Informational masking of speech by time-varying competitors: Effects of frequency region and number of interfering formants

Brian Roberts, and Robert J. Summers

Citation: The Journal of the Acoustical Society of America **143**, 891 (2018);
View online: https://doi.org/10.1121/1.5023476
View Table of Contents: http://asa.scitation.org/toc/jas/143/2
Published by the Acoustical Society of America

Informational masking of speech by time-varying competitors: Effects of frequency region and number of interfering formants

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

(Received 31 October 2017; revised 17 January 2018; accepted 23 January 2018; published online 13 February 2018)

This study explored the extent to which informational masking of speech depends on the frequency region and number of extraneous formants in an interferer. Target formants—monotonized three-formant (F1+F2+F3) analogues of natural sentences—were presented monaurally, with target ear assigned randomly on each trial. Interferers were presented contralaterally. In experiment 1, single-formant interferers were created using the time-reversed F2 frequency contour and constant amplitude, root-mean-square (RMS)-matched to F2. Interferer center frequency was matched to that of F1, F2, or F3, while maintaining the extent of formant-frequency variation (depth) on a log scale. Adding an interferer lowered intelligibility; the effect of frequency region was small and broadly tuned around F2. In experiment 2, interferers comprised either one formant (F1, the most intense) or all three, created using the time-reversed frequency contours of the corresponding targets and RMS-matched constant amplitudes. Interferer formant-frequency variation was scaled to 0%, 50%, or 100% of the original depth. Increasing the depth of formant-frequency variation and number of formants in the interferer had independent and additive effects. These findings suggest that the impact on intelligibility depends primarily on the overall extent of frequency variation in each interfering formant (up to ~100% depth) and the number of extraneous formants.

https://doi.org/10.1121/1.5023476

[VB] Pages: 891–900

I. INTRODUCTION

An important requirement for successful communication in the auditory scenes often encountered in everyday life is the ability of the listener to attend to the speech of the talker despite the presence of interfering sounds, including other speech (Cherry, 1953; see Bregman, 1990; Darwin, 2008). The interference produced by these sounds can be considered to fall into three categories. The first is energetic masking, in which the auditory-nerve response to the key acoustic features of the target speech is swamped by the response to other concurrent sounds. The second is modulation masking, in which amplitude variation in the interferer lowers sensitivity to similar rates of variation in the target (e.g., Stone *et al.*, 2012; Stone and Moore, 2014). The third is informational masking (see, e.g., Kidd *et al.*, 2008), in which the interferer compromises the effectiveness with which the listener forms a coherent auditory object from the constituents of the target speech, is able to attend to it, or has access to the general cognitive resources required to process its critical features (see, e.g., Shinn-Cunningham, 2008). Speech is a sparse signal in a frequency × time representation, and so there are often circumstances in which the interference arises mainly from informational masking—e.g., when the target talker is accompanied by one competing talker with a similar level (Brungart, 2001; Brungart *et al.*, 2006; see Darwin, 2008).

The experiments reported here concern the informational component of speech-on-speech masking. A number of studies have focused on the linguistic aspects of this interference. For example, there is evidence that interfering speech has more impact when it is in the native language of the listener rather than in another language (e.g., Brouwer *et al.*, 2012), and that syntactic constraints influence the abilities of listeners to separate target speech from masking speech under conditions of spatial uncertainty (Kidd *et al.*, 2014). The focus here, however, is on using unintelligible interferers with precisely controlled properties to elucidate further the impact that informational masking has on processing the acoustic-phonetic features of the target speech. Both kinds of informational masking of speech occur under natural listening conditions.

A simple means of fully isolating the informational from the energetic component of masking is to present the target and masker to opposite ears. For example, some studies have investigated the ability to listen with independent ears by asking listeners to identify monaural target speech when presented alone or accompanied by a contralateral masker whose properties have been manipulated in various ways (e.g., Brungart *et al.*, 2005; Gallun *et al.*, 2007). This general approach has been extended to an arrangement in which a simplified three-formant analogue of a sentence-length utterance (F1+F2+F3) is accompanied in the contralateral ear by a single-formant interferer (Roberts and Summers, 2015; Summers *et al.*, 2016). In these studies, the properties of the single-formant interferer were derived from

[a]Electronic mail: b.roberts@aston.ac.uk, ORCID: 0000-0002-4232-9459.

those of the target F2, by time reversal or inversion of the F2 frequency contour, and so the interferer has mainly been conceptualized as a competitor for the second formant (termed F2C). Recent research using this stimulus configuration, or similar ones in which energetic masking of the target was largely (if not entirely) eliminated, have shown that the impact of the interferer on intelligibility is governed mainly by the time-varying properties of the F2C frequency contour, whereas the F2C amplitude contour has relatively little effect or none at all (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015). Only radical differences between formants in acoustic source characteristics, such as harmonic vs sine-wave analogues, have been shown to modulate the impact of an interferer on intelligibility to an extent comparable with that of formant-frequency variation in the interferer (Roberts *et al.*, 2015; Summers *et al.*, 2016).

Studies using the F2C paradigm, or variants of it, have shown that the impact of an interfering formant tends to rise as either the rate or extent (depth) of its formant-frequency variation increases (Summers *et al.*, 2012; Roberts *et al.*, 2014). To date, the effect of rate has been explored less extensively, but it has been shown that the impact of a two-formant interferer grows progressively as the time-base for its formant-frequency variation is increased from zero to at least twice the natural rate (Summers *et al.*, 2012). For the depth of formant-frequency variation, it has been established that the intelligibility of the target speech falls progressively as the depth of F2C increases from 0% (i.e., constant) up to 100% scaling (i.e., equal to that of the target F2), but that there is little further fall in intelligibility when F2C depth is increased to 200% scaling (Roberts and Summers, 2015). Note that the rise in masking that occurs as the depth of F2C is increased from 0% to 100% cannot be explained in terms of target-masker similarity between the corresponding formants. This is because halving the depth of variation in the target formants did not change the relationship between F2C depth and masker impact over the 0%–100% range, even though the point of maximum target-similarity occurred when F2C was set to 50% depth (Roberts *et al.*, 2014). Furthermore, similarity between F2 and F2C in their general pattern of movement was not important—F2Cs whose frequency contours were derived by inverting and rescaling the frequency contour of F2 had the same impact as depth-matched versions that instead followed a stylized triangular contour (Roberts *et al.*, 2014).

