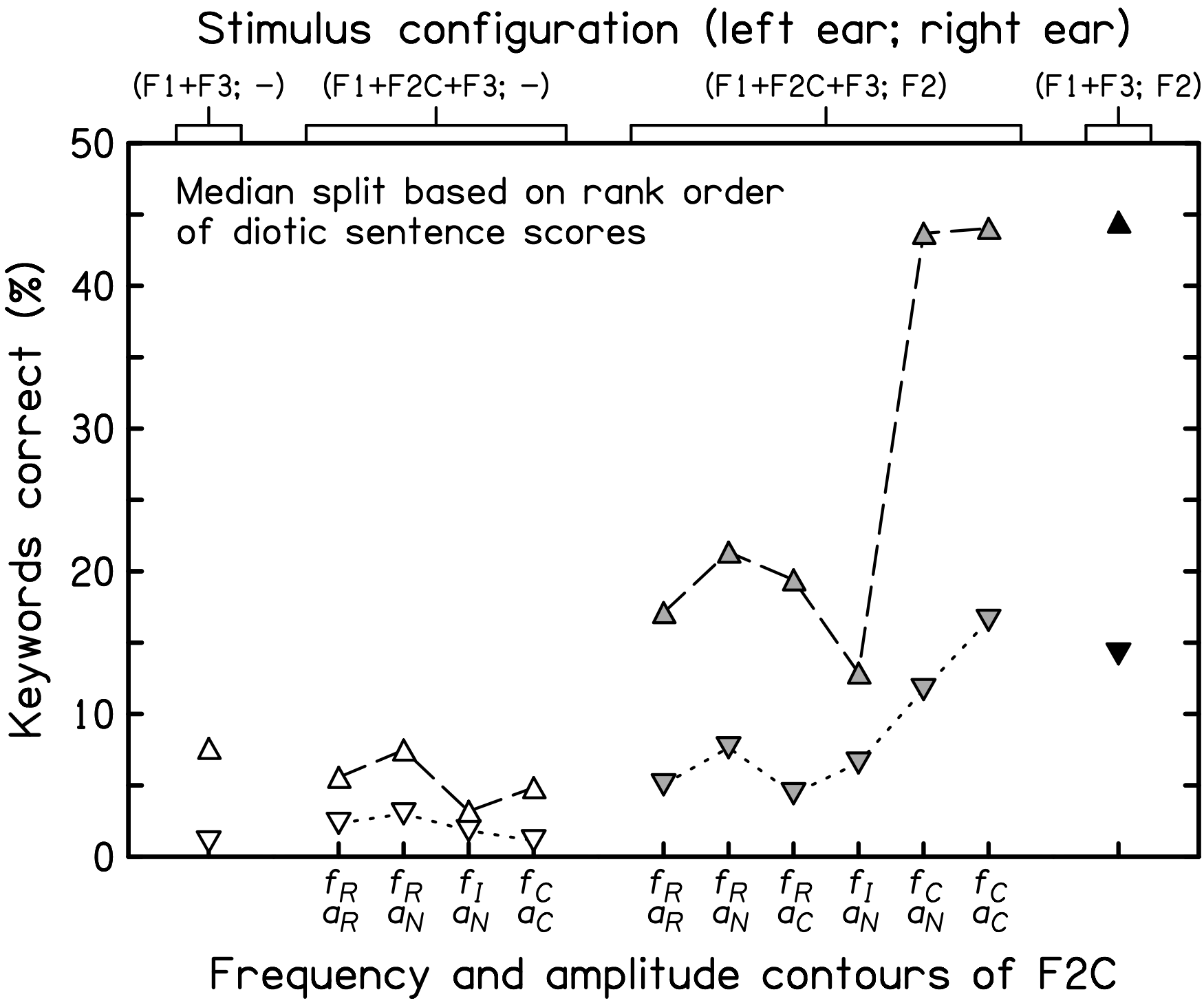
FIG. 4. Results for experiment 1 – influence of frequency and amplitude contour on the effect of competitors (F2Cs) on the intelligibility of sine-wave analogues of sentences spoken with almost continuous voicing. Mean scores are shown following a median split of the data by rank order of the diotic scores for all 72 sentences. The results for the more intelligible sentences are indicated by upward-pointing triangles and dashed lines. The results for the less intelligible sentences are indicated by downward-pointing triangles and dotted lines. Axis labels and symbol shadings correspond to their counterparts in Fig. 3.

