Informational masking of speech depends on masker spectro-temporal variation but not on its coherence

Brian Roberts, and Robert J. Summers

Citation: The Journal of the Acoustical Society of America **148**, 2416 (2020); doi: 10.1121/10.0002359

View online: https://doi.org/10.1121/10.0002359

View Table of Contents: https://asa.scitation.org/toc/jas/148/4

Published by the Acoustical Society of America

JASA ARTICLE

Informational masking of speech depends on masker spectro-temporal variation but not on its coherence[a]

Brian Roberts[b] and Robert J. Summers[c]
School of Psychology, Aston University, Birmingham B4 7ET, United Kingdom

ABSTRACT:
The impact of an extraneous formant on intelligibility is affected by the extent (depth) of variation in its formant-frequency contour. Two experiments explored whether this impact also depends on masker spectro-temporal coherence, using a method ensuring that interference occurred only through informational masking. Targets were monaural three-formant analogues (F1+F2+F3) of natural sentences presented alone or accompanied by a contralateral competitor for F2 (F2C) that listeners must reject to optimize recognition. The standard F2C was created using the inverted F2 frequency contour and constant amplitude. Variants were derived by dividing F2C into abutting segments (100–200 ms, 10-ms rise/fall). Segments were presented either in the correct order (coherent) or in random order (incoherent), introducing abrupt discontinuities into the F2C frequency contour. F2C depth was also manipulated (0%, 50%, or 100%) prior to segmentation, and the frequency contour of each segment either remained time-varying or was set to constant at the geometric mean frequency of that segment. The extent to which F2C lowered keyword scores depended on segment type (frequency-varying vs constant) and depth, but not segment order. This outcome indicates that the impact on intelligibility depends critically on the overall amount of frequency variation in the competitor, but not its spectro-temporal coherence.

https://doi.org/10.1121/10.0002359

(Received 8 June 2020; revised 6 October 2020; accepted 7 October 2020; published online 28 October 2020)

[Editor: Karen S. Helfer] Pages: 2416–2428

I. INTRODUCTION

Spoken communication often takes place in complex auditory scenes in which listeners must segregate and attend to target speech when it is accompanied by other sounds, including interfering speech [see, e.g., Bregman (1990), Darwin (2008), and Mattys *et al.* (2012)]. The masking produced by other sounds can arise either through loss of encoding fidelity for the target in the auditory-nerve response (energetic masking, EM) or—even when there is a good representation of the target's critical features in the peripheral response—from a variety of causes in the central auditory system (informational masking, IM). Specifically, IM can arise from failures of object formation or selection, or from general capacity limitations on information processing [e.g., Shinn-Cunningham (2008)]. Given that the speech signal is sparse on a frequency × time representation, speech-on-speech interference is often dominated by IM, particularly when there is only one interfering voice and its level is similar to or lower than that of the target voice [e.g., Brungart *et al.* (2006)]. These are circumstances often experienced by listeners.

In a recent study of speech-on-speech IM, using simplified analogues of speech derived from sentence-length utterances, Summers and Roberts (2020) compared the impact on target intelligibility of acoustically matched intelligible and unintelligible interferers presented in the contralateral ear. It was estimated that about two thirds of the impact of these interferers on keyword scores arose from acoustic-phonetic interference (i.e., those aspects of IM that hinder the extraction or integration of information about speech articulation carried by the time-varying formant-frequency contours) whereas the rest arose from linguistic interference (i.e., those aspects of IM that occur after lexical objects have been formed, such as intrusion of words from an interfering sentence). Much research has focused on the linguistic components of this interference, but we still know relatively little about which acoustic properties of interfering speech govern the IM it causes. This is an important research question given the evidence that acoustic-phonetic interference makes a major contribution to IM. The study reported here focuses on the contribution of two acoustic properties of interfering formants—the amount and coherence of formant-frequency change over time. There is a growing body of evidence showing that formant-frequency variation plays an important role in speech-on-speech IM (Roberts *et al.*, 2010, 2014; Roberts and Summers 2015, 2018; Summers *et al.*, 2012), but precisely which aspects of this variation are critical remains unclear. In addition, to our knowledge, the relationship between the continuity of formant

[a] Poster presentations on the first experiment were given at the 171st Meeting of the Acoustical Society of America, Salt Lake City, UT, May 2016 and at the Basic Auditory Science Meeting of the British Society of Audiology, Cambridge, UK, September 2016.
[b] Electronic mail: b.roberts@aston.ac.uk, ORCID: 0000-0002-4232-9459.
[c] ORCID: 0000-0003-4857-7354.

movements in a speech-like masker and the IM it causes has so far received little or no attention. More generally, we are not aware of any IM studies, using either speech or non-speech stimuli, that have made direct comparisons between otherwise-matched coherent and incoherent maskers.

The second-formant competitor (F2C) paradigm (Remez *et al.*, 1994; Roberts *et al.*, 2010) offers a convenient method for exploring speech-on-speech IM with a high degree of stimulus control. It involves accompanying a three-formant analogue of a target sentence with an extraneous formant originally considered to act as a competitor by providing an alternative candidate for the target F2. F2C is always presented in the opposite ear to F2 and so causes relatively little EM for dichotic targets (i.e., F1 and F3 in the same ear as F2C) and none at all for monaural targets. Hence, the impact of the extraneous formant arises mainly or exclusively from IM. F2C, which is typically derived from the properties of F2 (e.g., by time reversing or inverting the F2 frequency contour), must be rejected by the listener to optimize recognition of the target speech. Research using the F2C paradigm or variants of it has shown that the impact of extraneous formants on intelligibility is governed mainly by the time-varying properties of their formant-frequency contours, whereas their amplitude contours have relatively little or no effect (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015, 2018; Summers *et al.*, 2012). To the extent that a constant-amplitude extraneous formant can have a greater impact than one with a time-varying amplitude contour (Roberts and Summers, 2015), this is probably a secondary effect arising from greater illumination of the underlying formant-frequency contour owing to the absence of low-amplitude intervals. To date, the only other factors shown to modulate the impact of an interferer on intelligibility to an extent comparable with that of formant-frequency change are radical differences between formants in acoustic source properties, such as harmonic vs sine-wave analogues (Roberts *et al.*, 2015; Summers *et al.*, 2016) and, to a lesser extent, differences in fundamental frequency (ΔF0) (Summers *et al.*, 2010, 2017). Nonetheless, IM can remain considerable even when there is a clearly discernible ΔF0 between the target and interfering formants (4 semitones) (Summers and Roberts, 2020).

Most studies using the F2C paradigm have focused on manipulating the depth of formant-frequency change, which is directly related to the *range* of variation. Specifically, adjusting the depth of frequency variation around the geometric mean of the F2C frequency contour from 0% (constant at the geometric mean) to 100% (full scale) progressively increases its impact on intelligibility (Roberts *et al.*, 2014; Roberts and Summers, 2015, 2018). Thereafter, further rises in IM are less pronounced and plateau around 150% depth. Summers *et al.* (2012) instead explored the effect of manipulating the *rate* of formant-frequency variation while preserving 100% depth. They found that IM rose as formant-frequency variation in the interferer was increased from one-quarter to twice natural rate, the highest rate reported. The parallel outcomes suggest that, at least up to some limit, it may be the *velocity* of the formant transitions in the interferer, rather than range or rate *per se*, that is the primary influence on speech-on-speech IM.