It remains unclear why the impact on intelligibility plateaus when the depth of F2C formant-frequency variation exceeds 100%. One possibility concerns the frequency region occupied by F2C, which has the same geometric mean frequency as that of the target F2, but begins to overlap with the ranges of the other formants once depth exceeds 100%. There is good reason to expect F2 to make a greater contribution to intelligibility than any of the other formants. The second formant is typically associated with the front cavity and carries critical information about front-back tongue position for vowels, and F2 transitions provide critical information about place of articulation for plosives and other consonants (see, e.g., Stevens, 1998). Furthermore, at least for sine-wave analogues of three-formant sentences (see Remez *et al.*, 1981), there is evidence of a hierarchy of contributions by different formants to intelligibility. Specifically, removal of F2 typically lowers intelligibility the most and removal of F3 the least; also, for single-formant stimuli, F2 is typically the most intelligible and F3 is the least (Han and Chen, 2017). Hence, one possible explanation for why the impact of F2C plateaus once its depth of formant-frequency variation exceeds 100% is because the presence of extraneous acoustic-phonetic information in frequency regions outside the typical F2 range caused little additional interference. Another possibility is that the observed plateau in impact as the depth of formant-frequency variation was increased is a peculiarity arising from the use of single-formant interferers. Therefore, the current study explored the extent to which informational masking is governed by the frequency region occupied by an interferer, the extent of formant-frequency variation in each interfering formant, and the number of interfering formants.

II. EXPERIMENT 1

In this experiment, the effect of interferer frequency region on target intelligibility was tested using a single-formant interferer whose geometric mean frequency was matched to that of F1, F2, or F3. Three versions of the interferer were created by taking the frequency contour of F2C and transposing it down into the F1 range or up into the F3 range, while maintaining its root-mean-square (RMS) level and the extent of its formant-frequency variation (100% depth for F2C) on a log scale. Note that this approach differs in three important ways from a comparison of the effects of single-formant interferers derived independently from F1, F2, and F3. First, in contrast with the usual effect of spectral tilt in speech stimuli, all versions of the interferer are matched for total energy; second, all versions share the same depth and pattern of movement about their mean frequencies but, third, using log transposition of the F2C contour to maintain 100% depth typically leads to an interferer with a smaller frequency range than F1 and a larger range than F3. The first two aspects were intended to facilitate comparison between the effects of the three versions of the interferer, but note that the third aspect—an inevitable consequence of the second—may complicate the comparison. This aspect is considered further below.

A. Method

1. Listeners

All listeners were students or members of staff at Aston University and 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 listeners 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 listeners completed the training successfully and took part in the main experiment, but five did not meet the additional criterion of a mean score

of ≥20% keywords correct in the main experiment, when collapsed across conditions, and so were replaced. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. Twenty listeners (five males) successfully completed the experiment (mean age = 23.2 yr, range = 18.3–44.2). 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 (mostly British) 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 Bamford–Kowal–Bench (BKB) sentences (Bench *et al.*, 1979) spoken by a British male talker of "Received Pronunciation" English. To enhance the intelligibility of the synthetic analogues, the 30 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.

Synthetic-formant analogues of each sentence were created using the corrected 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 with a fundamental frequency (F0) of 140 Hz was used in the synthesis of all stimuli used in the training and main experiment; note that no noise source was used and so all phonetic segments in these analogues were rendered fully as voiced, regardless of their original source characteristics. 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 at least 80 Hz away 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 interferers was created by time-reversing the frequency contour of F2 and transposing it such that its geometric mean frequency matched that of F1, F2, or F3. These variants were termed $F2C_1$, $F2C_2$, and $F2C_3$, respectively. The transposition maintained the extent of formant-frequency variation of $F2C_n$ on a log scale. The amplitude was constant and set in all cases to the RMS level of the target F2. All competitors were rendered as the outputs of a second-order resonator. The excitation source (Rosenberg pulses), F0 (140 Hz), 3-dB bandwidth (70 Hz), and output sign (–) were identical to those used to synthesize the target F2. When present, $F2C_n$ was always sent to the contralateral ear.

There were five conditions in the main experiment (see Table I). One condition (C1) was a control, for which only F1 and F3 were presented. Three conditions (C2–C4) were experimental, for which the stimuli comprised all three target formants accompanied by one version of $F2C_n$. The final condition (C5) was the reference case, for which only the monaural target formants were presented. The stimuli are illustrated in Fig. 1 using the narrowband spectrogram of a synthetic analogue of an example sentence in the top panel and the three versions of the corresponding interferers in the lower panels ($F2C_1$, $F2C_2$, and $F2C_3$, in descending order). For each listener, the 30 sentences were divided equally across conditions (i.e., 6 per condition), such that there were 18–20 keywords per condition. Allocation of sentences to conditions was counterbalanced by rotation across each set of five listeners tested. Hence, the total number needed to produce a balanced dataset was a multiple of five listeners.

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

TABLE I. Stimulus properties for the conditions used in experiment 1 (main session). The frequency contour of each variant of the single-formant interferer ($F2C_n$, where $n = 1$, 2, or 3) was derived by time-reversing the frequency contour of the target F2 and transposing it such that the geometric mean frequency matched that for the corresponding *n*th target formant.

