Across-formant integration and speech intelligibility: Effects of acoustic source properties in the presence and absence of a contralateral interferer

Robert J. Summers,[1] Peter J. Bailey,[2] and Brian Roberts[1,a]
[1] *Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, United Kingdom*
[2] *Department of Psychology, University of York, Heslington, York YO10 5DD, United Kingdom*

(Received 11 April 2016; revised 15 July 2016; accepted 15 July 2016; published online 19 August 2016)

The role of source properties in across-formant integration was explored using three-formant (F1+F2+F3) analogues of natural sentences (targets). In experiment 1, F1+F3 were harmonic analogues (H1+H3) generated using a monotonous buzz source and second-order resonators; in experiment 2, F1+F3 were tonal analogues (T1+T3). F2 could take either form (H2 or T2). Target formants were always presented monaurally; the receiving ear was assigned randomly on each trial. In some conditions, only the target was present; in others, a competitor for F2 (F2C) was presented contralaterally. Buzz-excited or tonal competitors were created using the time-reversed frequency and amplitude contours of F2. Listeners must reject F2C to optimize keyword recognition. Whether or not a competitor was present, there was no effect of source mismatch between F1+F3 and F2. The impact of adding F2C was modest when it was tonal but large when it was harmonic, irrespective of whether F2C matched F1+F3. This pattern was maintained when harmonic and tonal counterparts were loudness-matched (experiment 3). Source type and competition, rather than acoustic similarity, governed the phonetic contribution of a formant. Contrary to earlier research using dichotic targets, requiring across-ear integration to optimize intelligibility, H2C was an equally effective informational masker for H2 as for T2.

[http://dx.doi.org/10.1121/1.4960595]

[VB] Pages: 1227–1238

I. INTRODUCTION

The frequencies of the first three formants and their patterns of change over time are a critical source of information for identifying the phonetic segments being articulated by a talker and hence for understanding speech (see, e.g., Roberts *et al.*, 2011). Precisely how the information carried by different formants is integrated across frequency into a phonetic percept is not fully understood, especially in contexts where more than one talker is speaking at once (see, e.g., Darwin, 2008). In such circumstances, successful communication depends on the extent to which the formant ensemble reaching the ears can be separated into a figure (target) and background (interferer). In principle, any primitive grouping cues facilitating the perceptual segregation of target and masker may lessen speech-on-speech interference (Bregman, 1990). However, while it has long been known that acoustic cues such as differences in onset time and fundamental frequency (F0) can influence the ability to group and segregate formants, these influences can be complex and context-dependent. For example, if the F0 of one formant is different from that of the others, then that formant is usually heard as coming from a different source, but nonetheless may still contribute to the perceived identity of the speech sounds (Cutting, 1976). Typically, imposing a difference in F0 on one formant in an ensemble reduces its phonetic contribution to the speech percept only in circumstances where there is competition between alternative candidates for one or more of the lower formants (Darwin, 1981; Gardner *et al.*, 1989; Summers *et al.*, 2010).

Roberts *et al.* (2015) have recently investigated the effects of more radical differences in acoustic source characteristics between individual formants in an ensemble. They used sentence-length speech analogues and the second-formant competitor (F2C) paradigm (e.g., Remez *et al.*, 1994; Roberts *et al.*, 2010). The crux of the F2C paradigm is the dichotic presentation of two versions of F2; intelligibility is enhanced by the phonetic integration of one version (target F2) with the other formants (F1+F3) but impaired by the integration of the other (F2C). Hence, the listener must reject the competitor to optimize recognition of the utterance. The use of dichotic presentation allows competition between the two candidates for F2 in a context where any interference must arise primarily through informational masking (see, e.g., Durlach *et al.*, 2003; Kidd *et al.*, 2008). Although there are circumstances in which informational masking of speech arising from target-masker confusions can be small (Westermann and Buchholz, 2015), these effects are often substantial (e.g., Brungart *et al.*, 2006). In the version of the F2C paradigm used by Roberts *et al.* (2015), the target F2

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

was presented in the ear opposite F1+F3 and repeat listening was permitted.