The notion of the *perceptual coherence* of an acoustic stimulus merits some clarification, particularly in the context of speech and speech-like stimuli. Some researchers use the term to refer to phonetic coherence, an abstract property of the acoustical speech signal that is assumed to derive from the kinematic and dynamic coherence of the articulatory gestures involved in speech production [e.g., Remez *et al.* (1994) and Remez (2005)]. In the context of the F2C paradigm, there is evidence that this aspect of coherence does not govern the impact of an extraneous formant on intelligibility. Specifically, an F2C whose frequency contour was derived by inverting and rescaling the frequency contour of F2 (plausibly speech-like) had the same impact as depth-matched versions which instead followed a stylized triangular contour [not plausibly speech-like; see Roberts *et al.* (2014)]. Other researchers use the term coherence to refer to more concrete properties of the stimulus such as the extent to which change over time is continuous or discontinuous, a factor assumed to influence the likelihood that the stimulus will be heard as a single stream [see Bregman (1990)]. Although we acknowledge that the two approaches to perceptual coherence may share some common roots, in the study reported here we intend this latter meaning. Notably, several studies have indicated a critical role for formant transitions and continuity of the pitch contour in holding together the acoustically diverse and rapidly changing speech signal as a coherent stream (Cole and Scott, 1973; Dorman *et al.*, 1975; Darwin and Bethell-Fox, 1977; Stachurski *et al.*, 2015).

The importance of maintaining the coherence of target speech as a single stream is self-evident—without it intelligibility will be impaired—but it is less obvious how the degree of spectro-temporal coherence of a speech-like interferer might influence the IM it produces. Consider, for example, the possible effects of dividing F2C into segments and shuffling their order, a manipulation that introduces abrupt discontinuities into the formant-frequency contour. These discontinuities might lead to an increase in the extent of IM for one or more of three reasons. First, although the overall frequency range is unaffected by shuffling the segments, some or all of the discontinuities may be perceived as frequency jumps, akin to adding rapid formant transitions. In effect, this would be like increasing the overall amount of formant-frequency change in the interferer, which is a factor already known to cause greater IM. Second, they may be perceptually salient and so likely to draw attention away from the target. Third, they may compromise the coherence of the interferer, such that it becomes harder to hold together, preventing it from being ignored or rejected as a whole. On the other hand, segmentation and reordering of F2C reduces target-masker similarity in terms of their sequential properties (smooth and continuous change vs abrupt discontinuities). There is evidence from studies of IM using non-speech stimuli that at least some types of qualitative differences between target and masker can lead to partial release from IM, such as differences in source properties [e.g., pure tone vs narrow-band noise; see Neff (1995)] and direction of frequency sweep (Durlach *et al.*, 2003).

As noted by Roberts *et al.* (2014), the properties of stimuli typically used in the F2C paradigm differ in important ways from those often used in studies of IM for non-speech stimuli. The latter usually employ narrowband targets and broadband maskers, and often use the multitone, multi-burst paradigm [Neff and Green (1987); for a review see Kidd *et al.* (2008)], in which maskers are constructed by concatenating a sequence of multiple tone bursts such that frequency variation in the masker across time is generated by making an independent constrained-random draw of frequencies for each successive burst. Hence, increases in the extent of masker frequency variation across time are confounded with decreases in its overall spectro-temporal coherence, arising from larger discontinuities on average between adjacent segments of the stimulus. In contrast, our previous studies using the F2C paradigm, or variants thereof, have involved changing the amount of frequency variation in the target and competitor formants while preserving formant-frequency contours with coherent trajectories. Continuous trajectories inevitably have a fairly high degree of predictability from moment-to-moment, but this property would be disrupted by breaking the formant-frequency contour into segments and shuffling the order of those segments.

The experiments reported here extend our previous work on the effects of manipulating the depth of formant-frequency variation in a contralateral single-formant interferer (F2C) by exploring the effects of segmenting F2C and then randomly ordering the segments, with a view to establishing whether masker impact depends not only on the amount of formant-frequency variation in the interferer but also on its spectro-temporal coherence. While coherence of the target is clearly important for phoneme perception, and for maintaining a one-stream percept, coherence of the extraneous formant may not necessarily determine how much interference it causes. Cutting F2C into segments also offers a convenient method of exploring which aspects of frequency change over time are important for IM. If the velocity of formant transitions in the interferer is important, as well as the overall range of frequency variation, then replacing the time-varying frequency contour of each segment of F2C with a piecewise constant value may substantially reduce F2C impact on intelligibility, despite the segment-by-segment changes present in the distribution of energy across frequency.

II. EXPERIMENT 1

This experiment focused on the effects of manipulating the spectro-temporal coherence of the extraneous formant, F2C, on the IM arising when it accompanies a target sentence in the contralateral ear. The effects of three factors were investigated. First, the effect of segmentation *per se* was explored by comparing the impact on intelligibility of a continuous F2C with cases where F2C was divided into segments of equal duration without reordering. This comparison is important because the manipulation imposes a regular pulsatile structure on F2C without introducing abrupt discontinuities in formant frequency. As noted above, it is possible that the impact of F2C may nonetheless be affected because the perceptual salience of these changes grabs the attention of listeners or because these changes lower the similarity in timbre between F2C and the target formants. Second, the effect of introducing sudden discontinuities into the F2C frequency contour was explored by comparing the effects of presenting the segments in the correct order or in randomized order. Third, the extent to which any effect of introducing these discontinuities on F2C impact is mediated by stream segregation, rather than by some other aspect of overall spectro-temporal coherence, was explored by comparing the effects of segment reordering when F2C is divided into either shorter or longer segments. Studies using repeating sequences of pure tones have shown that large differences in frequency between consecutive tones lead to strong and obligatory stream segregation for a 100-ms tone repetition time, but that there is little tendency towards obligatory stream segregation when the tone repetition time is increased to 200 ms (Bregman and Campbell, 1971; van Noorden, 1975).

A. Method

1. Listeners

All listeners were students or staff members at Aston University and received either course credits or payment 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 in either ear. 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 passed the training and took part in the main experiment, but five did not meet the additional criterion of a mean score of $\geq$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. Twenty-four listeners (two males) successfully completed the experiment (mean age = 25.2 years, range = 18.5–52.4). 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.

https://doi.org/10.1121/10.0002359

2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of a collection of short sentences spoken by a male talker with "Received Pronunciation," which is the accent traditionally regarded as the standard for British English. The text for these sentences was provided by Patel and Morse (2010) and consisted of variants created by rearranging words in sentences taken from the Bamford-Kowal-Bench (BKB) lists (Bench *et al.*, 1979) while maintaining semantic simplicity. To enhance the intelligibility of the synthetic analogues, the 48 sentences used were 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 British male talker with a similar accent 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 for the corrected formant frequencies were extracted automatically from the stimulus spectrograms. To facilitate creation of the interferers (see below), the set of formant contours for each sentence was slightly compressed or expanded in time to ensure that the target duration was set to the nearest multiple of 200 ms and up-scaled to a sampling rate of 40 kHz using linear interpolation.