Condition	Stimulus configuration (target ear; other ear)
C1	(F1+F3; —)
C2	(F1+F2+F3; $F2C_1$)
C3	(F1+F2+F3; $F2C_2$)
C4	(F1+F2+F3; $F2C_3$)
C5	(F1+F2+F3; —)

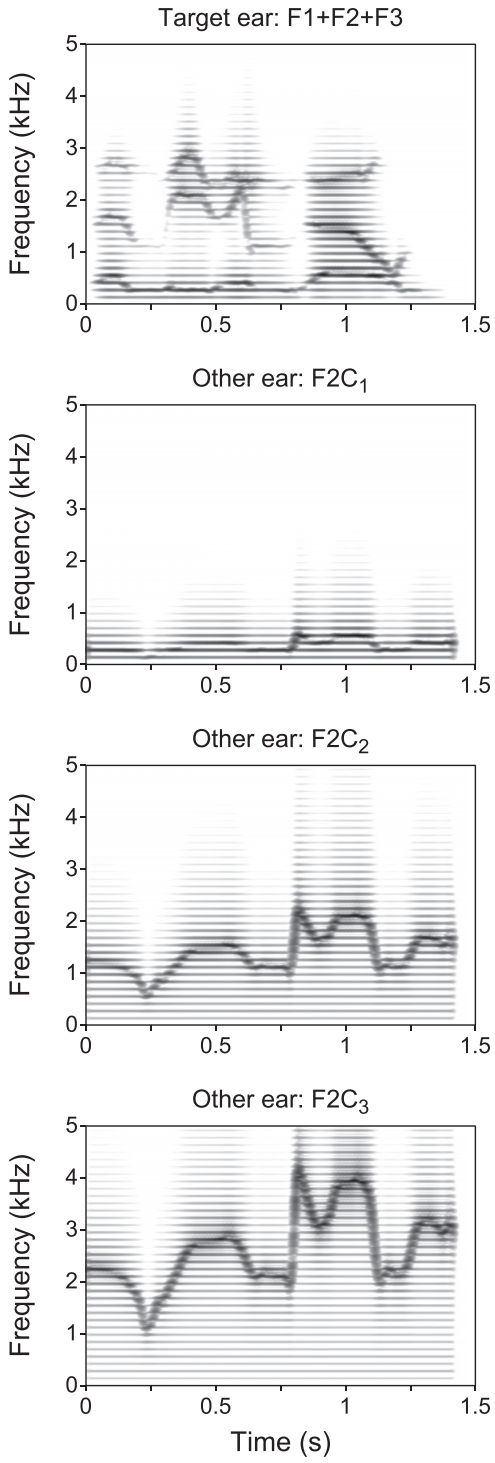

FIG. 1. Stimuli for experiment 1—narrowband spectrograms for a three-formant analogue of the sentence "They're kneeling down" (top) accompanied in the contralateral ear by one of three variants of a single-formant interferer ($F2C_n$, where $n = 1$, 2, or 3). These interferers were derived from the time-reversed frequency contour of the target F2 by transposing it to match the geometric mean frequency of the corresponding nth target formant (lower panels, in descending order). The amplitude of $F2C_n$ was constant and matched to the RMS power of the target F2. Note how the linear frequency range of $F2C_n$ increases with n; the log-frequency range remains constant.

session and typically took about 45 minutes 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 interferers and a new sentence was used for each trial. On each of the first ten trials, listeners heard diotic presentations of the synthetic version (S) and the original (clear, C) recording of a sentence in the order SCSCS; no response was required but listeners were asked to attend 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 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-like stimuli.

During the final ten 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 stimuli delivered only to the selected ear. Listeners continued on to the main session if they met either or both of two criteria: (1) $\geq$50% keywords correct across all 40 trials needing a transcription (30 diotic with repeat listening; 10 monaural, random selection of ear, no repeat listening); (2) $\geq$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; $F2C_n$ (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 Ltd, Singapore), programmable attenuators [Tucker-Davis Technologies (TDT) 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 ear receiving $F2C_n$ (mean $\approx$ 67 dB SPL) depending on the RMS power of the target F2. In the training session, the presentation level of the diotic materials (original recordings and first 40 sentences) was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation. The last ten sentences in the training session were presented monaurally at the reference level.

4. Data analysis

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; homonyms were accepted. As in our previous studies (e.g., Roberts *et al.*, 2010; Roberts and Summers, 2015), we classified responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly. Except where stated, the values and statistics reported here are based on these tight keyword scores. All statistical analyses were computed using SPSS (SPSS statistics version 21, IBM Corp., Armonk, NY). The measures of effect size reported here are eta squared (η^2) and partial eta squared (η^2_p). All pairwise comparisons (two tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967; Keppel and Wickens, 2004).

Following Roberts *et al.* (2014), we used phonemic scoring as an additional measure to the tight scoring of keywords. Typed responses were converted automatically into phonemic representations using eSpeak (Duddington, 2014), which generates phonemic representations of the input text using a pronunciation dictionary and a set of generic pronunciation rules for English orthography. The mean percentage of phonemes correctly identified across all words in the sentences was computed using an algorithm that finds an optimal alignment between the sequence of phonemes for the original sentence and its transcription through insertion and removal of transcribed segments (see Needleman and Wunsch, 1970). The mean percentage of phonemes correctly identified—the phonemic score—is defined as 100 × (number of correctly aligned phonemes)/(number of phonemes in the original sentence).

B. Results and discussion

Figure 2 shows the mean percentage scores (and inter-subject standard errors) across conditions for keywords correctly identified. The black, gray, and white bars indicate the results for the control, target-plus-interferer, and target-only conditions, respectively; the corresponding means for the phonemic scores are shown in the caption. A one-way within-subjects analysis of variance (ANOVA) of the keyword scores across the five conditions showed a highly significant effect of condition on intelligibility [$F(4,76) = 18.418$, $p < 0.001$, $\eta^2 = 0.492$].[1] Scores for the control condition show that intelligibility was low in the absence of F2; the mean score for C1 was significantly different from those for all other conditions ($p < 0.001$ in all cases). Performance was best when all three target formants were presented alone (C5); note that the relatively low intelligibility compared with natural speech is to be expected given the simple source properties and three-formant parallel vocal-tract model used to synthesize the sentences. When the monaural target was accompanied by any of the contralateral interferers (C2–C4), intelligibility was lowered significantly relative to C5 (overall mean difference = 16.5% pts; $p = 0.015$ – $p < 0.001$). Clearly, all three versions of $F2C_n$ were effective interferers.