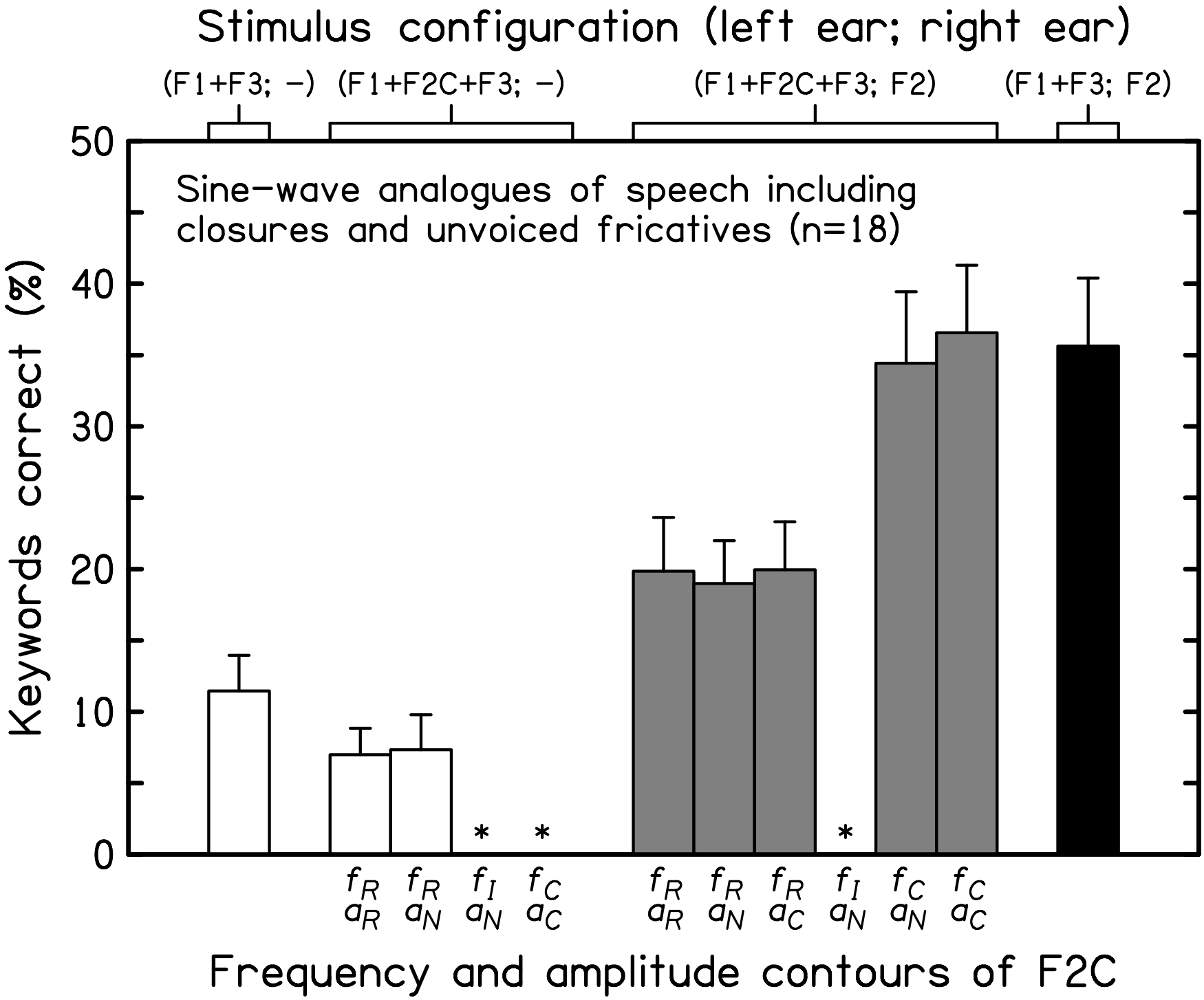
FIG. 5. Results for experiment 2 – influence of frequency and amplitude contour on the effect of competitors (F2Cs) on the intelligibility of sine-wave analogues of sentences involving closures and unvoiced fricatives. Mean scores and inter-subject standard errors (n=18) are shown for the control conditions (white bars), experimental conditions (grey bars), and the dichotic reference condition (black bar). The top axis indicates which formants were presented to each ear; the bottom axis indicates the frequency and amplitude contours of F2C (when present). The frequency contour was either time reversed (R) or constant at its geometric mean (C). The amplitude contour was time reversed (R), time forward (i.e., normal, N), or constant at a value preserving the RMS power (C). Asterisks indicate conditions from experiment 1 that were not included in experiment 2.

