# Informational masking of monaural target speech by a single contralateral formant

Brian Roberts[a] and Robert J. Summers
*Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, United Kingdom*

(Received 27 January 2015; revised 14 April 2015; accepted 14 April 2015)

Recent research suggests that the ability of an extraneous formant to impair intelligibility depends on the variation of its frequency contour. This idea was explored using a method that ensures interference cannot occur through energetic masking. Three-formant (F1 + F2 + F3) analogues of natural sentences were synthesized using a monotonous periodic source. Target formants were presented monaurally, with the target ear assigned randomly on each trial. A competitor for F2 (F2C) was presented contralaterally; listeners must reject F2C to optimize recognition. In experiment 1, F2Cs with various frequency and amplitude contours were used. F2Cs with time-varying frequency contours were effective competitors; constant-frequency F2Cs had far less impact. To a lesser extent, amplitude contour also influenced competitor impact; this effect was additive. In experiment 2, F2Cs were created by inverting the F2 frequency contour about its geometric mean and varying its depth of variation over a range from constant to twice the original (0%–200%). The impact on intelligibility was least for constant F2Cs and increased up to ~100% depth, but little thereafter. The effect of an extraneous formant depends primarily on its frequency contour; interference increases as the depth of variation is increased until the range exceeds that typical for F2 in natural speech.
 [http://dx.doi.org/10.1121/1.4919344]

[JFC] Pages: 2726–2736

## I. INTRODUCTION

Listeners are often faced with circumstances in which they must direct their attention to one talker in the presence of other talkers, a situation known as the cocktail party problem (Cherry, 1953). Spectral prominences corresponding to the acoustic resonances of the vocal tract are an important feature of the speech signal; the frequencies and amplitudes of these formants change as the shape of the vocal tract is changed by movements of the articulators, particularly the tongue, lips, and jaw. Most notably, the frequencies of the first three formants and their patterns of change over time are a critical source of information for identifying the phonetic segments articulated by a talker (e.g., Roberts *et al.*, 2011). Hence, when more than one talker is speaking at once, perceptually separating the formant ensemble reaching the ears into a figure (target) and background (interferer) is necessary for successful communication.

Interfering speech can affect the intelligibility of target speech through energetic masking, in which the auditory-nerve response to the target is swamped by the response to the masker, through modulation masking, in which amplitude variation in the masker reduces the sensitivity of the auditory system to similar rates of variation in the target, or through informational masking, which encompasses all other forms of interference of central origin (e.g., Durlach *et al.*, 2003; Kidd *et al.*, 2008). Speech is a spectro-temporally sparse signal, so when there are two talkers, energetic masking usually affects only parts of the target speech, limited in both frequency and time (e.g., Cooke, 2006). Consequently, unless the signal-to-noise ratio is poor, separating two voices is mainly a problem of assigning readily detectable frequency-time regions to the correct source rather than one of detecting parts of the target signal (e.g., Darwin, 2008). Indeed, at least in circumstances where there is one competing voice rather than many, the impact of the interferer on the intelligibility of target speech typically arises primarily through informational masking (e.g., Brungart *et al.*, 2006).

Informational masking is an umbrella term for a broad range of effects falling in three classes—failures of object formation, failures of object selection, and capacity limitations on cognitive processing (Shinn-Cunningham, 2008). In principle, acoustic cues facilitating the perceptual separation of target and masker (Bregman, 1990) may lessen any of these kinds of interference. These cues are typically differences between target and masker in basic acoustic properties—e.g., differences in fundamental frequency (F0) can be used to separate formant ensembles (Gardner *et al.*, 1989; Summers *et al.*, 2010). In contrast, recent findings suggest that across-formant grouping is not governed by similarity in the dynamic properties of the formant-frequency contours (Summers *et al.*, 2012; Roberts *et al.*, 2014). Note, however, that the extent of informational masking produced by an interferer depends on its acoustic properties even in the absence of useful segregation cues. For example, extraneous formants with time-varying frequency contours have a greater impact on intelligibility than those with constant-frequency contours (e.g., Roberts *et al.*, 2010, 2014). The experiments reported here explore further the role of the time-varying properties of speech when separating a target voice from an interfering voice under conditions of informational masking.

One approach to investigating the informational component of speech-on-speech masking is the use of a binary

[a]Electronic mail: b.roberts@aston.ac.uk

 0001-4966/2015/137(5)/2726/11/$30.00 

mask to retain all frequency-time regions in the mixture dominated by the target and to eliminate those dominated by the masker. This approach has proved influential and can be used regardless of whether the masker consists of one or several competing voices. The concept of a binary mask has its origins in computational auditory scene analysis (e.g., Brown and Cooke, 1994; Wang and Brown, 1999), in which the aim is to use acoustic grouping cues extracted from the stimulus mixture to inform the construction of the binary mask. In practice, however, most studies have used ideal binary masks (see Wang, 2005) based on prior knowledge of the signals contributing to the mixture (e.g., Cooke *et al.,* 2001; Brungart *et al.,* 2006). Studies of this kind have established, for example, that the effects of target-masker similarity on speech intelligibility arise primarily from informational masking (e.g., same-sex vs different-sex competing voices; Brungart *et al.,* 2009).

Another approach to isolating the informational component of masking is to configure the stimulus so as to minimize energetic masking of the target speech by the interfering speech. Several studies have used the second-formant competitor (F2C) paradigm (Remez *et al.,* 1994; Roberts *et al.,* 2010) and analogues of sentence-length materials to investigate the ability of listeners to attend to a set of target formants in the presence of an extraneous formant. F2C may be considered as an alternative candidate for the second formant, which must be rejected to optimize intelligibility. Central to the F2C paradigm is the presentation of the target F2 and F2C to opposite ears, an arrangement that greatly reduces energetic masking of the target speech by the competitor. This approach has proved fruitful, but there are two design features that have constrained the stimulus manipulations possible and the generality of the conclusions drawn from the results. First, to our knowledge, all previous studies using the F2C paradigm or variants thereof have split the target formants between ears—e.g., left ear = F1+F2C; right ear = F2+F3—so that listeners must integrate phonetic information across ears, as well as frequency, to optimize performance. Given the challenging nature of this task, requiring dichotic integration under competitive conditions, previous studies have allowed participants to listen to each stimulus more than once, typically up to six times, before transcribing it. Second, the presence of F2C in the same ear as F1 limits the extent to which the competitor's properties can be varied across conditions. The current study uses a new version of the F2C paradigm, one adapted to overcome these limitations and to provide a closer approximation to realistic listening conditions.

The adapted F2C paradigm involves presenting all the target formants in the same ear (monaural speech) and the extraneous formant in the opposite ear. This arrangement avoids the need to integrate information across ears and completely eliminates energetic masking of the target formants by the extraneous formant. The new version also uses one-shot trials (a single stimulus presentation on each trial) with random allocation of the target speech to the left or right ear. The lack of opportunity for repeat listening further increases the ecological validity of the approach and the uncertainty from trial to trial about the lateralization of the target speech prevents listeners from attending selectively to one ear, increasing the extent of informational masking (see Kidd *et al.,* 2008). Finally, the isolation of the extraneous formant in the contralateral ear removes constraints on the frequency range over which F2C can vary, which is of particular relevance to the design of experiment 2. Here, we report two experiments using the adapted method to examine further the effects of formant frequency and amplitude variation on the informational component of speech-on-speech masking. The results of these experiments confirm and extend those of the earlier studies and increase the generality of their conclusions.