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 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 (Rosenberg, 1971). 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 in the main experiment, the "standard" second-formant competitor (F2C) was created by inverting the F2 frequency contour about its geometric mean and setting the amplitude to a constant root-mean-square (RMS) level matching that of the target F2. Roberts and Summers (2015) have shown that RMS-matched constant-amplitude competitors have at least as much impact on the intelligibility of monaural targets as competitors with time-varying amplitude contours. Four other competitors were derived from the standard case by segmenting the amplitude contour such that, at the junction of each abutting segment, the amplitude fell to zero and rose back to constant over 20 ms (2 × 10-ms raised-cosine envelope). Segment duration was either 100 or 200 ms; these values were informed by typical syllable durations, by studies of the effects of tone repetition time on auditory stream segregation, and by the need to have a sufficient number of segments given the relatively short duration of the sentences (∼2 s). These segments were presented either in the correct order (coherent F2C frequency contour) or the order of the segments was randomized subject to the constraint that no two consecutive segments were in their original order (incoherent contour). For the random-order conditions, segment ordinal positions were shuffled anew for each rotation of sentence allocations (see below). Overall, the mean absolute difference in frequency at the boundary between adjacent segments in the incoherent sequences was 5.0 semitones (ST) (range = 2.6–9.9) for the 100-ms segments and 5.1 ST (range = 1.3–11.1) for the 200-ms segments. All competitors were rendered as the outputs of a second-order resonator. The excitation source, F0 (140 Hz), 3-dB bandwidth (70 Hz), and synthesizer configuration were identical to those used to synthesize the target F2. When present, F2C was always sent to the ear contralateral to that receiving the target speech.

There were eight conditions in the main experiment [see part (a) of Table I]. Two of the conditions were controls (C1 and C2), for which the target F2 was absent; C2 also included the continuous (i.e., unsegmented) competitor. There were five experimental conditions (C3–C7), for which the stimuli comprised the target formants accompanied by one of the versions of F2C; the competitor was either continuous (C3) or segmented (C4–C7, either in order or shuffled). Examples of the competitors used can be found in the supplementary material.[1] The final condition (C8) was the reference case, for which only the monaural target formants were presented. The frequency and amplitude contours of the competitors used in each experimental condition—including their segmentation—are illustrated schematically in Fig. 1. For each listener, the 48 sentences were divided equally across conditions (i.e., six per condition), such that there were 18 or 19 keywords per condition. Allocation of sentences to conditions was counterbalanced by rotation across

TABLE I. Part (a) summarizes the stimulus properties for the conditions used in experiment 1 (main session). The formant-frequency contour of each version of F2C was derived from the contour created by inverting that of the target F2 about its geometric mean. The amplitude contour of the unsegmented version (∞) was constant and matched to the RMS level of F2. For the other versions, the amplitude contour was also constant but was divided into segments (duration = 100 or 200 ms) and, defined by these segments, the frequency contour could be presented in the correct order (coherent) or in random order. Part (b) summarizes the mean results for the corresponding conditions.

	Part (a)				Part (b)	
Condition	Stimulus configuration (target ear; other ear)	F2C amplitude contour	F2C segment duration (ms)	F2C segment order	Mean keyword scores	Mean phoneme scores
C1	(F1+F3; —)	—	—	—	15.4%	34.8%
C2	(F1+F3; F2C)	Continuous	∞	—	6.1%	25.8%
C3	(F1+F2+F3; F2C)	Continuous	∞	—	40.4%	57.4%
C4	(F1+F2+F3; F2C)	Segmented	100	Correct	41.7%	59.9%
C5	(F1+F2+F3; F2C)	Segmented	100	Random	39.2%	56.7%
C6	(F1+F2+F3; F2C)	Segmented	200	Correct	37.8%	55.6%
C7	(F1+F2+F3; F2C)	Segmented	200	Random	38.1%	54.2%
C8	(F1+F2+F3; —)	—	—	—	55.6%	70.2%

each set of eight listeners tested. Hence, the total number of listeners needed to produce a balanced dataset was a multiple of 8.

3. Procedure

During testing, listeners were seated in front of a computer screen and a keyboard in a double-walled sound-attenuating chamber (Industrial Acoustics 1201A, Winchester, UK). The experiment consisted of a training session followed by the main session and typically took ∼50–60 min to complete; listeners were free to take a break whenever they wished. In both parts of the experiment, trials were presented in a newly randomized order for each listener.

The training session comprised 50 trials; stimuli were presented without interferers and a new sentence was used for each trial. Diotic presentation was used for the first 40 trials. For the first ten of these trials, listeners heard 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 the synthetic version of a given sentence, which they were asked to transcribe using the keyboard. They were allowed to listen to the stimulus up to six times before entering 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. The strategy of providing feedback using alternating presentations of the synthetic and original versions provides an efficient way of enhancing the perceptual learning of speech-like stimuli (Davis *et al.*, 2005).

During the final ten training trials, the target sentence 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) ≥50% keywords correct across all 40 trials requiring a transcription (30 diotic with repeat listening; 10 monaural with random selection of ear and 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 was selected randomly; F2C (when present) was always presented in the other ear. Listeners were allowed to hear each stimulus only once before entering their transcription and no feedback was given.

All speech analogues were synthesized using MITSYN (Henke, 2005) at a sample rate of 40 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), a pair of 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 reference level (long term average) of 75 dB sound pressure level (SPL); there was some variation in the ear receiving F2C (mean ≈ 69 dB SPL) depending on the RMS power of the target F2 from which F2C was derived. In the training session, the presentation level of the diotic materials used (first 40 trials) was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation. The final ten sentences in the training session were presented monaurally at the reference level.

4. Data analysis

For each listener, stimulus intelligibility was quantified using keyword scoring as the primary measure. 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. Responses were classified using tight scoring, in which a response is scored as correct only if it matches the keyword exactly; obvious misspellings

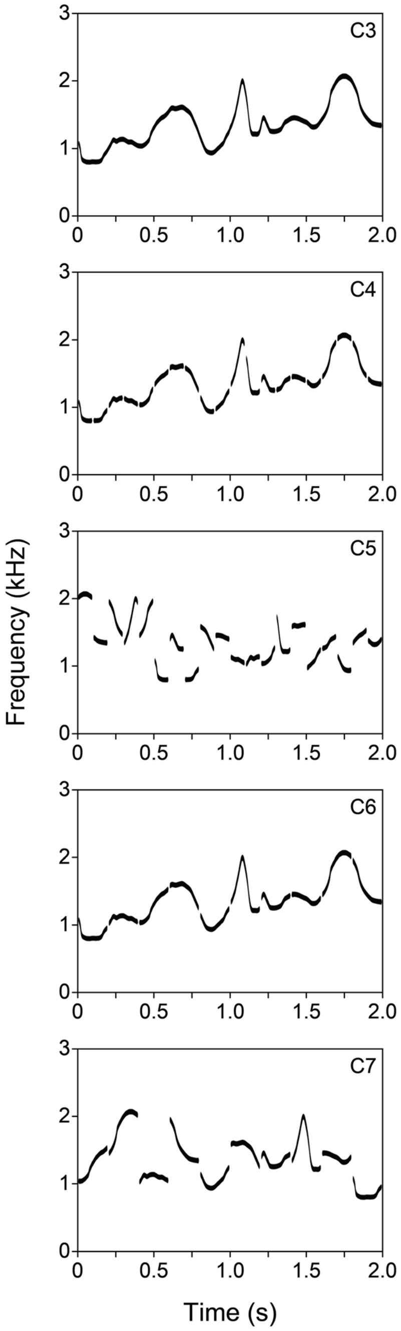

FIG. 1. Stimuli for experiment 1—schematic spectrograms illustrating formant-frequency contours and segmentation for the five versions of the competitor derived from the example sentence "The cat lies on the ground" (conditions C3-C7). The top panel illustrates the coherent and unsegmented version of F2C (C3), which was created using the inverted F2 frequency contour and a constant amplitude RMS-matched to the target F2. The next two panels illustrate the versions of F2C where it was divided into 100-ms-long segments, presented either in the correct order (C4) or random order (C5). The final two panels illustrate the versions of F2C where it was divided into 200-ms-long segments, presented either in the correct order (C6) or random order (C7). For all competitors involving division into segments (C4-C7), the constant amplitude contour was divided such that each segment was defined by raised-cosine amplitude ramps of ±10 ms.

or typographical errors were corrected and homonyms were accepted [see, e.g., Foster *et al.* (1993) and Roberts *et al.* (2010)]. Following Roberts *et al.* (2014), phoneme scoring was used as an additional measure of intelligibility. Typed responses were converted automatically into phoneme strings 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 insertions, substitutions, and deletions as required [see Needleman and Wunsch (1970)]. The mean percentage of phonemes correctly identified—the phoneme score—is defined as 100 × (number of correctly aligned phonemes)/(number of phonemes in the original sentence).