The differences between target formants in the overall extent to which they carry useful acoustic-phonetic information led to the expectation that the contralateral interferer would have the greatest impact when it fell in the same frequency range as F2 and least when it fell in the same range as F3. Figure 2 shows a pattern consistent with this prediction, but a one-way ANOVA of the keyword scores restricted to the three experimental conditions (C2–C4) was not significant [$F(2,38) = 2.072$, $p = 0.140$]. This outcome is perhaps unsurprising given the relatively small differences between conditions ($F2C_1$ vs $F2C_2 = 5.1\%$ pts; $F2C_1$ vs $F2C_3 = 3.4\%$ pts; $F2C_2$ vs $F2C_3 = 8.5\%$ pts) and the inevitable compromise to the sensitivity of the analysis arising from the need to rotate the allocation of sentences across conditions in this design, which increased the extent of uncontrolled variance in our dataset. The phonemic scores follow a similar pattern to the keyword scores, but for these scores the effect of frequency region was significant [$F(2,38) = 3.427$, $p = 0.043$, $\eta^2 = 0.153$]. Pairwise comparisons revealed a significant difference between C3 and C4 [$F2C_2$ vs $F2C_3 = 9.0\%$ pts; $t(19) = 2.252$, $p = 0.036$]; the difference between C2 and C3 did not quite reach significance [$F2C_1$ vs $F2C_2 = 5.4\%$ pts; $t(19) = 1.918$, $p = 0.070$]. In all other respects, the outcomes of the supplementary analysis using the phonemic scores were consistent with the main analysis.

It is also worth noting that the log frequency range for F3 was considerably smaller than that for F2, and so transposing F2C upward while preserving its log range resulted in considerable overlap between the frequency ranges of $F2C_3$ and F2 (on average, ~40% of the F2 range fell within the

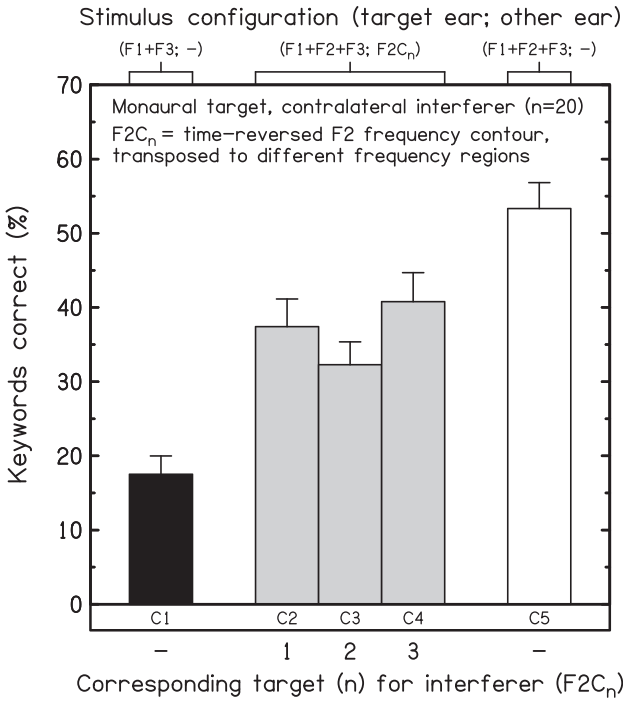

FIG. 2. Results for experiment 1—effect of frequency region on the impact of an interferer ($F2C_n$, where $n = 1$, 2, or 3) on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and inter-subject standard errors ($n = 20$) are shown for the control condition (black bar), the target-plus-interferer conditions (gray bars), and the target-only condition (white bar). The top axis indicates which formants were presented to each ear; the bottom axis indicates to which target formant the geometric mean frequency of $F2C_n$ was matched (when present). For ease of reference, condition numbers are included immediately above the bottom axis. The corresponding means for phonemic scoring across conditions were 32.1% (C1), 51.0% (C2), 45.6% (C3), 54.6% (C4), and 65.0% (C5).

range for $F2C_3$). The extent of this overlap probably reduced the difference in impact between $F2C_2$ and $F2C_3$. In addition, the frequency response of the headphones used simulated the broad resonance of the ear canal and so interferers transposed upward into the F3 range were boosted by ~2.5 dB on average, which is likely to have reduced still further the difference in impact between $F2C_2$ and $F2C_3$. On balance, it seems reasonable to conclude that the impact of a single-formant interferer is influenced by the frequency region into which it is transposed, but that this effect is rather broadly tuned around the F2 region. Why this might be the case is considered further in Sec. IV.

III. EXPERIMENT 2

Although one previous study (Summers *et al.*, 2012) used a variant of the F2C paradigm involving two-formant interferers (specifically F2C+F3C), their impact on intelligibility was not compared with that for single-formant interferers. In this experiment, the effects of single- and three-formant interferers on target intelligibility were compared directly. These interferers (F1C and F123C) were derived from the target F1—the most intense formant—and from all three target formants, respectively. The spectral tilt characteristic of speech ensured that the inclusion of the higher formants had little effect on the total energy of the interferer, such that any significant increase in impact observed when they were included could not be explained in terms of changes in overall power. The depth of formant-frequency variation in these interferers was also manipulated, without exceeding the natural range, to provide a benchmark against which the impact of increasing the number of interfering formants could be assessed. Note that adding formants with different mean frequencies is a way of extending the frequency range over which there is formant-frequency variation in the interferer without increasing the depth of that variation.