## II. EXPERIMENT 1

Recent research using sentence-length analogues for which the target formants are presented dichotically suggests that the ability of an extraneous formant to impair speech intelligibility depends on the variation of its frequency contour, but not its amplitude contour. This has been reported both for sine-wave analogues (Roberts *et al.,* 2010) and synthetic-formant analogues of speech (Summers *et al.,* 2012). Experiment 1 examined whether competitor impact is influenced similarly by the frequency and amplitude contours of F2C when all three target formants are presented in the same ear and F2C is in the opposite ear, such that any interference cannot occur through energetic masking.

### A. Method

#### 1. Listeners

All volunteers were students or members of staff at Aston University who received either course credit or cash for taking part. They were first tested using a screening audiometer (Interacoustics AS208, Assens, Denmark) 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 the speech analogues used (see Sec. II A 3). About two thirds of these volunteers completed the training successfully and took part in the main experiment. With one exception (who was replaced), all of these listeners met the additional criterion of a mean score of $\geq$ 20% keywords correct, when collapsed across all conditions in the main experiment, and so their results were included in the final dataset. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. Twenty-seven listeners (five males) successfully completed the experiment (mean age = 19.9 yr, range = 18.3–30.1). To our knowledge, none of the listeners had heard any of the sentences used in the main experiment in any previous study or assessment of their speech perception. All were native speakers of English and gave informed consent. The research was approved by the Aston University Ethics Committee.

#### 2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of a collection of sentences spoken by a British

male talker of "Received Pronunciation" English. The text for these sentences was provided by Patel and Morse (personal communication) and consisted of variants created by rearranging words from the Bamford–Kowal–Bench (BKB) sentence lists (Bench *et al.*, 1979). To enhance the intelligibility of the synthetic analogues, the 54 sentences used were semantically simple and selected to contain ≤25% phonemes involving vocal tract closures or unvoiced frication. A set of keywords was chosen for each sentence; most designated keywords were content words. The stimuli for the training session were derived from 50 sentences spoken by a different talker and taken from commercially available recordings of the Harvard sentence lists (IEEE, 1969). These sentences were also selected to contain ≤25% phonemes involving closures or unvoiced frication.

For each sentence, the frequency contours of the first three formants were estimated from the waveform automatically every 1 ms from a 25-ms-long Gaussian window, using custom scripts in Praat (Boersma and Weenink, 2010). In practice, the third-formant contour often corresponded to the fricative formant rather than F3 during phonetic segments with frication; these cases were not treated as errors. Gross errors in automatic estimates of the three formant frequencies were hand-corrected using a graphics tablet; artifacts are not uncommon and manual post-processing of the extracted formant tracks is often necessary (Remez *et al.*, 2011). Amplitude contours corresponding to the corrected formant frequencies were extracted automatically from the stimulus spectrograms; these contours were used to generate synthetic analogues of each sentence.

Synthetic-formant analogues of each sentence were created using the extracted frequency and amplitude contours to control three parallel second-order resonators whose outputs were summed. Following Klatt (1980), the outputs of the resonators corresponding to F1, F2, and F3 were summed using alternating signs (+, –, +) to minimize spectral notches between adjacent formants in the same ear. A monotonous periodic source (F0 = 140 Hz) was used in the synthesis of all stimuli used in the training and main experiment. The excitation source was a periodic train of simple excitation pulses modeled on the glottal waveform, which Rosenberg (1971) has shown to be capable of producing synthetic speech of good quality. 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. Stimuli were selected such that the frequency of the target F2 was always ≥80 Hz from the frequencies of F1 and F3 at any moment in time. Hence, there were no approaches between formant tracks close enough to cause audible interactions between corresponding harmonics exciting adjacent formants.

For each sentence used in the main experiment, a set of competitors was created by various manipulations of the frequency and amplitude contours of F2. The frequency contour of F2C could be time reversed ($f_R$), inverted about its geometric mean ($f_I$), or constant at its geometric mean ($f_C$). The amplitude contour could be time reversed ($a_R$), time forward (i.e., normal, $a_N$), or constant at a value that preserved the root mean square (RMS) power ($a_C$). The set of contours used to construct the variants of F2C is illustrated for an example sentence in Fig. 1. To keep the experiment within acceptable bounds, not every possible combination of the available frequency and amplitude contours was used. All competitors were rendered as the outputs of a second-order resonator. The excitation source (Rosenberg pulses), F0 frequency (140 Hz), 3-dB bandwidth (70 Hz), and output sign (–) were identical to those used to synthesize the target F2. Note that instances where time-reversed frequency and/or amplitude contours were used did not involve time reversal of the excitation source for F2C. When present, F2C was always sent to the ear contralateral to that receiving the target formants.

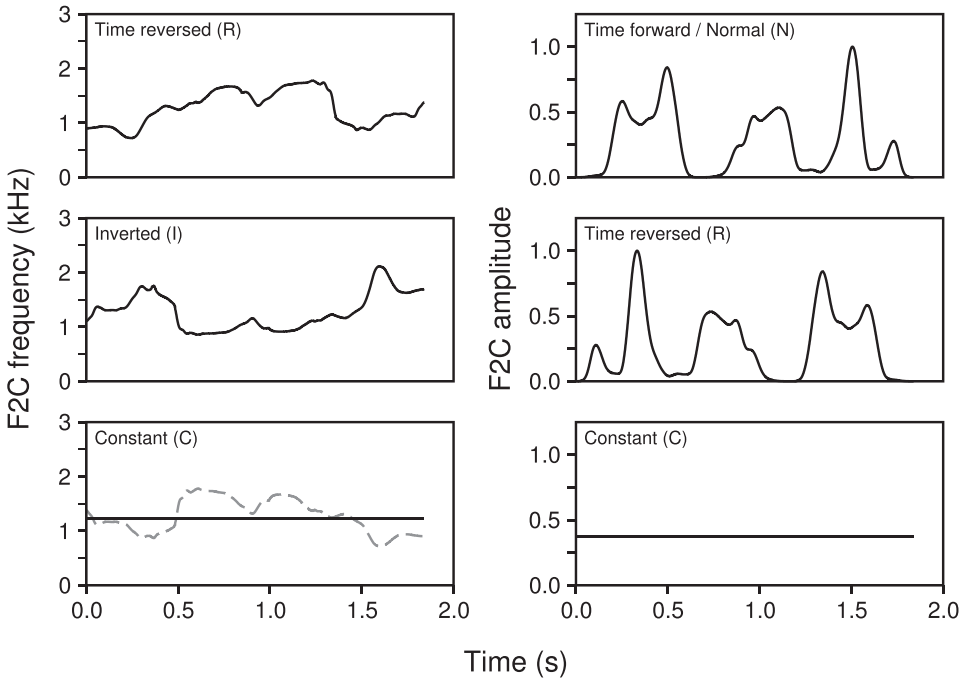


FIG. 1. Stimuli for experiment 1—frequency and amplitude contours for the different competitors (F2Cs) added to the synthetic-formant analogue of the example sentence "Her long hair is brown." The left- and right-hand panels show, respectively, the set of frequency and amplitude contours for F2C derived from F2. For reference, the frequency contour of the target F2 is included in the bottom-left panel (dashed gray line). Amplitude contours are shown normalized to the maximum value in the original F2 contour. Relative to the target F2, the F2C frequency contour was time reversed ($f_R$), inverted about the geometric mean frequency ($f_I$), or constant at the geometric mean frequency ($f_C$). The F2C amplitude contour was time reversed ($a_R$), time forward (i.e., normal, $a_N$), or constant at a value preserving the RMS power ($a_C$).