All statistical analyses were computed using R 3.5.3 (R Core Team, 2019) and the *ez* analysis package (Lawrence, 2016). To provide a more sensitive analysis than standard within-subjects analysis of variance (ANOVA), linear mixed-effects models were used to take into account random effects arising not only from differences in performance between listeners but also from differences in intelligibility between sentences. The significance of the effects of the experimental factors was evaluated following the approach of Luke (2017); the package *lmerTest* (Kuznetsova *et al.*, 2017) was used for its implementation of the Satterthwaite approximation to estimate the degrees of freedom of the denominator term in the F statistic. Standardized effect sizes are not reported because, owing to the way variance is partitioned in linear mixed-effects models, there is no agreed method of computing them for individual model terms such as main effects or interactions [see, e.g., Rights and Sterba (2019)]. All *a posteriori* pairwise comparisons (two tailed) were computed using the package *multcomp* (Hothorn *et al.*, 2008) and the estimated degrees of freedom; these comparisons were evaluated using the restricted least-significant-difference test (Snedecor and Cochran, 1967; Keppel and Wickens, 2004). Unless otherwise stated, all statistics reported here were computed using keyword scores; statistics computed using phoneme scores are presented only when the two measures disagree on whether or not a given comparison was significant. This happened only on one occasion.

B. Results and discussion

Figure 2 shows the mean percentage scores (and intersubject standard errors) across conditions for keywords correctly identified. The black, gray, and white bars indicate the results for the control, experimental (i.e., target-plus-interferer), and target-only conditions, respectively; for the experimental conditions, dark and light gray bars indicate the results for the coherent- and random-order cases, respectively. For ease of comparison, part (b) of Table I presents the mean keyword scores and corresponding mean phoneme scores side by side; the two sets of scores follow a similar pattern across conditions. A linear mixed-effects model corresponding to a one-way within-subjects ANOVA across

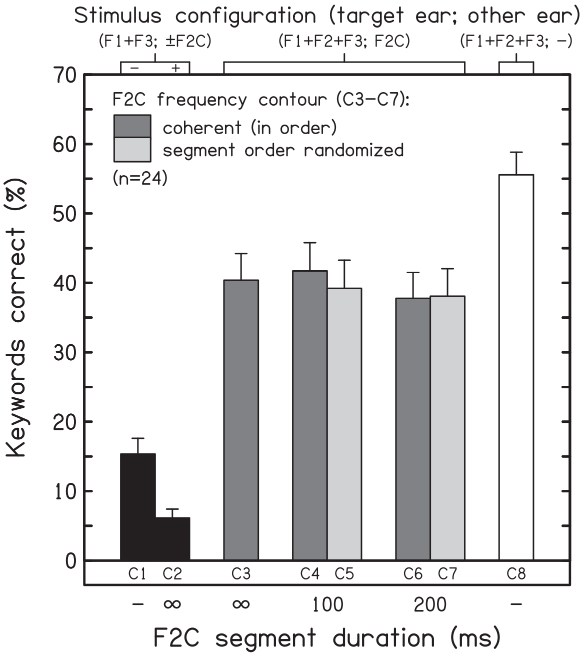

FIG. 2. Results for experiment 1—effects of segment order (coherent vs random) and duration (100 vs 200 ms) on the impact of competitors (F2Cs) on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and inter-subject standard errors (n = 24) are shown for the control conditions (black bars), the target-plus-competitor conditions (gray bars; dark = F2C segments in correct order, light = F2C segments in random order), and the target-only reference condition (white bar). The top axis indicates which formants were presented to each ear; the bottom axis indicates segment duration for F2C (when present). The symbol ∞ indicates indefinite duration (i.e., unsegmented). For ease of reference, condition numbers are included immediately above the bottom axis.

all eight conditions showed a highly significant effect of condition on intelligibility [F(7, 1074.0) = 48.954, p < 0.001]. As expected, the control conditions indicated that intelligibility was low for F1+F3 alone (C1; mean keywords correct = 15.4%), and near floor when the standard unsegmented competitor was added to F1+F3 in the absence of the target F2 (C2; mean keywords correct = 6.1%), indicating that F2C was not a good surrogate for F2 in supporting intelligibility. Pairwise comparisons indicated that the scores for the two control conditions differed from those for all other conditions, including each other (range: p = 0.003–p < 0.001).

Intelligibility of the target speech in the reference condition (C8; 55.6% keywords correct) was as might be expected given the simple source properties and three-formant parallel vocal-tract model used to synthesize the sentences. All versions of F2C were effective interferers; intelligibility was lowered significantly relative to the reference case when the monaural target (F1+F2+F3) was accompanied by any of the contralateral interferers (C3–C7; overall mean fall = 16.1 percentage points [% pts]; p < 0.001 in all cases). A linear mixed-effects model corresponding to a one-way within-subjects ANOVA restricted to the five experimental conditions (C3–C7) revealed neither a significant effect of condition nor a trend towards one [F(4, 645.32) = 0.851, p = 0.493]. Critically, the relatively flat profile of scores across the experimental conditions suggests that the impact of F2C was largely unaffected either by segmentation of the amplitude contour (unbroken vs divided into abutting 100- or 200-ms segments) or by segment order (coherent vs incoherent). Clearly, introducing sudden discontinuities into the formant-frequency contour between adjacent segments does not change the extent of interference that F2C produces, regardless of whether the segments are short enough to be likely to influence their perceptual organization. The outcomes of the supplementary analysis using the phoneme scores were fully consistent with the main analysis.

III. EXPERIMENT 2

The results of experiment 1 suggest that informational masking of target speech by the contralateral competitor does not depend on masker spectro-temporal coherence. Experiment 2 evaluated this interpretation further by manipulating not only F2C segment order but also the overall extent of formant-frequency variation in F2C, a factor which has been shown previously to be critically important in determining the impact of F2C on target intelligibility (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015, 2018). One aspect of this evaluation was to examine whether the apparent absence of an effect of segment order was in fact a ceiling effect arising from the use of full-depth F2C frequency contours in experiment 1. For example, it is known that increasing F2C depth from 100% to 200% has relatively little additional impact on target intelligibility (Roberts and Summers, 2015). The other aspect was to explore whether within-segment changes in formant-frequency contour *per se* are important in determining competitor impact or whether F2C impact depends primarily on the overall amount of formant-frequency variation across the whole sequence of segments. This issue was explored using conditions in which the time-varying frequency contour of each segment was replaced with a locally constant value.