Roberts *et al.* (2015) used stimuli with an F1+F3 'frame' that was either buzz-excited (harmonic, H) or sine-wave (tonal, T); each of the target F2 and F2C could take either form and F2 and F2C were matched for root mean square (RMS) power. The properties of F2C were derived from those of the target F2 by time-reversing its frequency and amplitude contours; this manipulation preserves the rate and depth of frequency and amplitude variation found in F2 but changes its pattern. Without F2C, intelligibility was little affected by whether or not the source properties of the target F2 matched those of the F1+F3 frame. This outcome is consistent with earlier findings that across-formant differences in F0 typically have little or no impact on speech recognition in the absence of competition. In contrast, the impact on intelligibility was often substantial when the target F2 was accompanied by F2C in the opposite ear. Most notably, intelligibility was always lowest when F2C was harmonic and F2 was tonal, regardless of the acoustic source properties of F1+F3.

The outcomes of the study by Roberts *et al.* (2015) are interesting for two reasons. First, the effects of target-masker similarity on informational masking for non-speech stimuli (Neff, 1995; Lee and Richards, 2011) suggest that across-formant integration should be facilitated when the source properties of the formants match but hindered when there is a mismatch. Clearly, the dominance of the harmonic candidate for F2 over the tonal candidate irrespective of the source properties of the other formants seems incompatible with a major role for target-masker similarity in determining across-formant grouping. Rather, the results suggest that the type of acoustic source is more important than acoustic similarity between formants in governing intelligibility. Second, the results add to a growing body of evidence from studies and simulations of combined acoustic and electro-acoustic hearing that phonetic information can be integrated across radically different modes of stimulation (e.g., Turner *et al.*, 2004; Qin and Oxenham, 2006; Verschuur *et al.*, 2013). In principle, useful information about formant-frequency variation could be carried by a wide variety of source characteristics, extending well beyond those that might plausibly be produced by a human talker. However, the results of Roberts *et al.* (2015) suggest that the integration of phonetic information across different modes of stimulation may be greatly affected by the presence of interferers, even in circumstances where masking is primarily informational. Hence, these findings may have implications for enhancing mixed-mode listening in clinical contexts.

Before accepting as a general conclusion the notion that the integration of phonetic information across formants is governed by the type of acoustic source, rather than acoustic similarity between formants, two aspects of the experimental design used by Roberts *et al.* (2015) merit further investigation. First, the version of the F2C paradigm used (left ear = F1 ± F2C+F3; right ear = F2) involved dichotic targets and so optimum intelligibility required integration of the target formants across ears, as well as across frequency. Since that study, Roberts and Summers (2015) developed a version of the F2C paradigm that involves presenting all the target formants in the same ear (i.e., monaural speech) and the extraneous formant in the opposite ear. This arrangement completely eliminates energetic masking of the target formants by the interferer (and vice versa); it also avoids the need for listeners to integrate information across ears. The adapted version also uses a single stimulus presentation on each trial, with random allocation of the target speech to the left or right ear. The lack of opportunity for repeat listening further increases the ecological validity of the approach and the uncertainty from trial to trial about the lateralization of the target speech discourages listeners from attending selectively to one ear, increasing the extent of informational masking (see Kidd *et al.*, 2008). Second, harmonic formants are wideband and so are louder than their tonal counterparts when they are matched for RMS power. Hence, a possible alternative account of the findings reported by Roberts *et al.* (2015) is that, when in competition, the louder candidate formant dominates in contributing phonetic information to the speech percept. The three experiments reported here address these issues.