There were nine conditions in the main experiment (see Table I). Two conditions (C1 and C2) were controls, for which the target F2 was absent. The stimuli for C1 comprised F1 and F3 only. The stimuli for C2 also contained F2C; its parameters ($f_I$, $a_N$) were chosen as representative of cases where the competitor has time-varying contours. Six conditions (C3−C8) were experimental cases, for which the stimuli contained the target F2 and an F2C with one of the six pre-selected combinations of frequency and amplitude contours, including time-varying and constant cases. The final condition (C9) was the reference case, for which only the monaural target formants were presented. For each listener, the 54 sentences were divided equally across conditions (i.e., six per condition), such that there were always 18 or 19 keywords per condition. Allocation of sentences was counterbalanced by rotation across each set of nine listeners tested. Hence, the total number of listeners needed to produce a balanced dataset was a multiple of nine.

### 3. Procedure

During testing, listeners were seated in front of a computer screen and a keyboard in a sound-attenuating chamber (Industrial Acoustics 1201A; Winchester, UK). The experiment consisted of a training session followed by the main session and typically took about 50 min to complete; listeners were free to take a break whenever they wished. In both parts of the experiment, stimuli were presented in a new quasi-random order for each listener.

The training session comprised 50 trials; stimuli were presented without competitors and a new sentence was used for each trial. On each of the first 10 trials, participants heard diotic presentations of the synthetic version (degraded, D) and the original recording (clear, C) of a sentence in the order DCDCD; no response was required but participants were asked to listen to these sequences carefully. On each of the next 30 trials, listeners heard a diotic presentation of the synthetic version of a sentence, which they were asked to transcribe using the keyboard. 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 was provided by playing the original recording (44.1 kHz sample rate) followed by a repeat of the synthetic version. Davis *et al.* (2005) found this strategy to be an efficient way of enhancing the perceptual learning of speech analogues.

During the final 10 training trials, the sentence analogue was delivered monaurally; the ear receiving it was selected randomly on each trial. Listeners heard the stimulus only once before entering their transcription. Feedback was provided as before, in this case with the original and synthetic versions delivered only to the selected ear. Listeners continued on to the main session if they met either or both of two criteria: (1) ≥ 50% keywords correct across all 40 trials needing a transcription (30 trials = diotic with repeat listening; 10 trials = monaural, random selection of ear, no repeat listening); (2) ≥ 50% keywords correct for the final 15 diotic-with-repeat-listening trials. On each trial in the main experiment, the ear receiving the target formants (F1 + F2 + F3 or F1 + F3) was selected randomly; F2C (when present) was always presented in the other ear. Listeners were allowed to hear each stimulus once only before entering their transcription. No feedback was given.

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 (Hannover, Germany) via a Sound Blaster X-Fi HD sound card (Creative Technology, Singapore), programmable attenuators (Tucker-Davis Technologies PA5; Alachua, FL), and a headphone buffer (TDT HB7). Output levels were calibrated using a sound-level meter (Brüel and Kjaer, type 2209; Nærum, Denmark) coupled to the earphones by an artificial ear (type 4153). All target sentences were presented at a long-term average of 75 dB sound pressure level (SPL); there was some variation in the sound level at the ear receiving F2C (mean ≈ 65 dB SPL), depending on the RMS power of the corresponding F2. In the training session, the presentation level of the diotic materials (first 40 target sentences plus original recordings) was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation. The last 10 sentences in the training session were presented monaurally at the reference level.

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 target F2. The frequency contour could be time reversed ($f_R$), inverted about the geometric mean of F2 ($f_I$), or constant at the geometric mean of F2 ($f_C$). The amplitude contour could be time reversed ($a_R$), time forward (i.e., normal, $a_N$), or constant at a value that preserved the RMS power ($a_C$).

| Condition | Stimulus configuration (target ear, other ear) | F2C frequency ($f$) and amplitude ($a$) contours |
|---|---|---|
| C1 | (F1+F3; –) | – |
| C2 | (F1+F3; F2C) | $f_I$, $a_N$ |
| C3 | (F1+F2+F3; F2C) | $f_I$, $a_N$ |
| C4 | (F1+F2+F3; F2C) | $f_I$, $a_C$ |
| C5 | (F1+F2+F3; F2C) | $f_R$, $a_R$ |
| C6 | (F1+F2+F3; F2C) | $f_R$, $a_C$ |
| C7 | (F1+F2+F3; F2C) | $f_C$, $a_N$ |
| C8 | (F1+F2+F3; F2C) | $f_C$, $a_C$ |
| C9 | (F1+F2+F3; –) | – |

### 4. Data analysis

For each listener, the intelligibility of each stimulus was quantified in terms of the percentage of keywords identified correctly; homonyms were accepted. The stimuli for each condition comprised six sentences. Given the variable number of keywords per sentence (2–4), the mean score for each listener in each condition was computed as the percentage of keywords reported correctly giving equal weight to all the keywords used. As in our previous studies (Roberts *et al.*, 2010, 2014, 2015; Summers *et al.*, 2010, 2012), 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). All statistical analyses were computed using

SPSS (SPSS statistics version 20, IBM Corp.). The measure of effect size reported here is partial eta squared ($\eta_p^2$).