A. Method

Except where described, the same method was used as for experiment 1. Forty-four listeners (12 males) passed the training and successfully completed the experiment (mean age = 21.1 years, range = 18.3–43.8); none of these listeners took part in experiment 1. Only one listener who successfully passed the training needed to be replaced for failing to meet the additional criterion of at least 20% keywords correct in the main experiment. The training session was identical to that used in experiment 1; no extraneous formants were presented. The stimuli for the main experiment were derived from 66 sentences drawn from the same collection of recordings as used in experiment 1. Testing took place in a single-walled sound-attenuating chamber (Industrial Acoustics 401A, Winchester, UK) housed within a quiet room.

Given that no effect of F2C segment duration (100 vs 200 ms) was found in experiment 1, it was possible here to reduce the number of conditions by using a segment duration of 150 ms. Hence, the set of formant contours for each target sentence was adjusted to ensure that the stimulus duration was set to the nearest multiple of 150 ms prior to creating the target and competitors. For each version of F2C, the constant amplitude contour was divided every 150 ms and each segment was defined by ±10 ms rise/fall ramps. As for experiment 1, the F2C frequency contour was derived from the inverted F2 contour and the constant amplitude used for each version of F2C was RMS-matched to that of the corresponding target F2.

A set of nine competitors was created for each sentence in the main experiment; these are illustrated schematically in Fig. 3. The effects of three factors on the impact of segmented interferers on target intelligibility were explored—the depth of F2C frequency variation [0% (i.e., constant), 50%, or 100%; see Roberts *et al.* (2014), Roberts and Summers (2015, 2018)], segment type (frequency-varying vs constant), and segment order (coherent vs random). Rescaling of the F2C formant-frequency contour was performed on a log scale about the geometric mean frequency. The rescaled frequency at time t, $s(t)$, is given by

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

where x ($0 \leq x \leq 1$) is a proportional scale factor determining the maximum frequency range relative to the original range for the target F2 (the natural depth of variation), $f(t)$ is the formant frequency at time t, and g is the geometric mean of the whole formant-frequency contour. Contour segmentation was applied after rescaling, following which the frequency contour of each segment was either left as time-varying or was set to a constant value at the local geometric mean frequency for that segment. Both segment types were presented either in the correct order (coherent) or in shuffled order (incoherent). The 0%-depth F2C case provides a single comparator for the effect of all these manipulations, because there is no distinction of type or order when all F2C segments have the same constant formant frequency. The coherent versions of the 50%- and 100%-depth constant-frequency conditions are somewhat analogous to pointillistic speech, in which the speech signal is replaced by a matrix of steady tone bursts presented in consecutive order (Kidd *et al.*, 2009).

For competitors composed of frequency-varying segments, there were no discontinuities in frequency between adjacent segments when presented in the correct order, irrespective of F2C depth. When the frequency-varying segments were presented in random order, the mean absolute difference in frequency between adjacent segments was 2.8 ST (range = 1.0–5.7) when F2C depth was 50% and 5.6 ST (range = 1.7–12.5) when F2C depth was 100%. For competitors composed of constant segments in the correct order, the mean absolute difference in frequency between adjacent segments was 1.7 ST (range = 0.9–2.8) when F2C depth was 50% and 3.4 ST (range = 1.8–5.6) when F2C depth was 100%. When the constant segments were presented in random order, the mean absolute difference in frequency between adjacent segments was 2.5 ST (range = 0.9–5.2) when F2C depth was 50% and 5.0 ST (range = 2.0–11.1) when F2C depth was 100%.

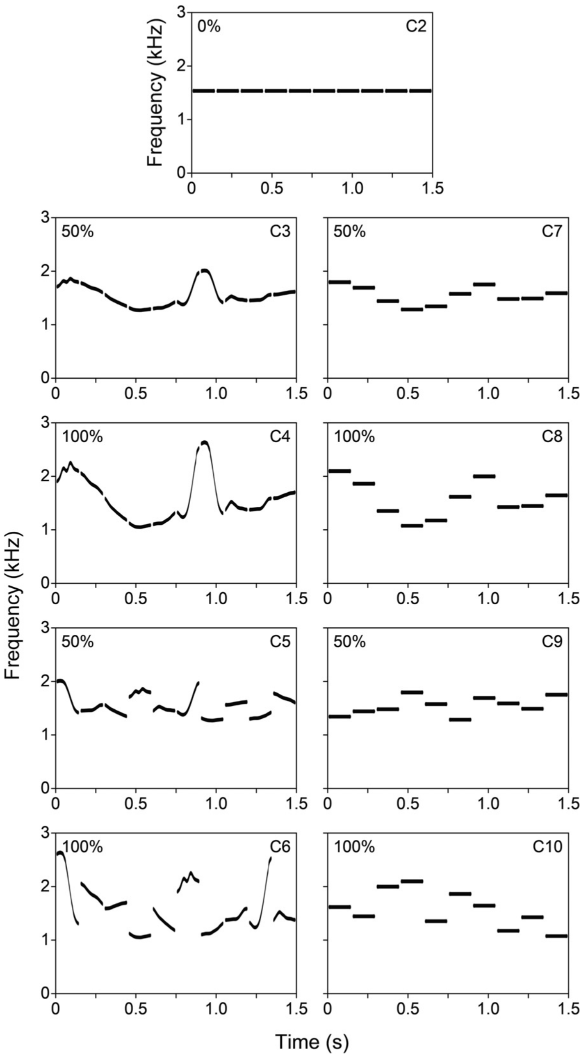

FIG. 3. Stimuli for experiment 2—schematic spectrograms illustrating formant-frequency contours and segmentation for the nine versions of the competitor derived from the example sentence "Mother ate the oranges" (conditions C2–C10). For all competitors, the constant amplitude contour was RMS-matched to the target F2 and divided into 150-ms-long segments; each segment was defined by raised-cosine amplitude ramps of ±10 ms. In each panel, the depth of the example F2C and the condition from which it was drawn are identified. The top panel (C2) illustrates the case where the F2C frequency contour was set to a constant value at the geometric mean frequency for the target F2 (constant segments, 0% depth). The four lower panels on the left (C3–C6) illustrate the versions of F2C in which the time-varying inverted F2 frequency contour was used after scaling it relative to that for the target F2 (50% or 100% depth) and the segments were presented either in the correct order or random order. The four lower panels on the right (C7–C10) differ from their counterparts on the left only in that each segment was set to a constant value at the geometric mean frequency for that segment.

There were eleven conditions in the main experiment [see part (a) of Table II]. One condition (C1) was a control for which the target F2 was absent. Nine conditions (C2–C10) were experimental, for which the target formants

TABLE II. Part (a) summarizes the stimulus properties for the conditions used in experiment 2 (main session). The depth of F2C frequency variation refers to the scale factor applied to its frequency contour. A scale factor of 0% indicates a constant frequency for F2C, corresponding to the geometric mean frequency of the target counterpart. Note that there is no distinction between frequency-varying and constant cases or between in-order and random-order cases when depth = 0% (C2). The amplitude contour (constant and matched to the RMS level of F2) was divided into 150-ms-long segments, and the frequency contour of each segment was either time-varying or set to a constant frequency at the geometric mean for that segment. Defined by these segments, the frequency contour could be presented in the correct order (coherent) or in random order. Part (b) summarizes the mean results for the corresponding conditions.