A. Method

Except where described, the same method was used as for experiment 1. There were nine conditions in experiment 2; hence, the total number required to produce a balanced dataset was a multiple of nine listeners. Twenty-seven listeners (four males) passed the training and successfully completed the experiment (mean age = 21.8 yr, range = 18.3–31.9); one listener who passed the training failed to meet the additional criterion of at least 20% keywords correct overall in the main experiment and so was replaced. The training session was identical to that used in experiment 1. The stimuli for the main experiment were derived from recordings of 54 sentences spoken by the same talker as for the BKB sentences used in experiment 1. The text for these sentences was provided by Patel and Morse (2010) and consisted of variants created by rearranging words from BKB sentence lists while maintaining semantic simplicity. Sentences were allocated to conditions in the same way as for experiment 1 (18–19 keywords per condition).

All stimuli were generated using the same excitation source, F0 (140 Hz), resonator bandwidths, and output signs as for experiment 1. A set of six interferers was created for each sentence in the main experiment. Three versions of the single-formant interferer (F1C) were created using the time-reversed frequency contour of F1 and applying different scale factors to adjust the extent of formant-frequency variation to 0% (constant), 50%, or 100% of the original depth (Roberts *et al.*, 2014; Roberts and Summers, 2015). Rescaling of the formant-frequency contour was performed on a log scale about the geometric mean frequency. 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 ($0 \le x \le 1$) is a proportional scale factor determining the maximum frequency range relative to the original (the depth of variation), $f(t)$ is the formant frequency at time t, and g is the geometric mean of the whole formant-frequency contour. In all cases, F1C had a constant amplitude and was matched to the RMS level of its counterpart target F1. Three versions of the three-formant interferer (F123C) were created using the time-reversed frequency contours of all three target formants. Each formant in F123C had its depth of formant-frequency variation scaled to 0%, 50%, or 100% of that for the corresponding target formant. The amplitude for each formant in the interferer was constant and matched to the RMS level of the corresponding target formant. After the three interfering formants were synthesized and added together, the overall RMS level of F123C was adjusted to match that of the complete target (F1 + F2 + F3) to take into account differences in harmonic cancellation across formants between target and interferer. Given that F1 contained most of the energy of the target stimulus, there was little difference in level between F1C and the corresponding F123C (always <1 dB).

There were nine conditions in the main experiment (see Table II). Two conditions (C1 and C2) were controls for which the target F1 was absent (i.e., F2 + F3); C2 also contained F1C (at 100% depth) in the contralateral ear. Six conditions (C3–C8) were experimental, for which the target formants were accompanied in the contralateral ear by either F1C (C3–C5) or F123C (C6–C8). Across these conditions, the depth of formant-frequency variation in the interferer was scaled to 0% (C3 and C6), 50% (C4 and C7), or 100% (C5 and C8). The final condition (C9) was the reference case, for which only the monaural target formants were presented (i.e., no interfering formants). The stimuli are illustrated in Fig. 3 using the narrowband spectrogram of a synthetic analogue of an example sentence (top panel) accompanied by the corresponding single-formant interferer (F1C, middle panel) and three-formant interferer (F123C, bottom panel) when presented at 100% depth.

B. Results and discussion

Figure 4 shows the mean percentage keyword scores (and intersubject standard errors) for each condition. The black, gray, and white bars indicate the results for the control, target-plus-interferer, and target-only conditions, respectively; within the target-plus-interferer conditions, light and dark gray bars

TABLE II. Stimulus properties for the conditions used in experiment 2 (main session). The frequency contours of the single- (F1C) and three-formant (F123C) interferers were derived from time-reversed versions of the corresponding target formants. The scale factor for the interferer refers to the depth of variation in formant frequency for each interfering formant, relative to that for the corresponding unscaled target formant(s). A scale factor of 0% indicates a constant frequency contour for each interfering formant, corresponding to the geometric mean frequency of the target counterpart.

Condition	Stimulus configuration (target ear; other ear)	Scale factor (depth) of interfering formants relative to corresponding target formants (%)
C1	(F2+F3; —)	—
C2	(F2+F3; F1C)	100
C3	(F1+F2+F3; F1C)	0
C4	(F1+F2+F3; F1C)	50
C5	(F1+F2+F3; F1C)	100
C6	(F1+F2+F3; F123C)	0
C7	(F1+F2+F3; F123C)	50
C8	(F1+F2+F3; F123C)	100
C9	(F1+F2+F3; —)	—

indicate the results for single- and three-formant interferers, respectively. A one-way within-subjects ANOVA over all nine conditions demonstrated a highly significant effect of condition on intelligibility [$F(8,208) = 59.394$, $p < 0.001$, $\eta^2 = 0.696$]. Intelligibility was low in the absence of the target F1 (C1) and was near floor when F2 + F3 was accompanied in the other ear by F1C (C2). Keywords scores for C1 and C2 were significantly different from each other ($p = 0.003$) and from those for all other conditions ($p \leq 0.001$ in all cases). Performance was best when all three target formants were presented alone (C9). Intelligibility was significantly lowered, often substantially, when the target formants were accompanied by any of the contralateral interferers (C9 vs C3–C8, overall mean difference = 24.3% pts, $p < 0.003$ in all cases). Note that the impact of the 100%-depth F1C (31.4% pts) was considerably larger than that of its second-formant counterpart in experiment 1 ($F2C_2$, 21.1% pts) or of F2C in any of our previous studies; this is probably because F1C was much more intense than F2C.[2]

The effect of the experimental manipulations of the interfering formants was explored using a two-way within-subjects ANOVA restricted to the target-plus-interferer conditions (C3–C8). The two factors were the number of interfering formants (two levels: one or three) and the depth of formant-frequency variation (three levels: 0%, 50%, or 100%). This analysis revealed significant main effects of the number of interfering formants [$F(1,26) = 11.955$, $p = 0.002$, $\eta^2_p = 0.315$] and the depth of formant-frequency variation [$F(2,52) = 38.713$, $p < 0.001$, $\eta^2_p = 0.598$]. There was no interaction between these factors [$F(2,52) = 0.123$, $p = 0.884$], indicating that their effects were independent and additive. Increasing the depth of formant-frequency variation in the interferer had the greater effect on keyword scores (average fall for 0% vs 100% depth = 22.5% pts) but there was also an appreciable effect, about one-third the size, of increasing the number of formants (average fall for F1C vs F123C = 7.9% pts). This was the case despite the fact that the additional formants hardly changed the presentation level of the interferer. Overall, the impact of the interferer was greatest for F123C at 100% depth (39.3% pts) and least for F1C at 0% depth (9.5% pts). The outcomes of the supplementary analysis using the phonemic scores were fully consistent with the main analysis.