### B. Results

Figure 2 shows the mean percentage scores (and intersubject standard errors) across conditions in terms of keywords identified correctly. The white, gray, and black bars indicate the results for the control, experimental, and target-only reference conditions, respectively; within the experimental conditions, dark and light gray bars indicate the results for cases with time-varying and constant *amplitude* contours, respectively. A one-way within-subjects analysis of variance (ANOVA) over all nine conditions showed a highly significant effect of condition on intelligibility [$F(8,208) = 44.763$, $p < 0.001$, $\eta^2_p = 0.633$].[1] All pairwise comparisons (two tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967). The control conditions indicated that intelligibility was reduced substantially in the absence of the target F2 (C1) and was near floor when F2 was replaced by F2C (C2). Keyword scores for C1 and C2 were significantly different from those for all other conditions and from each other

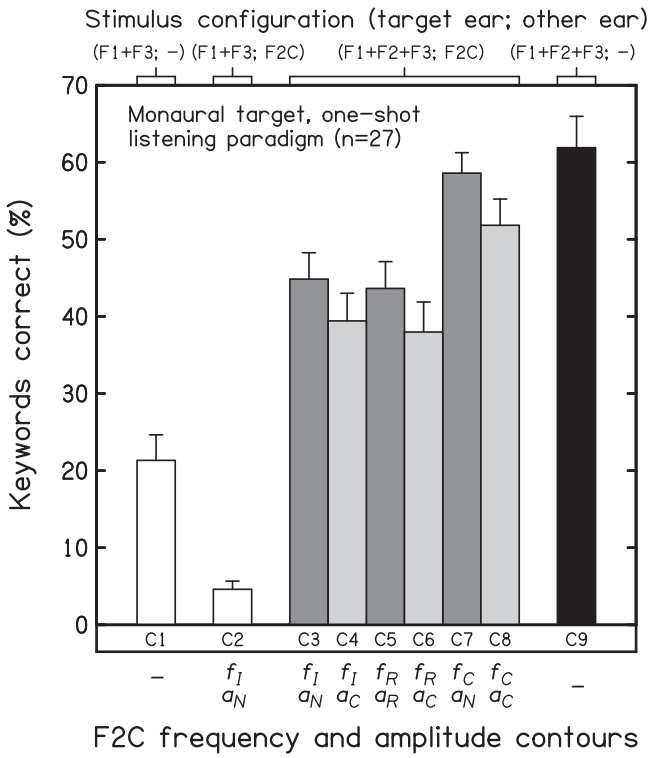


FIG. 2. Results for experiment 1—influence of frequency and amplitude contour on the effect of competitors (F2Cs) on the intelligibility of formant analogues of the target sentences. Mean keyword scores and intersubject standard errors ($n = 27$) are shown for the control conditions (white bars), experimental conditions (gray bars), and target-only 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). For ease of reference, condition numbers are included immediately above the bottom axis. Relative to the target F2, the F2C frequency contour was time reversed ($f_R$), inverted about the geometric mean frequency ($f_I$), or constant at the geometric mean frequency ($f_C$). The F2C amplitude contour was time reversed ($a_R$), time forward (i.e., normal, $a_N$), or constant at a value preserving the RMS power ($a_C$).

($p < 0.001$ in all cases). With one exception (C7 vs C9, $p = 0.343$), intelligibility was significantly lower when the monaural target was accompanied by a contralateral competitor (range: $p = 0.011 - p < 0.001$).

The effect of F2C properties on competitor impact was explored using a two-way within-subjects ANOVA restricted to the set of experimental conditions (C3–C8). The two factors were frequency contour (three levels: inverted, time reversed, or constant) and amplitude contour (two levels: dynamic or constant). This analysis revealed significant main effects of frequency contour [$F(2,52) = 19.255$, $p < 0.001$, $\eta^2_p = 0.425$] and amplitude contour [$F(1,26) = 7.729$, $p = 0.010$, $\eta^2_p = 0.229$], but no interaction between them [$F(2,52) = 0.035$, $p = 0.966$]. The primary outcome was that competitors with either type of time-varying frequency contour (inverted or time reversed) were significantly more effective than those with constant frequency contours ($p < 0.001$ in both cases); which dynamic contour was used made no difference ($f_I$ vs $f_R$, $p = 0.568$). Relative to the target-only reference condition (C9), competitors with time-varying and constant frequency contours caused scores to fall on average by 20.4 and 6.7 percentage points, respectively, corresponding to a difference of 13.7 percentage points. There was also a smaller and additive effect of amplitude contour, such that constant amplitude contours were significantly more effective than dynamic ones (time forward or reversed). Relative to the reference condition (C9), competitors with time-varying and constant amplitude contours caused scores to fall on average by 12.9 and 18.8 percentage points, respectively, corresponding to a difference of 5.9 percentage points.

### C. Discussion

Despite the differences from earlier studies, in which the target formants were split between ears and repeat listening was permitted, the adapted method was effective at distinguishing the relative impacts of different F2Cs on performance. Keyword intelligibility is typically reduced when monaural speech is accompanied by an extraneous formant in the contralateral ear; this interference cannot arise from energetic masking. The results indicate that competitor impact depends primarily on the dynamic properties of the F2C frequency contour—competitors with time-varying frequency contours, whether derived from F2 by spectral inversion or time reversal, have a much greater effect on intelligibility than competitors with constant frequency contours. This outcome is in accord with earlier findings for similar materials using configurations where the target formants were presented dichotically and energetic masking was controlled but not eliminated completely (Roberts *et al.*, 2014; Summers *et al.*, 2010, 2012). In addition, the magnitudes of the different impacts on keyword scores are broadly similar to those reported previously. While it is acknowledged that fusion by common F0 between target speech and competitor may have contributed to the overall extent of dichotic interference observed here (cf. Summers *et al.*, 2010), the results clearly indicate that frequency variation in

an extraneous formant is a major factor governing the extent of informational masking that it produces.

A novel aspect of the results is that F2C amplitude contour makes a small but significant contribution to competitor impact in this context; the effect was about half the size of that observed for F2C frequency contour and was additive. Previous studies using dichotic configurations of target formants found that whether the amplitude contour of an extraneous formant was time-varying or constant (matched for RMS power) had no effect whatsoever on competitor impact, either for sine-wave (Roberts *et al.,* 2010) or synthetic-formant analogues of speech (Summers *et al.,* 2012). Somewhat surprisingly, the new findings indicate that a competitor with a constant amplitude contour ($a_C$) has more impact on intelligibility than one with a time-varying amplitude contour, whether time forward ($a_N$) or reversed ($a_R$). It is not clear why this is the case, but one possibility is that constant-amplitude competitors tend to draw attention to the ear receiving them. This could occur because they are more salient than the target formants (subjectively, a constant-amplitude F2C tends to stand out against the time-varying target formants) or because they become audible earlier (constant-amplitude F2Cs reach maximum after only 10 ms, whereas the target formants may remain at low amplitude for substantially longer). A change in the balance of spatial attention away from the ear receiving the monaural speech might plausibly lower intelligibility. However, such a change is likely to have little or no effect when the target formants are distributed across both ears, particularly when repeat listening is permitted as was the case in the previous studies.

## III. EXPERIMENT 2

A dynamic property of the speech signal carrying critical phonetic information is the velocity of formant-frequency change, which is affected both by the rate and depth of formant-frequency variation. Rate and depth are associated with speech rate (syllables/s) and the extent of movements of the articulators, respectively (e.g., Lindblom and Sundberg, 1971; Weismer and Berry, 2003). Recent research suggests that increasing either the rate or depth of formant-frequency variation in a competitor increases its impact on intelligibility, but that differences in these properties between the target and interfering formants do not provide a basis for their perceptual segregation (Summers *et al.,* 2012; Roberts *et al.,* 2014). Also, it does not seem to matter whether the pattern of this variation is plausibly speech-like (inverted F2 frequency) or not (triangle wave). Roberts *et al.* (2014) concluded that target-masker similarity in these dynamic properties is not important for the segregation or selection of a subset of formants from an ensemble because there was no evidence of a maximum in interference when the depth of formant-frequency variation for F2C matched that for the target formants. Rather than any evidence of tuning in this variable, interference simply increased as the average depth of formant-frequency variation in the competitor increased, suggesting that larger frequency variations in F2C have a greater effect on the extraction of phonetic information from the target formants.