II. EXPERIMENT 1

F1 and F3 were always generated by passing a monotonous periodic source through second-order resonators. The target F2 was either generated in the same way or was a sine-wave (tonal) analogue. When present, the extraneous competitor (F2C) received in the ear contralateral to the target could take either form. This approach allowed exploration of the effects of matches and mismatches in acoustic source characteristics within formant ensembles, both in target-only and target-plus-interferer listening contexts. Although there are inherent differences in the amount of phonetic information carried by harmonic and tonal analogues of formants with the same frequency and amplitude contours, as evidenced by the lower intelligibility of sine-wave speech than of otherwise comparable harmonic analogues (Bailey *et al.*, 1977; Remez *et al.*, 1981), these differences can be controlled by making an appropriate choice of comparisons across conditions.

A. Method

1. Listeners

Listeners 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. All of them met the additional criterion of a mean score of ≥20% keywords correct in the main experiment, when collapsed across all conditions, and so their results were included in the final dataset. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. Twenty-four listeners (four males) successfully completed the experiment (mean age = 19.6 years, range = 18.2–34.8). To our knowledge, none of the listeners had heard any of the sentences used in the main part of the experiment in any previous study or assessment. All listeners were native speakers of English and gave informed consent. The research was approved by the Aston University Ethics Committee.

2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of the Bamford-Kowal-Bench (BKB) sentence lists (Bench *et al.*, 1979), spoken by a British male talker of "Received Pronunciation" English. To enhance the intelligibility of the speech analogues, 48 semantically simple sentences were used; these sentences 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. To facilitate comparison, the sentences and designated keywords were identical to those used in the corresponding experiment reported by Roberts *et al.* (2015), which used dichotic targets. The stimuli for the training session (see Sec. II A 3) 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 also contained ≤25% phonemes involving closures or unvoiced frication.

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

In all conditions, the frequency and amplitude contours of F1 and F3 were used to control two parallel buzz-excited second-order resonators. Hence, F1+F3 provided a common "harmonic frame" shared by all conditions. The excitation source was a monotonous periodic train of pulses (F0 = 140 Hz) modeled on the glottal waveform (Rosenberg, 1971). In the all-harmonic target conditions (H1+H2+H3), the frequency and amplitude contours of F2 were used to control a third parallel buzz-excited resonator receiving the same excitation source. 3-dB bandwidths of the resonators corresponding to F1, F2, and F3 were set to constant values of 50, 70, and 90 Hz, respectively. 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 hybrid-target conditions (H1+T2+H3), the frequency and amplitude contours of F2 were used to control the properties of a time-varying sinusoid. This tonal analogue of F2 (T2) was matched to the RMS power of its harmonic counterpart (H2) before being combined with the harmonic F1+F3 frame.

For each sentence in the main experiment, second-formant competitors (F2Cs) were generated using the time-reversed frequency and amplitude contours of the corresponding target F2; as noted above, this manipulation preserves the rate and depth of frequency and amplitude variation found in the F2 contour. These competitors were rendered either as the output of a buzz-excited resonator (H2C) or as an RMS-matched time-varying sinusoid (T2C). In the former case, the excitation source (Rosenberg pulses), F0 frequency (140 Hz), 3-dB bandwidth (70 Hz), and output sign (–) were identical to those used to synthesize the target F2. The waveform of the excitation source for F2C was not time reversed. When present, F2C was always delivered in the ear opposite to that receiving the monaural target.

There were eight conditions in the main experiment (see Table I). C1 and C2 were the F2-absent conditions. The stimuli for C1 comprised the F1+F3 frame alone; C2 differed only in that F2C (harmonic version) was present in the contralateral ear. The stimuli for C3–C6 comprised all three target formants plus the contralateral competitor. This set represents all four combinations of acoustic properties for F2 and F2C; the study by Roberts *et al.* (2015) did not include the case where both F2 and F2C were tonal analogues, for which neither matched the harmonic F1+F3 frame. The stimuli for the remaining conditions (C7–C8) comprised only the target formants. C7 constitutes the all-harmonic reference case. Figure 1 illustrates both versions of the three-formant monaural target (left panels: all-harmonic = top, hybrid = bottom) and both versions of the competitor (right panels: harmonic = top, tonal = bottom). The 48 sentences were divided equally across conditions (i.e., six per condition), such that there were always 18 or 19 keywords per condition. Allocation of sentences to conditions was counterbalanced by rotation across each set of eight listeners tested. Hence, the total number of listeners required to produce a balanced dataset was a multiple of eight.