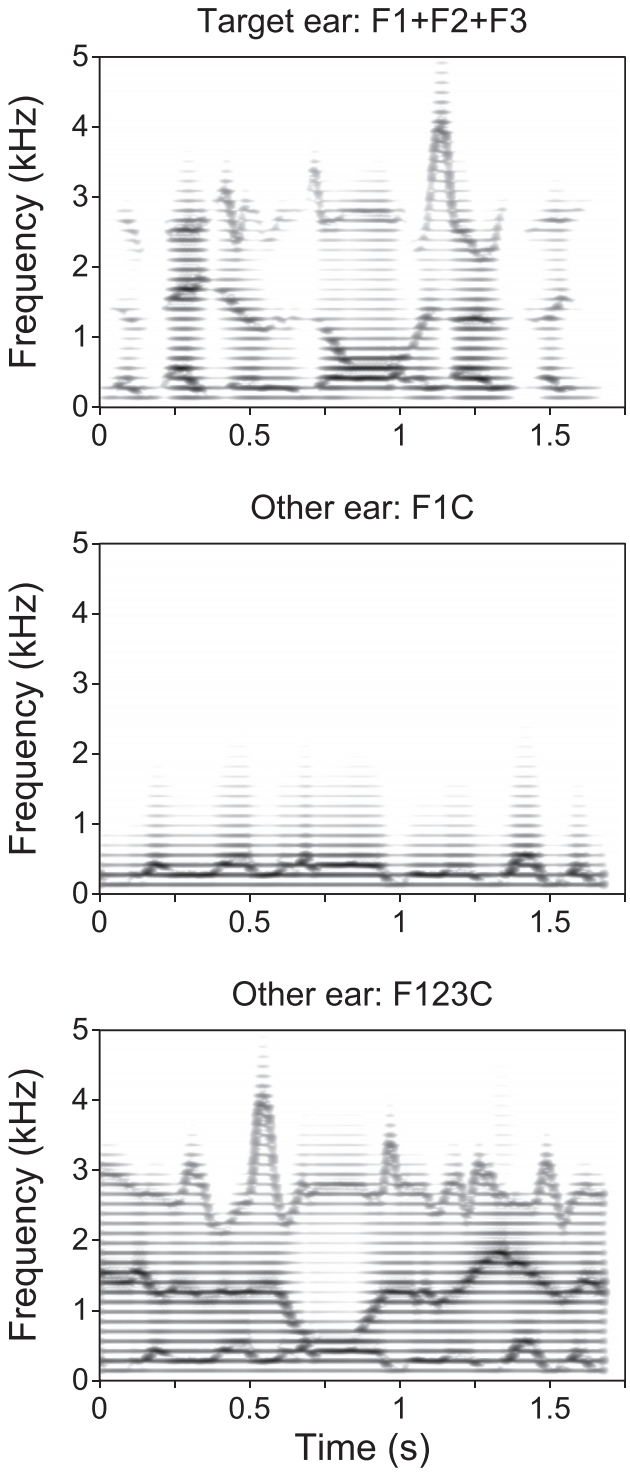

FIG. 3. Stimuli for experiment 2—narrowband spectrograms for a three-formant analogue of the sentence "The bedroom door was open" (top) accompanied in the contralateral ear by either a single-formant interferer (F1C; middle) or a three-formant interferer (F123C; bottom), both shown here when scaled to 100% depth. The frequency contours of these interferers were time-reversed versions of those for the corresponding target formants. The amplitude of each interfering formant was constant and matched to the RMS power of its target counterpart.

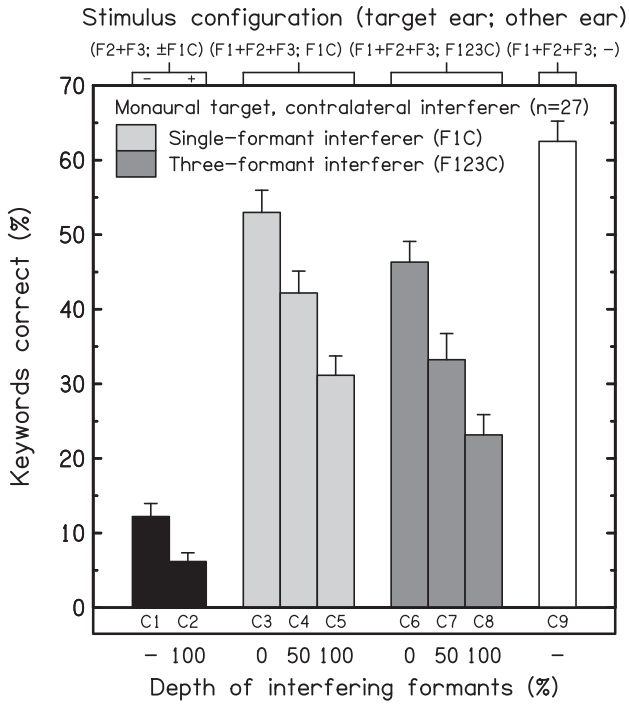

FIG. 4. Results for experiment 2—effect of the number of formants in an interferer (1 or 3; F1C or F123C) and their depth of formant-frequency variation (0%, 50%, or 100%) on the impact of that interferer on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and intersubject standard errors ($n = 27$) are shown for the control conditions (black bars), the target-plus-interferer conditions (gray bars), and the target-only condition (white bar). Light gray and dark gray bars indicate performance in the single- and three-formant interferer conditions, respectively. The top axis indicates which formants were presented to each ear; the bottom axis indicates the depth of formant-frequency variation in the interferer (when present). For ease of reference, condition numbers are included immediately above the bottom axis. The corresponding means for phonemic scoring across conditions were 27.2% (C1), 20.0% (C2), 66.3% (C3), 57.8% (C4), 47.9% (C5), 58.3% (C6), 46.7% (C7), 36.7% (C8), and 75.2% (C9).