The experiments reported by Roberts *et al.* (2014) involved presenting F1 and the competitor in the same ear (F1 + F2C; F2 + F3), and so depth of formant-frequency variation in the competitor could not be increased above 100% without F2C approaching or crossing the track of the target F1. Even within this limit, greater masking between these formants cannot be ruled out as F2C depth is increased, because this configuration controlled but did not completely eliminate energetic masking. The only way to test a range of depths for F2C above and below that for the target F2 without violating this constraint was to scale down the frequency variation in all the target formants to 50% (this manipulation had little impact on intelligibility in the absence of F2C). Therefore, it is possible that the apparent absence of tuning for depth of F2C frequency variation was an artifact of this constraint. In particular, note that the greatest depth used for F2C (100%) corresponds to the original depth for the target F2 in the natural utterances. Hence, it is possible that interference is maximal not when the depth of frequency variation in F2C matches that for the rescaled target F2, but when it matches the original depth. The adapted method allows substantially higher scale factors to be applied to the formant-frequency variation in the extraneous formant and so experiment 2 addressed this possibility by presenting the target speech at 100% depth and the F2C in the contralateral ear at depths ranging from 0% (constant) to twice the natural depth (200%).

### A. Method

Except where described, the same method was used as for experiment 1. Twenty-seven listeners (10 males) passed the training and successfully completed the experiment (mean age = 29.9 yr, range = 19.9–48.7). The training session was identical to that used in experiment 1. The stimuli for the main experiment were derived from the same set of 54 BKB-like sentences and were allocated to conditions in the same way. Consequently, none of the listeners who took part in experiment 1 were eligible to take part in this experiment.

All stimuli were generated using the same excitation source (Rosenberg pulses), F0 frequency (140 Hz), and resonator bandwidths as for experiment 1. A set of F2 competitors was created for each sentence in the main experiment. The frequency contour of each F2C was created by inverting the frequency contour of the target F2 about its geometric mean and applying a range of scale factors to adjust the depth of formant-frequency variation in the competitor. In all cases, the amplitude contour was identical to that used for the target F2; note that the effect of amplitude contour observed in experiment 1 was additive and hence the specific choice made should not influence the effect of formant-frequency change. Inversion and rescaling of F2C frequency contours was performed on a log-frequency scale. Each F2 contour was converted to a vector specifying, frame by frame, the frequency as a deviation from the geometric mean frequency of the whole track. Contour inversion was achieved by flipping the sign of each element in the vector.

TABLE II. Stimulus properties for the conditions used in experiment 2 (main session). The frequency contour of the competitor (F2C), when present, was inverted. The scale factor for F2C refers to the depth of variation in formant frequency, relative to that for the unscaled target F2. A scale factor of 0% indicates a constant frequency contour for F2C, corresponding to the geometric mean frequency of the target F2. The same amplitude contour was used for F2C as for the target F2 (i.e., F2C type $= f_I, a_N$). Hence, when 100% scaling was used (C5), the stimuli were identical to those used in C3 for experiment 1.

| Condition | Stimulus configuration (target ear, other ear) | F2C scale factor (%) relative to target F2 |
|---|---|---|
| C1 | (F1+F3; F2C) | 100 |
| C2 | (F1+F2+F3; F2C) | 0 |
| C3 | (F1+F2+F3; F2C) | 50 |
| C4 | (F1+F2+F3; F2C) | 75 |
| C5 | (F1+F2+F3; F2C) | 100 |
| C6 | (F1+F2+F3; F2C) | 125 |
| C7 | (F1+F2+F3; F2C) | 150 |
| C8 | (F1+F2+F3; F2C) | 200 |
| C9 | (F1+F2+F3; –) | – |

The depth of frequency variation around the geometric mean was then adjusted by multiplying the vector using a scale factor in the range 0 (i.e., constant at the geometric mean frequency) to 2 (i.e., twice the original depth). Scale factors $< 1$ compress the depth of F2C frequency variation, which has the effect of reducing the extent and velocity of formant-frequency change (the "formant squash" manipulation; Roberts *et al.*, 2014); scale factors $> 1$ expand the depth of frequency variation. In formal terms, the rescaled frequency for each formant at time $t$, $s(t)$, is given by

$$\log s(t) = \log g + x\left(\log \frac{f(t)}{g}\right), \quad (1)$$

where $x$ ($0 \leq x \leq 2$) is a proportional scale factor determining the maximum possible frequency range (depth of variation), $f(t)$ is the formant frequency at time $t$, and $g$ is the geometric mean of the whole formant-frequency contour. The frequency contours of the three target formants were not adjusted for depth of frequency variation.

There were nine conditions in the main experiment (see Table II). One condition (C1) was a control, for which F2C was present (100% depth) but the target F2 was absent. Seven conditions (C2−C8) were experimental, for which the target formants were accompanied by F2C in the contralateral ear. Across this set of conditions, the depth of variation in the F2C frequency contour around its geometric mean was scaled to 0%, 50%, 75%, 100%, 125%, 150%, and 200%. The final condition (C9) was the reference case, for which only the target formants were presented (i.e., the no-F2C case). The range of stimuli for the experimental conditions is illustrated in Fig. 3 using the narrowband spectrogram of a synthetic analogue of an example sentence accompanied by an F2C whose frequency contour is scaled to 0%, 100%, or 200% (top, middle, and bottom right-hand panels, respectively).

### B. Results

Figure 4 shows the mean keyword scores (and intersubject standard errors) for the control condition (C1, asterisk), experimental conditions (C2−C8, filled circles), and reference condition (C9, open circle). The mean scores for the