TABLE I. Stimulus properties for the conditions used in experiment 1 (main session). H and T denote harmonic and tonal formant analogues, respectively. The F1+F3 frame was harmonic in all conditions. Instances where F2 and/or F2C were rendered using different source characteristics from the frame are shown in bold. The F0 frequency of the Rosenberg source for the harmonic analogues of F1, F2, F3, and F2C was always 140 Hz.

Condition	Stimulus configuration (target ear)	Stimulus configuration (other ear)
C1	H1+H3	—
C2	H1+H3	H2C
C3	H1+H2+H3	H2C
C4	H1+H2+H3	**T2C**
C5	H1+**T2**+H3	H2C
C6	H1+**T2**+H3	**T2C**
C7	H1+H2+H3	—
C8	H1+**T2**+H3	—

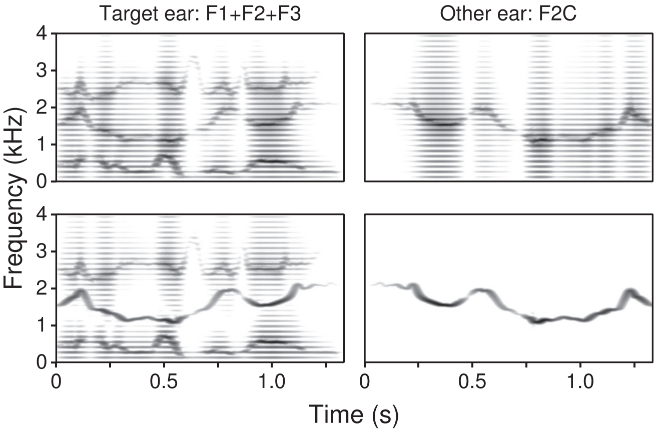

FIG. 1. Stimuli for experiment 1—narrowband spectrograms of the example sentence "They had a lovely day" (left panels) accompanied by a competitor for F2 (F2C) in the contralateral ear (right panels). The formants constituting the F1+F3 frame were rendered as harmonic analogues in all conditions; the F0 frequency of the harmonic analogues was 140 Hz. F2 and F2C were rendered either as harmonic (H) or tonal (T) analogues. On the left, the upper and lower panels illustrate the all-harmonic target (H1+H2+H3) and hybrid target (H1+T2+H3) cases, respectively. On the right, the upper and lower panels illustrate the harmonic (H2C) and tonal (T2C) competitor cases, respectively.

3. Procedure

During testing, listeners were seated in front of a computer screen and a keyboard in a sound-attenuating chamber (Industrial Acoustics 1201A; Winchester, UK). The experiment consisted of a training session followed by the main session and took about 45–60 min to complete; listeners were free to take a break whenever they wished. In both parts of the experiment, stimuli were presented in a new quasi-random order for each listener.

The training session comprised 50 trials; stimuli were presented without competitors and a new sentence was used for each trial. Half of the sentences were rendered as all-harmonic analogues (H1+H2+H3); the others differed only in that F2 was instead rendered as a sine-wave analogue (H1+T2+H3). On each of the first ten trials, listeners heard diotic presentations of the synthetic version (S) and the original recording (clear, C) of a given 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 new 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 analogues.