Part (a)					Part (b)	
Condition	Stimulus configuration (target ear; other ear)	Depth of F2C formant-frequency variation (%)	F2C segment type	F2C segment order	Mean keyword scores	Mean phoneme scores
C1	(F1+F3; —)	—	—	—	20.0%	36.0%
C2	(F1+F2+F3; F2C)	0	(Constant)	(Correct)	55.2%	68.9%
C3	(F1+F2+F3; F2C)	50	Frequency-varying	Correct	51.3%	66.0%
C4	(F1+F2+F3; F2C)	100	Frequency-varying	Correct	44.7%	59.6%
C5	(F1+F2+F3; F2C)	50	Frequency-varying	Random	47.7%	62.1%
C6	(F1+F2+F3; F2C)	100	Frequency-varying	Random	43.8%	57.6%
C7	(F1+F2+F3; F2C)	50	Constant	Correct	50.1%	64.7%
C8	(F1+F2+F3; F2C)	100	Constant	Correct	50.7%	65.0%
C9	(F1+F2+F3; F2C)	50	Constant	Random	53.4%	68.3%
C10	(F1+F2+F3; F2C)	100	Constant	Random	48.5%	63.4%
C11	(F1+F2+F3; —)	—	—	—	58.6%	73.1%

were accompanied by one of the nine versions of the competitor in the contralateral ear, including manipulations of the depth of F2C frequency variation, segment type, and segment order. Examples of the competitors used can be found in the supplementary material.[1] The final condition (C11) was the reference case, for which only the target formants were presented (i.e., no interfering formant). For each listener, the 66 sentences were divided equally across conditions (i.e., six per condition), such that there were 18 or 19 keywords per condition. Allocation of sentences was counterbalanced by rotation across each set of eleven listeners tested.

B. Results and discussion

Figure 4 shows the mean percentage keyword scores (and intersubject standard errors) across conditions. Black, gray, and white bars indicate the results for the control, experimental, and target-only conditions, respectively. Within the set of experimental conditions, the results for conditions involving F2Cs composed of frequency-varying or constant segments are shown on the left and right, respectively; dark and light gray bars display the results for the coherent- and random-order cases, respectively. Part (b) of Table II presents the mean keyword scores and corresponding mean phoneme scores; once again, the two sets of scores follow a similar pattern across conditions. A linear mixed-effects model corresponding to a one-way within-subjects ANOVA across all eleven conditions showed a highly significant effect of condition on intelligibility [F(10, 2785.0) = 34.366, $p < 0.001$]. Results for the control condition indicate that intelligibility was low for F1+F3 alone (C1; mean keywords correct = 20.0%); pairwise comparisons showed that this score was significantly different from those for all other conditions ($p < 0.001$ in all cases). As expected, performance was best when all three target formants were presented alone (C11; mean keywords correct = 58.6%). The profile of scores across conditions suggests that there was considerable variation in the impact of the various competitors used in the experimental conditions.

The effect of the different experimental manipulations of the interfering formants was explored using a linear mixed-effects model corresponding to a three-way within-subjects ANOVA restricted to the nine target-plus-interferer conditions (C2–C10). The three factors manipulated were the depth of formant-frequency variation in F2C (three levels: 0%, 50%, or 100%), F2C segment type (two levels: frequency-varying vs constant), and F2C segment order (two levels: correct vs random).[2] This analysis revealed significant main effects of F2C depth [F(2, 3047.9) = 25.780, $p < 0.001$] and segment type [F(1, 3047.9) = 6.920, p = 0.009], but there was no effect of F2C segment order [F(1, 3047.9) = 0.788, p = 0.375]. Furthermore, none of the interactions involving segment order were significant [Depth × Order: F(2, 3047.9) = 0.293, p = 0.746; Type × Order: F(1, 3047.9) = 0.769, p = 0.381; Depth × Type × Order: F(2, 3047.9) = 1.588, p = 0.205]. The final interaction term approached but did not quite reach significance for the keyword scores [Depth × Type: F(2, 3047.9) = 2.540, p = 0.079], but the interaction was significant for the phoneme scores [Depth × Type: F(2, 3048.0) = 4.367 p = 0.013]. In all other respects, the outcomes of the supplementary analysis using the phoneme scores were fully consistent with the main analysis.

The effects of F2C depth, F2C segment type, and their interaction were explored further using mixed-effects models computed only for those factors; mean changes in keyword scores reported below are those computed from the

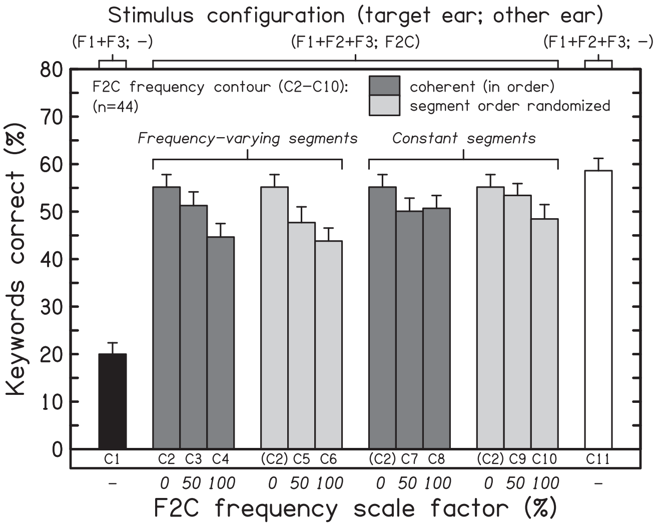

FIG. 4. Results for experiment 2—effects of segment type (frequency-varying vs constant), segment order (coherent vs random), and depth of formant-frequency variation (0%, 50%, or 100%) on the impact of competitors (F2Cs) on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and inter-subject standard errors (n = 44) are shown for the control condition (black bar), the target-plus-competitor conditions (gray bars; dark = F2C segments in correct order, light = F2C segments in random order), and the target-only reference condition (white bar). 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). For the target-plus-competitor conditions (gray bars), the results for interferers with frequency-varying segments and with constant segments are shown on the left- and on the right-hand side, respectively. For ease of reference, condition numbers are included immediately above the bottom axis. Note that the results for C2 are shown four times, because there are no distinctions between different segment types or order for the 0%-depth case and so C2 acts as the comparator case for all four combinations of these factors for the 50%- and 100%-depth cases.

underlying data rather than the estimates provided by the mixed-effects models. Relative to the target-only reference case (C11), adding an F2C with frequency-varying segments (C3–C6) had a significant impact on target intelligibility for an F2C depth of 50% [mean fall = 9.1% pts, t(1209.3) = 3.950, p < 0.001) and 100% (mean fall = 14.4% pts, t(1209.3) = 6.657, p < 0.001]. Adding an F2C with constant segments (C7–C10) also had a significant impact on target intelligibility for an F2C depth of 50% [mean fall = 6.9% pts, t(1209.7) = 2.856, p = 0.004] and 100% (mean fall = 9.1% pts, t(1209.7) = 4.122, p < 0.001), but adding an F2C with 0% depth—i.e., constant frequency throughout—did not (C2; mean fall = 3.5% pts, t(2785.0) = 1.109, p = 0.268). For competitors composed of frequency-varying segments, all pairwise comparisons between levels of F2C depth were significant [0% vs 50%, mean fall = 5.7% pts, t(1209.3) = 2.650, p = 0.008; 50% vs 100%, mean fall = 5.3% pts, t(1209.3) = 3.310, p < 0.001; 0% vs 100%, mean fall = 10.9% pts, t(1209.3) = 5.352, p<0.001]. For competitors composed of constant segments, only the comparison between 0% and 100% depth was significant [0% vs 50%, mean fall = 3.4% pts, t(1208.9) = 1.553, p = 0.121; 50% vs 100%, mean fall = 2.2% pts, t(1208.9) = 1.549, p = 0.122; 0% vs 100%, mean fall = 5.6% pts, t(1208.9) = 2.818, p = 0.005]. Pairwise comparisons between the effects of frequency-varying and constant F2C segments for corresponding F2C depths revealed a significant effect of segment type at 100% depth [mean difference = 5.3% pts, t(947.0) = 3.213, p = 0.001 but not at 50% depth [mean difference = 2.3% pts, t(947.0) = 1.367, p = 0.172]. Overall, this pattern suggests that constant-segment F2Cs are less effective as interferers than F2Cs with frequency-varying segments and that the difference in impact grows as F2C depth increases.