Increasing the number of interfering formants from one (F1C) to three (F123C) in the contralateral masker extended the frequency range over which formant-frequency variation took place and increased the impact of the masker on intelligibility by about one-third. This suggests that extending the frequency range of the time-varying masker beyond that of the target F1 caused greater interference, albeit more gradually than the effect of increasing masker depth within the F2 range. This outcome differs from our previous finding that increasing the depth of formant-frequency variation beyond 100% for a single-formant interferer (F2C) had little extra effect (Roberts and Summers, 2015). One possible reason for this difference is that increasing the number of interfering formants increased the total amount of formant-frequency variation in the masker without increasing (for a given formant) the depth of that variation—and hence also the velocity of formant transitions—beyond the range associated with natural articulation (see, e.g., Tjaden and Weismer, 1998; Weismer and Berry, 2003).

Another possible reason for the greater impact of the three-formant interferer is the greater overall complexity of its time-varying properties. While the movements of the three formants were not independent of one another, derived as they were from the trajectories of the three target formants, they nonetheless moved in complex ways in relation to one another. It may also be relevant that, in comparison with the single-formant interferer, the three-formant interferer had spectro-temporal and bandwidth characteristics more like those of typical speech sounds. To our knowledge, the relationship between the number and complexity of formant movements in a speech-like masker and the interference it causes has received little or no attention, but note that there is evidence that the extent of informational masking for a single-tone target embedded in a multi-tone masker typically increases as the number of tones in the masker increases from two up to at least ten (Neff and Green, 1987; Oh and Lutfi, 1998).

IV. SUMMARY AND CONCLUDING DISCUSSION

Previous studies in which the F2C paradigm (or variants thereof) has been used to examine the acoustic factors that may contribute to the acoustic-phonetic aspects of speech-on-speech informational masking have manipulated two classes of feature. One is the time-varying properties of the interfering formants, including the rate and depth of formant-frequency variation (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015; Summers *et al.*, 2012), the presence of formant amplitude variation (Roberts *et al.*, 2010; Summers *et al.*, 2012), and the presence of fundamental-frequency (F0) variation (Summers *et al.*, 2017). The other is the source properties of the target and interfering formants, including differences between formants in F0 (Summers *et al.*, 2010; Summers *et al.*, 2017) and stimulus configurations in which some formants were rendered as buzz-excited resonances and others as sine-wave analogues (Roberts *et al.*, 2015; Summers *et al.*, 2016). The experiments reported here have extended these investigations to include the frequency region within which interfering formant-frequency variation occurs and the number of interfering formants present. Under dichotic presentation and the conditions of uncertainty with respect to ear of presentation used here, the impact on intelligibility depended primarily on the overall extent of formant-frequency variation in the interferer (at least up to ~100% depth), but it was also influenced by the number of time-varying formants contributing to the interferer. Increasing the number of interfering formants increases the total amount of formant-frequency variation without exceeding the natural range of variation for any one formant; it also increases the overall complexity of the interferer and the properties that it has in common with typical speech. For a given extent of formant-frequency variation (on a log scale), the particular frequency region occupied by an extraneous formant had relatively little effect on the interference caused.

The F2C paradigm was originally conceptualized as involving competition between the target F2 and the extraneous formant to form a coherent group with F1+F3, in which a fall in intelligibility arises from the displacement (or dilution) of F2 from the perceptual organization of the target speech (Remez *et al.*, 1994; Roberts *et al.*, 2010). Roberts *et al.* (2014) noted that this interpretation attributes the informational masking to a specific corrupting influence of the interferer, in which there is a failure to exclude acoustic

variation in the extraneous formant from the perceptual evaluation of the acoustic-phonetic features of the target sentence, but that the impact of the interferer may instead represent a non-specific disrupting influence, in which the interferer acts as a cognitive load (see, e.g., Mattys *et al.*, 2012) that limits the general processing capacity available for the target sentence. Our finding that transposing an interfering formant of a given depth and pattern of formant-frequency variation across a wide range of frequencies had such a modest effect on the overall extent of informational masking is not conclusive regarding this distinction, but the outcome is in accord with the notion that the non-specific disruptive effects of the interferer make a substantial contribution to its overall impact.

When considering the results of the current study in the broader context of studies using the F2C paradigm, and its variants, there is an interesting parallel between the informational masking observed in these studies and the irrelevant sound effect (ISE). The ISE is an example of cross-modal interference, in which the presence of an acoustic distractor that participants are instructed to ignore nonetheless impairs serial recall of visually presented digits or words. The effect tends not to habituate over repeated testing (Hellbrück *et al.*, 1996). Most notably, the ISE requires frequency change in the distractor (Jones and Macken, 1993); it cannot be induced by amplitude change alone (Tremblay and Jones, 1999). Similarly, studies using the F2C paradigm have shown that the impact on intelligibility of the time-varying properties of an extraneous formant is governed mainly by formant-frequency variation, whereas formant-amplitude variation has relatively little effect, or none at all, depending on the particular stimulus configuration used (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015; Summers *et al.*, 2012). It is also merits note that the ISE is typically greatest when the acoustic distractor is speech or has speech-like properties (e.g., Viswanathan *et al.*, 2014), but that it can also be substantial in the presence of instrumental music, which implies that complexity of spectro-temporal change is an important contributor to the ISE (see Ellermeier and Zimmer, 2014, for a review). An important challenge for future research on the informational masking of speech will be to assess the relative contributions of non-specific disruptive effects (general capacity limitations) and specific corrupting effects (intrusions) to the overall interference caused.