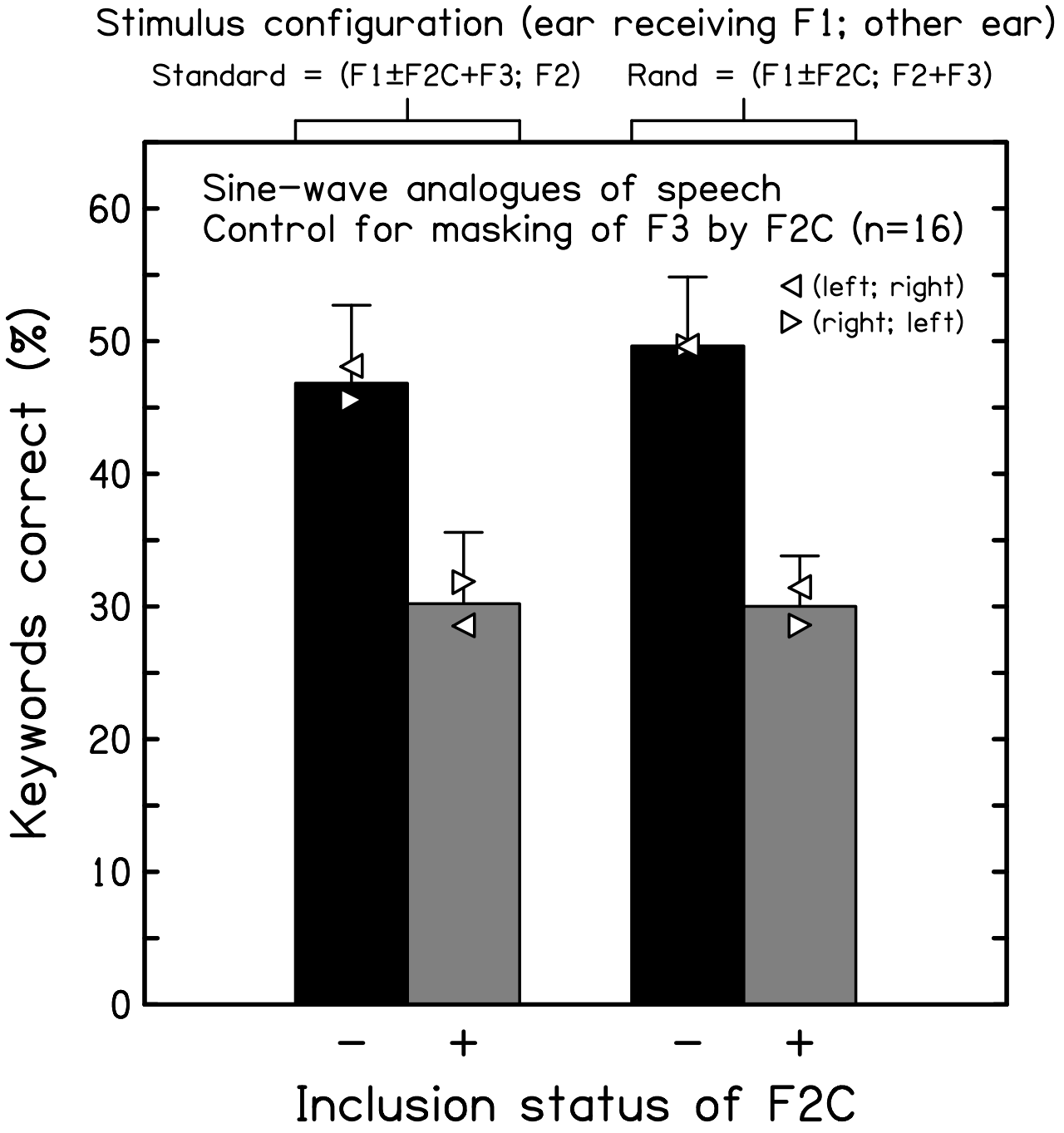
FIG. 6. Results for experiment 3 – effect of a competitor (F2C) with time-reversed frequency and amplitude contours on the intelligibility of sine-wave speech in two dichotic configurations. Mean scores and inter-subject standard errors (n=16) are shown in the absence (black bars) and presence (grey bars) of F2C. The paired columns on the left- and right-hand sides correspond to the results for the standard and Rand configurations, respectively (see top axis). For each condition, the leftward- and rightward-pointing triangles indicate the mean scores when the ear receiving F1 was left or right, respectively (see inset).