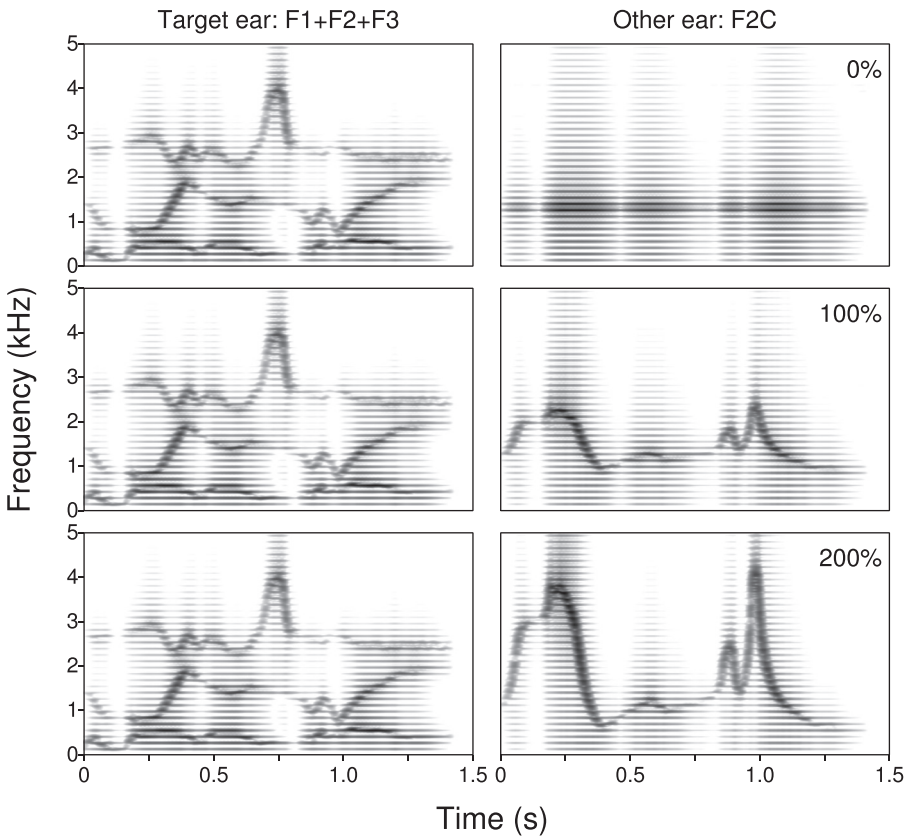


FIG. 3. Stimuli for experiment 2—narrowband spectrograms of a synthetic-formant analogue (F0 = 140 Hz) of the example sentence "The boy knows the way" (left panels) accompanied in the contralateral ear by one of three variants of a competitor (F2C) scaled to different depths of formant-frequency variation—0% (constant, top right), 100% (baseline, middle right), and 200% (maximum, bottom right). The frequency contour of F2C was created by inverting the F2 frequency contour about its geometric mean and scaling it as indicated. The amplitude contour of F2C was the same as that of the target F2. Note the wide frequency excursions made by F2C at 200% scaling.

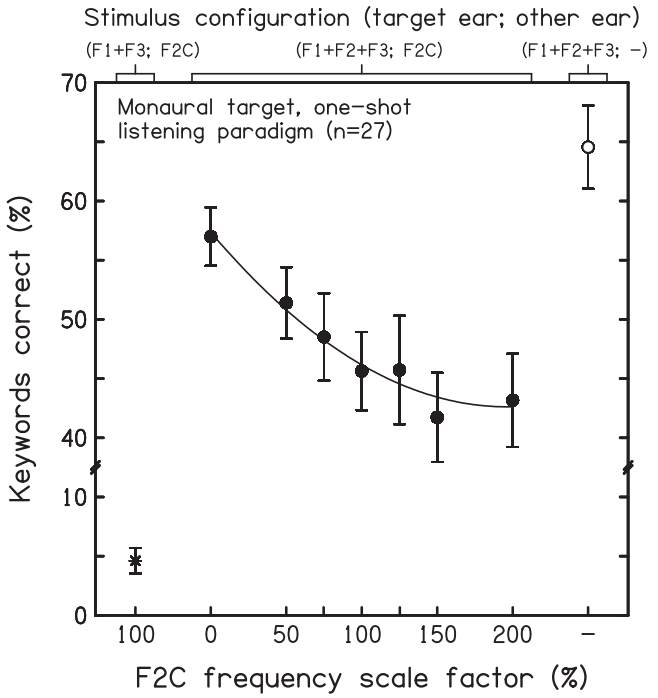


FIG. 4. Results for experiment 2—influence of the depth of formant-frequency variation in competitor formants (F2Cs) on the intelligibility of formant analogues of the target sentences. The frequency contour of F2C was created by inverting the F2 frequency contour about its geometric mean and scaling its depth to 0% (constant), 50%, 75%, 100% (baseline), 125%, 150%, or 200% (maximum). In all cases, the amplitude contour of F2C was the same as that used for the target F2. Mean keyword scores and intersubject standard errors ($n = 27$) are shown for the control condition (asterisk), experimental conditions (filled circles), and the target-only reference condition (open circle). A quadratic function was used to generate the curve fitted to the mean scores for the seven experimental conditions. The top axis indicates which formants were presented to each ear; the bottom axis indicates the scale factor controlling the depth of formant-frequency variation in F2C (when present).

experimental conditions have been fitted using a quadratic function describing the influence of depth of formant-frequency variation for F2C on the intelligibility of the target sentences. A one-way within-subjects ANOVA over all nine conditions showed a highly significant effect of condition on intelligibility [$F(8,208) = 41.566$, $p < 0.001$, $\eta^2_p = 0.615$].[1] The control condition indicated that intelligibility was near floor when F2C was present and the target F2 was absent (C1 vs C2 – C9: $p < 0.001$ in all cases). With the exception of the 0%-depth case (C2 vs C9, $p = 0.065$), all competitors had a significant impact on intelligibility (range: $p = 0.005$ – $p < 0.001$).

The effect of depth of formant-frequency variation for F2C on intelligibility was explored using a one-way ANOVA restricted to the set of experimental conditions (C2 to C8); this effect was highly significant [$F(6,156) = 4.242$, $p = 0.001$, $\eta^2_p = 0.140$]. Pairwise comparisons showed that the 0%-depth case was significantly different from each of the 75%- to 200%-depth cases (range: $p = 0.022$ – $p < 0.001$), and that the 200%-depth case was significantly different from the 0%- ($p = 0.001$) and 50%-depth ($p = 0.01$) cases, but not from the 100%-depth case ($p = 0.391$). Relative to the reference condition (C9), the inclusion of contralateral competitors with inverted frequency contours caused scores to fall by 7.6, 18.9, and 21.4 percentage points, for depths of 0% (C2), 100% (C5), and 200% (C8), respectively.

### C. Discussion

Once again, adding an extraneous formant in the ear contralateral to the monaural target speech reduced keyword intelligibility. The impact of F2C on intelligibility was least for constant-frequency F2Cs and increased up to ~100% depth, leveling off thereafter. Qualitative signs consistent with this pattern are evident in our earlier results (Roberts *et al.*, 2014) as F2C depth approaches 100% (the maximum tested in that study). Note that there is no sign of a minimum or inflection in keyword scores for the 100% depth case here, as would be expected if similarity in dynamic properties influences across-formant integration of phonetic information. The results confirm and extend those from experiments involving dichotic presentation of the target formants (Roberts *et al.*, 2014). Namely, it is the overall extent of variation in the formant-frequency contour of F2C, not the extent relative to that of the target formants, which governs competitor impact. Clearly, the absence of tuning reported by Roberts *et al.* (2014) is not an artifact of the constraints limiting the extent of F2C frequency variation in that study to the natural depth (100%).