ACKNOWLEDGMENTS

This research was supported by Research Grant No. ES/N014383/1 from the Economic and Social Research Council (UK), awarded to B.R. To access the research data underlying this publication, see http://doi.org/10.17036/researchdata.aston.ac.uk.00000309. We are grateful to Peter Bailey for his advice concerning this research and for his comments on drafts of this paper. We also thank Rob Morse and Meghna Patel for providing the BKB-like sentences, Quentin Summerfield for enunciating them, and Gerald Kidd for his suggestions on the informational masking literature. Poster presentations on this research were given at the 173rd Meeting of the Acoustical Society of America (Boston, MA, June 2017), and the annual Basic Auditory Science Meeting (Nottingham, UK, September 2017).

[1]As a precaution, given the low scores obtained in the control condition(s), all ANOVAs reported in this study were repeated using arcsine-transformed data [$Y' = 2\arcsin(\sqrt{Y})$, where Y is the proportion correct score; see Studebaker, 1985]. The results confirmed the outcome of the original analyses; applying the transform did not change any of the comparisons reported here from significant to non-significant or vice versa.

[2]Consistent with this interpretation, the impact of F1C was also considerably greater than that of F2C in experiment 1 after its transposition into the F1 region without changing the RMS power ($F2C_1$, 15.9% pts). Note, however, that comparing the impact of F1C with $F2C_1$ is not straightforward because the frequency range of F1 is usually greater than that of F2 on a log scale and because they have different patterns of movement.

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.

Boersma, P., and Weenink, D. (**2010**). "PRAAT, a system for doing phonetics by computer (version 5.1.28) [computer program]," http://www.praat.org/ (Last viewed 10 March 2010).

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA), pp. 1–790.

Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (**2012**). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," J. Acoust. Soc. Am. **131**, 1449–1464.

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

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., 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.

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.

Duddington, J. (**2014**). "eSpeak 1.48," available at http://espeak.sourceforge.net/ (Last viewed 15 September 2016).

Ellermeier, W., and Zimmer, K. (**2014**). "The psychoacoustics of the irrelevant sound effect," Acoust. Sci. Technol. **35**, 10–16.

Gallun, F. J., Mason, C. R., and Kidd, G. (**2007**). "The ability to listen with independent ears," J. Acoust. Soc. Am. **122**, 2814–2825.

Han, Y., and Chen, F. (**2017**). "Relative contributions of formants to the intelligibility of sine-wave sentences in Mandarin Chinese," J. Acoust. Soc. Am. **141**, EL495–EL499.

Hellbrück, J., Kuwano, S., and Namba, S. (**1996**). "Irrelevant background speech and human performance: Is there long-term habituation?," J. Acoust. Soc. Jpn. (E) **17**, 239–247.

Henke, W. L. (**2005**). "MITSYN: A coherent family of high-level languages for time signal processing, [software package]" (Belmont, MA).

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

Jones, D. M., and Macken, W. J. (**1993**). "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory," J. Exp. Psychol. Learn. Mem. Cognit. **19**, 369–381.

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., and Best, V. (**2014**). "The role of syntax in maintaining the integrity of streams of speech," J. Acoust. Soc. Am. **135**, 766–777.

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*, Vol. 29, edited by W. A. Yost and R. R. Fay (Springer, Berlin), pp. 143–189.
Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.
Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cognit. Process. **27**, 953–978.
Needleman, S. B., and Wunsch, C. D. (**1970**). "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Biol. **48**, 443–453.
Neff, D. L., and Green, D. M. (**1987**). "Masking produced by spectral uncertainty with multicomponent maskers," Percept. Psychophys. **41**, 409–415.
Oh, E., and Lutfi, R. A. (**1998**). "Nonmonotonicity of informational masking," J. Acoust. Soc. Am. **104**, 3489–3499.
Patel, M., and Morse, R. P. (**2010**). (personal communication).
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.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–950.
Roberts, B., and Summers, R. J. (**2015**). "Informational masking of monaural target speech by a single contralateral formant," J. Acoust. Soc. Am. **137**, 2726–2736.
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. (**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. **41**, 680–691.
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 University Press, Ames, IA), pp. 1–310.
Stevens, K. N. (**1998**). *Acoustic Phonetics* (MIT Press, Cambridge, MA), pp. 1–607.
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.
Stone, M. A., and Moore, B. C. J. (**2014**). "On the near non-existence of 'pure' energetic masking release for speech," J. Acoust. Soc. Am. **135**, 1967–1977.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
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.
Summers, R. J., Bailey, P. J., and Roberts, B. (**2016**). "Across-formant integration and speech intelligibility: Effects of acoustic source properties in the presence and absence of a contralateral interferer," J. Acoust. Soc. Am. **140**, 1227–1238.
Summers, R. J., Bailey, P. J., and Roberts, B. (**2017**). "Informational masking and the effects of differences in fundamental frequency and fundamental-frequency contour on phonetic integration in a formant ensemble," Hear. Res. **344**, 295–303.
Tjaden, K., and Weismer, G. (**1998**). "Speaking-rate-induced variability in F2 trajectories," J. Speech Lang. Hear. Res. **41**, 976–989.
Tremblay, S., and Jones, D. M. (**1999**). "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound," J. Exp. Psychol. Hum. Percept. Perform. **25**, 1005–1015.
Viswanathan, N., Dorsi, J., and George, S. (**2014**). "The role of speech-specific properties of the background in the irrelevant sound effect," Q. J. Exp. Psychol. **67**, 581–589.
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.