Two aspects of these outcomes merit comment here. First, it must be acknowledged that there are two differences between F2Cs composed of frequency-varying or constant segments that may contribute to the reduced impact of the latter. As well as the obvious difference—that frequency-varying segments involve formant transitions but constant ones do not—it is also the case that the range of frequency variation will inevitably be smaller for the constant-segment F2Cs used here than for their nominally depth-matched counterparts. This issue is considered further in Sec. IV. Second, the presence of a 0%-depth segmented F2C had only a small (and insignificant) impact on keyword scores. Only one of our previous studies included a condition in which the effect on intelligibility of a 0%-depth F2C with a constant amplitude contour RMS-matched to F2 was tested. Roberts and Summers (2015) (experiment 1) used a 0%-depth F2C that was not segmented but was otherwise similar to that used here. They also found that the 0%-depth F2C had less impact on intelligibility than the greater-depth cases tested, but nonetheless it had a larger (and significant) impact than did the corresponding case here. Albeit that some caution is needed when comparing results across different studies and listeners, it seems reasonable to conclude that 0%-depth segmented F2Cs are not more effective as interferers than their unsegmented counterparts used in our previous studies. This result is interesting because it was apparent from informal observations that the pulsatile nature of the segmented F2Cs made them more salient than their continuous counterparts but nonetheless this property did not increase the IM produced by them. Indeed, our previous study in which 100%-depth monaural three-formant target sentences were accompanied by 100%-depth contralateral F2Cs showed effects for continuous competitors that were as large, or larger, than those for the corresponding segmented case here (Roberts and Summers, 2015).

IV. GENERAL DISCUSSION

The presence of interfering speech or a speech-like masker can cause considerable IM even when the target speech is in the contralateral ear (Gallun *et al.*, 2007). Although it should be acknowledged that a contralateral sound can have effects other than IM, for example inducing cochlear gain reduction in the other ear by activating the medial olivocochlear reflex [see, e.g., Lopez-Poveda (2018)], these effects are unlikely to reduce intelligibility unless the level of the target speech is considerably below that of the interferer [cf. Roberts and Summers (2019)]. The results of the experiments reported here are in accord with previous findings that the depth of variation in the frequency contour of an extraneous formant (F2C), acting as an informational masker, strongly affects its impact on the intelligibility of a three-formant analogue of a target sentence (Roberts *et al.*, 2014; Roberts and Summers, 2015). Furthermore, the results extend those findings by indicating that formant transitions (i.e., within-segment changes) constitute an important part of that variation for generating IM. Critically, however, the impact of F2C on the intelligibility of target sentences did not depend on either the segmentation of the amplitude contour (unbroken vs divided into 100–200-ms segments) or the randomization of segment order (coherent vs incoherent F2C frequency contour). These findings are considered in turn.

Substituting the frequency-varying segments of F2C with the corresponding constant-frequency segments in experiment 2 lowered the impact on keyword scores by about half. The effect of segment type has two potential origins. First, as noted above, the method used to produce the constant segments involved setting the formant frequency of each segment to its local geometric mean. This inevitably reduces the overall formant-frequency range for F2C, in effect decreasing its depth. Second, the manipulation results in the loss of formant-frequency change within individual segments. To estimate qualitatively the extent to which the smaller range contributed to the loss of F2C impact in the constant-segment conditions, the mean minimum-to-maximum frequency range was computed across our set of competitors for each combination of segment type and depth (50% or 100%). For the frequency-varying segments, these ranges were 8.7 and 17.4 ST, and for the constant segments these ranges were 6.6 and 13.1 ST, respectively. These values indicate that, on average, the formant-frequency range for the constant-segment F2Cs was around 75% that of their frequency-varying counterparts, whereas the impact of the constant-segment F2Cs on target intelligibility was only around half that of their frequency-varying counterparts. Furthermore, the mean impact of the 50%-depth competitors with frequency-varying segments fell within 0.1% pts of that for the 100%-depth competitors with constant segments, despite the substantially larger formant-frequency range of the latter.

Previous studies have suggested a linear increase in the impact of F2C on keyword scores as F2C frequency variation is increased over the range 0% to 100% depth (Roberts *et al.*, 2014; Roberts and Summers, 2015). Therefore, although not conclusive, the outcome reported here for experiment 2 suggests that frequency variation, as well as frequency range, is probably an important contributor to the observed difference between the impact of constant-segment F2Cs and their frequency-varying counterparts. As noted earlier, increasing either the rate or depth of formant-frequency variation for an extraneous formant (within broad limits) increases the amount of IM that it causes (Summers *et al.*, 2012; Roberts *et al.*, 2014; Roberts and Summers, 2015, 2018). Both manipulations increase the velocity of transitions in an interferer's formant-frequency contours, implying more rapid movement of the articulators. Although it is not known why extraneous formants with rapid formant transitions can have such an impact on speech intelligibility, there is a considerable body of evidence that the time-varying formant-frequency contours are particularly important in conveying acoustic-phonetic information [see, e.g., Stevens (1998)]. Therefore, it is possible that extraneous formants with rapid formant transitions are particularly likely to interfere with the extraction and integration of that information from target speech because they either incur a greater processing load (disrupting effect) or intrude more into the percept of the target sentence (corrupting effect). Future research might explore further the role of rapid transitions in causing IM by comparing the impact of an F2C composed only of constant-frequency segments with that of one in which those segments are linked with linear glides in formant frequency [cf. Dorman *et al.* (1975)].

The absence of evidence for any effect of segment order on the amount of IM generated by F2C is noteworthy in many respects. Although it is possible that the lack of an effect is specific to the segmentation strategy used here, there is no obvious reason why using segment durations outside the range tested (100–200 ms) or cutting F2C unevenly to give individual segments of different durations might have led to a different outcome (except perhaps if very short segments were used). It is also possible that segmenting and reordering F2C leads to a mixture of IM-boosting and IM-releasing effects that cancel each other out near-perfectly, but this seems rather unlikely. In terms of the possible effects of these manipulations identified in Sec. I, the results obtained suggest the following. First, the frequency jumps occurring at the abrupt discontinuities introduced by segmenting and reordering the F2C frequency contour are not a surrogate for linking the segments with linear formant transitions and so do not contribute to the overall amount of formant-frequency variation. Second, although the pulsatile nature of regularly segmented F2Cs (whether in the correct or random order) is perceptually salient, it does not make them more effective distractors and nor does any resulting fall in target-masker similarity cause a release from IM. Third, the contention that a discontinuous F2C might be harder to separate from target speech because it fragments into multiple streams is incorrect; note that this statement is concordant with the assertion of Bregman (1990) that effective separation of target and background sounds depends on their segregation from one another but not on the perceptual organization of those elements attributed to the background. When the current results are considered in conjunction with our earlier finding that the impact of F2C on intelligibility does not depend on whether its pattern is plausibly speech-like (Roberts *et al.*, 2014), it seems that neither the more abstract nor the more concrete aspects of the perceptual coherence of interfering formants are relevant to the amount of IM they produce.