Time reversed (R)
Inverted (I)
Constant (C)
Time reversed (R)
Time forward / Normal (N)
Constant (C)
F2C frequency (Hz)
F2C amplitude
Time (s)
3000
2500
2000
1500
1000
500
0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0


Stimulus configuration (left ear; right ear)
(F1+F3; –)
(F1+F2C+F3; –)
(F1+F2C+F3; F2)
(F1+F3; F2)
Sine-wave analogues of speech with
almost continuous voicing (n=24)
Keywords correct (%)
0
10
20
30
40
f_R a_R
f_R a_N
f_I a_N
f_C a_C
f_R a_R
f_R a_N
f_R a_C
f_I a_N
f_C a_N
f_C a_C
Frequency and amplitude contours of F2C


Stimulus configuration (left ear; right ear)
(F1+F3; –)
(F1+F2C+F3; –)
(F1+F2C+F3; F2)
(F1+F3; F2)
Median split based on rank order
of diotic sentence scores
Keywords correct (%)
50
40
30
20
10
0
f_R a_R
f_R a_N
f_I a_N
f_C a_C
f_R a_R
f_R a_N
f_R a_C
f_I a_N
f_C a_N
f_C a_C
Frequency and amplitude contours of F2C


Stimulus configuration (left ear; right ear)
(F1+F3; –)
(F1+F2C+F3; –)
(F1+F2C+F3; F2)
(F1+F3; F2)
Sine-wave analogues of speech including
closures and unvoiced fricatives (n=18)
Keywords correct (%)
50
40
30
20
10
0
*
*
*
f_R a_R
f_R a_N
f_I a_N
f_C a_C
f_R a_R
f_R a_N
f_R a_C
f_I a_N
f_C a_N
f_C a_C
Frequency and amplitude contours of F2C


Stimulus configuration (ear receiving F1; other ear)
Standard = (F1±F2C+F3; F2)
Rand = (F1±F2C; F2+F3)
Sine-wave analogues of speech
Control for masking of F3 by F2C (n=16)
(left; right)
(right; left)
Keywords correct (%)
0
10
20
30
40
50
60
−
+
−
+
Inclusion status of F2C