The results obtained here show that an extraneous formant whose depth of frequency variation exceeds that of the formants in the natural utterance cannot be rejected more easily from the ensemble based on the mismatch in time-varying properties. This is consistent with the notion that, unlike qualitative differences in simple acoustic properties between target and masker (e.g., tonal vs noisy stimuli; Neff, 1995), differences in their dynamic properties cannot provide a basis for their concurrent segregation (Roberts *et al.*, 2014). Precisely why the extent of informational masking produced by F2C saturates once the depth of formant-frequency variation exceeds the natural range for the corresponding target F2 remains to be established; indeed, there may be more than one aspect of the natural range that is relevant in this context. Factors that merit consideration in future research include the range of formant-frequency variation on a log scale, the extent of overlap between the ranges for F2C and the target F2, and the mean or maximum velocity of formant transitions. The number of formants comprising the interferer, and the degree of correlation in across-formant change for multi-formant interferers, may also be important factors.

## IV. GENERAL DISCUSSION

The results of the experiments reported here support and extend the findings of earlier studies that the ability of an extraneous formant to impair intelligibility is critically dependent on the variation of its frequency contour (Roberts *et al.*, 2010, 2014; Summers *et al.*, 2012). In particular, frequency variation in the interferer remains important in circumstances where no across-ear fusion of target formants is required (monaural speech), the extraneous formant cannot act as an energetic masker (interferer contralateral to the target), and there is only one opportunity to listen to the stimulus, with no

prior knowledge of which ear to attend. This outcome is consistent with evidence that listeners attending to the quieter of two speech signals presented concurrently in one ear are highly susceptible to interference from normal or time-reversed speech presented in the other ear (Brungart *et al.,* 2005). The adapted task used here represents a closer approximation to realistic listening conditions than has been achieved before with the F2C paradigm, but note that there are nonetheless circumstances in which the dichotic-target configurations used previously—(F1+F3; F2) and (F1; F2+F3)—would be advantageous experimentally. Specifically, the peripheral isolation of F2 or F1 offered by the former and latter variants, respectively, is better when one wishes to examine the effects of manipulating the properties of these target formants rather than those of the interferer.

Roberts *et al.* (2014) demonstrated, using dichotic-target configurations, that increasing the depth of formant-frequency variation in F2C within the range 0%–100% simply increases its impact on intelligibility; this outcome was not influenced by whether or not the depth of formant-frequency variation in the target and interferer was the same. These results were interpreted as evidence against a grouping constraint based on target-masker similarity in this complex dynamic property, which contrasts with evidence that target-masker similarity is an important organizational property in the context of simpler acoustic properties, such as differences in F0 (e.g., Summers *et al.,* 2010) and onset time (e.g., Darwin, 1981). Rather, it was concluded that the extraneous formant more effectively corrupts or disrupts extraction of the phonetic properties of the target speech as the extent of frequency variation in that formant increases.

The results of experiment 2 strengthen this interpretation for two reasons. First, the complete elimination of energetic masking rules out an account based on increased partial masking interactions between the target and competitor as the depth of F2C frequency variation increases. Second, the pattern of results was maintained despite the change in target depth from 50% to 100%; this would not be the case if F2C impact were tuned to target-masker similarity in this property. While the results reported here show that an extraneous formant whose range of frequency variation exceeds that of the natural utterance cannot be rejected more easily from the formant ensemble, raising the upper limit on scaling to 200% has revealed that increasing F2C depth beyond 100% has little or no additional effect on the informational masking it produces. Establishing why the effect on intelligibility of increasing F2C depth levels off will require further investigation.

Speech, like many other environmental sounds, has peaks and valleys in intensity as a function of time for which the intensity trajectories show a high degree of correlation across frequency. The phenomenon of comodulation masking release (Hall *et al.,* 1984) suggests that listeners should be able to use coherent envelope fluctuations as a means of grouping together acoustic elements from a common source and segregating competing sound sources. However, the results of the current study are contrary to this proposal. Similar to recent observations with sine-wave speech (Roberts *et al.,* 2010), generating an F2C for these buzz-excited analogues using either the time-forward ($a_N$) or time-reversed ($a_R$) amplitude contour for F2 produced equally effective competitors. This indicates that the impact of F2C on intelligibility is not affected by the correlation of its amplitude contour with that of F2 and the other target formants. Note, however, that this outcome is consistent with evidence that the increased intelligibility associated with applying high-rate amplitude modulation (AM) to a sine-wave speech stimulus does not depend on whether the AM is coherent or conflicting across formants (Lewis and Carrell, 2007).

Although the dichotic configuration used in the current study precludes a contribution of energetic masking to the interference observed, it is not necessarily the case that the extraneous formant acts purely as an informational masker. There is evidence that amplitude variations in a masker can interfere with the processing of modulations in target speech (e.g., Stone *et al.,* 2011; Stone *et al.,* 2012). This outcome is an example of modulation masking, and models for predicting speech intelligibility based on modulation masking can be quite successful (e.g., Dubbelboer and Houtgast, 2008; Jørgensen and Dau, 2011). In the current study, note that frequency variation in the masker leads to within-channel envelope variation even for the constant-amplitude case. Nonetheless, the extent to which modulation masking contributes to the results reported here remains unclear. Although there is evidence that contralateral maskers can cause modulation detection interference, the magnitude of this effect is small compared with within-ear effects (Bacon and Opie, 1994; Lyzenga and Carlyon, 2000). Also, we are not aware of any studies in which envelope variations in a contralateral masker have been shown to interfere with the processing of modulations in a target stimulus when those modulations are substantially supra-threshold, as was the case for the target speech in the current study. Hence, we contend that a contralateral extraneous formant acts primarily as an informational masker, but acknowledge that a contribution from modulation masking cannot be ruled out.

An interesting difference from previous studies, which used dichotic-target stimuli, is the finding that constant-amplitude competitors are more effective than those with time-varying amplitude. The size of this effect is about half that observed when comparing time-varying with constant F2C frequency contours and is in the opposite direction. Owing to the nature of simple parallel synthesis with second-order, unity DC-gain resonators, there are inevitably some changes in formant amplitude over time associated with changes in formant frequency for F2Cs with $a_C$ contours, but fortunately these changes are small compared with those associated with the alternation between more open and more closed vocal-tract configurations in speech production. Hence, we can be confident that our estimate of the size of the amplitude effect is a reasonable one. Although not conclusive, it is worth noting that the greater impact of constant-amplitude competitors is the opposite outcome to what one would predict if modulation masking were a major contributor to the interference observed.

The reason for the greater efficacy of competitors with $a_C$ contours observed here is unclear but, as suggested earlier, it

may arise from differences in how attention is divided between ears when listening to formant ensembles containing monaural or dichotic targets. Note that any effect of amplitude is certainly not mediated by target-masker similarity, as all the target formants have time-varying amplitudes, and so similarity in the extent of amplitude variation would predict the opposite outcome. Consistent with the finding from experiment 1 that the effect of F2C amplitude contour is additive, the change from using constant-amplitude F2Cs (Roberts *et al.*, 2014) to F2Cs with time-forward amplitude contours in experiment 2 does not appear to have modulated the effect of changes in F2C frequency variation over the range 0% to 100% depth. Finally, it is also worth noting that the results obtained for F2Cs with different amplitude contours support the idea that the effect of the frequency sweeps in time-varying F2Cs is likely to be a direct consequence of formant-frequency variation, not a result of within-channel AM.