Roberts and Summers (2018) noted that the factors governing the IM produced by a contralateral speech-like interferer have similarities with those governing the irrelevant sound effect (ISE), a form of cross-modal interference in which an acoustic distractor that participants are asked to ignore nonetheless impairs their serial recall of visually presented digits or words [for a review, see Ellermeier and Zimmer (2014)]. The key similarity is that frequency change in the distractor is necessary for the ISE to occur (Jones and Macken, 1993); amplitude change alone is insufficient (Tremblay and Jones, 1999). Furthermore, the ISE is usually greatest when the acoustic distractor is speech or has speech-like properties [e.g., Viswanathan *et al.* (2014)], but it can also be strong when instrumental music is used as the distractor. This implies that the complexity of spectro-temporal change is an important contributor to the ISE, a notion supported by recent evidence that the ISE increases with the number of channels used to vocode a background speech-like masker (Dorsi *et al.*, 2018). Nonetheless, a simple acoustic distractor in which frequency change arises from a sequence of four steady pure tones spaced at one-octave intervals and presented in random order—a stimulus bearing some similarities with our constant-segment interferers—produces a clear ISE (Jones and Macken, 1993). Taken together, these observations are sufficient to warrant speculation that a common underlying mechanism may be involved in IM and the ISE. Indeed, it may be informative for further research to include an experiment analogous to the one proposed above for the F2C paradigm, in which the ISE produced by a sequence of a small number of short steady tones (i.e., no within-segment variation) is compared directly with that produced when linear tone glides are used to link the consecutive tones [i.e., rapid transitions are present; cf. Bregman and Dannenbring (1973)].

In conclusion, the results of the experiments reported here indicate that the impact of extraneous formants acting as informational maskers on speech intelligibility depends critically on the overall amount of frequency variation in the interferer, but not on its spectro-temporal coherence. The effect of formant-frequency change over time appears to have two components, one corresponding to the range over which the variation occurs, the other corresponding to the velocity of the formant transitions present in the interferer. Differentiating further these two components may offer new insights into the ways in which the extraction and integration of acoustic-phonetic information carried by the formants of the target speech can be disrupted or corrupted by informational masking.

ACKNOWLEDGMENTS

This research was supported by Research Grant Nos. ES/K004905/1 (experiment 1) and ES/N014383/1 (experiment 2) from the Economic and Social Research Council (UK), awarded to B.R. To access the research data underlying this publication, see https://doi.org/10.17036/researchdata.aston.ac.uk.00000477. We thank Peter Bailey for his comments on earlier drafts of this paper, Rob Morse and Meghna Patel for providing the BKB-like sentences, and Quentin Summerfield for enunciating them. We are also grateful to Eric Morecambe for inspiring the design of our incoherent maskers.

[1] See supplementary material at https://doi.org/10.1121/10.0002359 for examples of the various versions of F2C used in this research.

[2] When F2C depth is set to 0%, there is no distinction for F2C segment type between frequency-varying and constant and for F2C segment order between correct and random. Therefore, the data for C2 were reproduced in the analysis for all four combinations of these factors.

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) [software package]," Institute of Phonetic Sciences, University of Amsterdam, The Netherlands, http://www.praat.org/ (Last viewed 15 September 2016).

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

Bregman, A. S., and Campbell, J. (**1971**). "Primary auditory stream segregation and perception of order in rapid sequences of tones," J. Exp. Psychol. **89**, 244–249.

Bregman, A. S., and Dannenbring, G. L. (**1973**). "The effect of continuity on auditory stream segregation," Percept. Psychophys. **13**, 308–312.

Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.

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

Darwin, C. J. (**2008**). "Listening to speech in the presence of other sounds," Philos. Trans. R. Soc. B **363**, 1011–1021.

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

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.

Dorman, M. F., Cutting, J. E., and Raphael, L. J. (**1975**). "Perception of temporal order in vowel sequences with and without formant transitions," J. Exp. Psychol. Hum. Percept. Perform. **1**, 121–129.

Dorsi, J., Viswanathan, N., Rosenblum, L. D., and Dias, J. W. (**2018**). "The role of speech fidelity in the irrelevant sound effect: Insights from noise-vocoded speech backgrounds," Q. J. Exp. Psychol. **71**, 2152–2161.

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

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., and Kidd, G., Jr. (**2003**). "Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity," J. Acoust. Soc. Am. **114**, 368–379.

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

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

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

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

Hothorn, T., Bretz, F., and Westfall, P. (**2008**). "Simultaneous inference in general parametric models," Biometrical J. **50**, 346–363.

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. **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).

Kidd, G., Jr., 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, Boston, MA), Vol. 29, pp. 143–189.

Kidd, G., Jr., Streeter, T. M., Ihlefeld, A., Maddox, R. K., and Mason, C. R. (**2009**). "The intelligibility of pointillistic speech," J. Acoust. Soc. Am. **126**, EL196–EL201.

Klatt, D. H. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. **67**, 971–995.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest package: Tests in linear mixed effects models," J. Stat. Softw. **82**, 1–26.

Lawrence, M. A. (**2016**). "ez: Easy analysis and visualization of factorial experiments (R package version 4.4-0) [software]," https://cran.r-project.org/package=ez (Last viewed 30 July 2018).

Lopez-Poveda, E. A. (**2018**). "Olivocochlear efferents in animals and humans: From anatomy to clinical relevance," Front. Neurol. **9**, 197.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cogn. 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. (**1995**). "Signal properties that reduce masking by simultaneous, random-frequency maskers," J. Acoust. Soc. Am. **98**, 1909–1920.

Neff, D. L., and Green, D. M. (**1987**). "Masking produced by spectral uncertainty with multicomponent maskers," Percept. Psychophys. **41**, 409–415.

Patel, M., and Morse, R. P. (**2010**). (private communication).

R Core Team. (**2019**). "R: A language and environment for statistical computing [software package]," The R Foundation, Vienna, Austria, http://www.R-project.org/ (Last viewed 31 July 2019).

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

Remez, R. E., 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.

Rights, J. D., and Sterba, S. K. (**2019**). "Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures," Psychol. Meth. **24**, 309–338.

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., and Summers, R. J. (**2018**). "Informational masking of speech by time-varying competitors: Effects of frequency region and number of interfering formants," J. Acoust. Soc. Am. **143**, 891–900.

Roberts, B., and Summers, R. J. (**2019**). "Dichotic integration of acoustic-phonetic information: Competition from extraneous formants increases the effect of second-formant attenuation on intelligibility," J. Acoust. Soc. Am. **145**, 1230–1240.

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 Cogn. Sci. **12**, 182–186.

Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods*, 6th ed. (Iowa University Press, Ames, IA).

Stachurski, M., Summers, R. J., and Roberts, B. (**2015**). "The verbal transformation effect and the perceptual organization of speech: Influence of formant transitions and F0-contour continuity," Hear. Res. **323**, 22–31.

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

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.

Summers, R. J., and Roberts, B. (**2020**). "Informational masking of speech by acoustically similar intelligible and unintelligible interferers," J. Acoust. Soc. Am. **147**, 1113–1125.

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.

van Noorden, L. P. A. S. (**1975**). "Temporal coherence in the perception of tone sequences," Doctoral thesis, Eindhoven University of Technology, Eindhoven, the Netherlands.

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.