In conclusion, the adapted version of the F2C paradigm introduced here offers a useful experimental tool for investigating further the informational masking of speech by extraneous formants. The results of the experiments reported here demonstrate that there are circumstances in which the intelligibility of monaural speech can be reduced substantially by a contralateral interferer. The findings also provide further support for the proposal that it is the overall extent of formant-frequency variation in a competitor, not the extent relative to that of the target formants, which governs its impact on intelligibility. The effect of competitor amplitude contour is less well understood, but does not depend on across-formant correlation in amplitude variations or target-masker similarity in the extent of those variations. Elucidating the extent to which competitor impact is specific—e.g., the intrusion of competitor properties into perceptual estimates of target properties (cf. Porter and Whitaker, 1980)—or non-specific—e.g., capacity limitations for increased perceptual load (cf. Mattys *et al.*, 2012)—is an important challenge for future research.

## ACKNOWLEDGMENTS

This research was supported by Research Grant ES/K004905/1 from the Economic and Social Research Council (UK), awarded to B.R. (ORCID: 0000-0002-4232-9459). To access the research data underlying this publication, see http://dx.doi.org/10.17036/Roberts_20150427_A01. We are grateful to Peter Bailey for his advice concerning this research and for his comments on drafts of our paper, and to Brian Moore for his comments on this article. We also thank Rob Morse and Meghna Patel for providing the BKB-like sentences and Quentin Summerfield for enunciating them. A poster presentation on part of this research was given at the Annual Conference of the British Society of Audiology (Keele, UK, September 2014).

[1]As a precaution, given the low scores obtained in the control condition(s), all ANOVAs were 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 analyses; applying the transform did not change any of the comparisons reported from significant to non-significant or vice versa.

Bacon, S. P., and Opie, J. M. (**1994**). "Monotic and dichotic modulation detection interference in practiced and unpracticed subjects," J. Acoust. Soc. Am. **95**, 2637–2641.
Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol. **13**, 108–112.
Boersma, P., and Weenink, D. (**2010**). "PRAAT, a system for doing phonetics by computer, software package, version 5.1.28. Institute of Phonetic Sciences, University of Amsterdam, The Netherlands," Retrieved 10 March 2010 from http://www.praat.org/ (Last viewed 9/29/2014).
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 1–790.
Brown, G. J., and Cooke, M. (**1994**). "Computational auditory scene analysis," Comput. Speech Lang. **8**, 297–336.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with an ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2009**). "Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers," J. Acoust. Soc. Am. **125**, 4006–4022.
Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L., and Kidd, G. (**2005**). "Across-ear interference from parametrically degraded synthetic speech signals in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **117**, 292–304.
Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.
Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.
Cooke, M., Green, P., Josifovski, L., and Vizinho, A. (**2001**). "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun. **34**, 267–285.
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," Philos. Trans. R. Soc. B **363**, 1011–1021.
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.
Dubbelboer, F., and Houtgast, T. (**2008**). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," J. Acoust. Soc. Am. **124**, 3937–3946.
Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.
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," Br. J. Audiol. **27**, 233–246.
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.
Hall, J. W., Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectro-temporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.
Henke, W. L. (**2005**). "MITSYN: A coherent family of high-level languages for time signal processing, software package (Belmont, MA)," www.mitsyn.com (Last viewed 9/29/2014).
Institute of Electrical and Electronics Engineers (IEEE) (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**, 225–246.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.
Keppel, G., and Wickens, T. D. (**2004**). *Design and Analysis: A Researcher's Handbook*, 4th ed. (Pearson Prentice Hall, Englewood Cliffs, NJ), pp. 1–611.
Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Auditory Perception of Sound Sources, Springer Handbook of Auditory Research*, edited by W. A. Yost and R. R. Fay (Springer, Berlin), Vol. 29, pp. 143–189.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
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.
Lindblom, B. E. F., and Sundberg, J. E. F. (**1971**). "Acoustical consequences of lip, tongue, jaw, and larynx movement," J. Acoust. Soc. Am. **50**, 1166–1179.
Lyzenga, J., and Carlyon, R. P. (**2000**). "Binaural effects in center-frequency modulation detection interference for vowel formants," J. Acoust. Soc. Am. **108**, 753–759.
Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cognit. Proc. **27**, 953–978.
Neff, D. L. (**1995**). "Signal properties that reduce masking by simultaneous, random-frequency maskers," J. Acoust. Soc. Am. **98**, 1909–1920.
Porter, R. J., and Whittaker, R. G. (**1980**). "Dichotic and monotic masking of CV's by CV second formants with different transition starting values," J. Acoust. Soc. Am. **67**, 1772–1780.
Remez, R. E., Dubowski, K. R., Davids, M. L., Thomas, E. F., Paddu, N. U., Grossman, Y. S., and Moskalenko, M. (**2011**). "Estimating speech spectra for copy synthesis by linear prediction and by hand," J. Acoust. Soc. Am. **130**, 2173–2178.
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.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2010**). "The perceptual organization of sine-wave speech under competitive conditions," J. Acoust. Soc. Am. **128**, 804–817.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2011**). "The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes," Proc. R. Soc. London, Ser. B **278**, 1595–1600.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2014**). "Formant-frequency variation and informational masking of speech by extraneous formants: Evidence against dynamic and speech-specific acoustical constraints," J. Exp. Psychol. Hum. Percept. Perform. **40**, 1507–1525.
Roberts, B., Summers, R. J., and Bailey, P. J. (**2015**). "Acoustic source characteristics, across-formant integration, and speech intelligibility under competitive conditions," J. Exp. Psychol. Hum. Percept. Perform. (published online).
Rosenberg, A. E. (**1971**). "Effect of glottal pulse shape on the quality of natural vowels," J. Acoust. Soc. Am. **49**, 583–590.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cognit. Sci. **12**, 182–186.
Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods*, 6th ed. (Iowa Press, Ames, IA), pp. 1–310.
Stone, M. A., Füllgrabe, C., Mackinnon, R. C., and Moore, B. C. J. (**2011**). "The importance for speech intelligibility of random fluctuations in 'steady' background noise," J. Acoust. Soc. Am. **130**, 2874–2881.
Stone, M. A., Füllgrabe, C., and Moore, B. C. J. (**2012**). "Notionally steady background noise acts primarily as a modulation masker of speech," J. Acoust. Soc. Am. **132**, 317–326.
Summers, R. J., Bailey, P. J., and Roberts, B. (**2010**). "Effects of differences in fundamental frequency on across-formant grouping in speech perception," J. Acoust. Soc. Am. **128**, 3667–3677.
Summers, R. J., Bailey, P. J., and Roberts, B. (**2012**). "Effects of the rate of formant-frequency variation on the grouping of formants in speech perception," J. Assoc. Res. Otolaryngol. **13**, 269–280.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.
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.
Weismer, G., and Berry, J. (**2003**). "Effects of speaking rate on second formant trajectories of selected vocalic nuclei," J. Acoust. Soc. Am. **113**, 3362–3378.