During the final ten training trials, sentence analogues were delivered monaurally; the target ear was selected randomly on each trial. Listeners heard the stimulus only once before entering their transcription. Feedback was provided by playing the original recording in the selected ear. Following the procedure of Roberts and Summers (2015), listeners continued on to the main session if they met either or both of two criteria: (1) ≥50% keywords correct across all 40 trials needing a transcription (30 trials = diotic with repeat listening; 10 trials = monaural, random selection of ear, no repeat listening); (2) ≥50% keywords correct for the final 15 diotic-with-repeat-listening trials. On each trial in the main experiment, the ear receiving the target formants (F1+F2+F3 or F1+F3) was selected randomly; F2C (when present) was always presented in the opposite 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 external sound card (Creative Technology, model SB1240; Singapore), programmable attenuators (Tucker-Davis Technologies PA5; Alachua, FL), and a headphone buffer (TDT HB7). Output levels were calibrated using a sound-level meter (Brüel and Kjaer, type 2209; Nærum, Denmark) coupled to the earphones by an artificial ear (type 4153). All target sentences were presented at a long-term average of 75 dB sound pressure level (SPL); there was some variation in the sound level at the ear receiving F2C (mean ≈ 65 dB SPL), depending on the RMS power of the target F2. In the training session, the presentation level of the diotic materials (first 40 target sentences plus original recordings) was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation. The last ten sentences in the training session were presented monaurally at the reference level.

4. Data analysis

For each listener, the intelligibility of each stimulus was quantified in terms of the percentage of keywords identified correctly; homonyms were accepted. The stimuli for each condition comprised six sentences. Given the variable number of keywords per sentence (3 or 4), the mean score for each listener in each condition was computed as the percentage of keywords reported correctly giving equal weight to all the keywords used. As in our previous studies (Roberts *et al.*, 2010, 2014, 2015; Roberts and Summers, 2015; Summers *et al.*, 2010, 2012), we classified responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly. All statistical analyses reported here were computed using SPSS (SPSS statistics version 20, IBM Corp.). The measure of effect size reported here is partial eta squared (η_p^2).

B. Results

Figure 2 shows the mean percentage scores (and intersubject standard errors) across conditions for keyword identification. The black, white, and gray bars indicate the results for the frame ± F2C, all-harmonic target, and hybrid-target conditions, respectively. A one-way within-subjects analysis of variance (ANOVA) showed a highly significant effect of condition on intelligibility [$F(7,161) = 41.45$, $p < 0.001$, $\eta_p^2 = 0.64$]. Paired-samples comparisons (two-tailed) were computed using the restricted least-significant-difference test

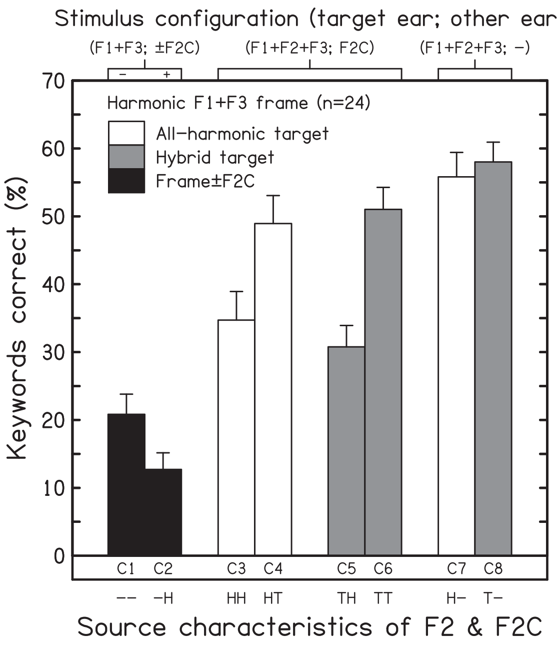

FIG. 2. Results for experiment 1—effects of source characteristics and competitors (F2Cs) on the intelligibility of sentence analogues when the F1+F3 frame was harmonic. Mean scores and inter-subject standard errors ($n = 24$) are shown for the F2-absent conditions (black bars), the conditions for which all target formants were harmonic analogues (matched, white bars), and the conditions for which the target speech was a mixed-source hybrid comprising the harmonic F1+F3 frame and a tonal analogue of F2 (mismatched, gray bars). The top axis indicates which formants were presented to each ear; the bottom axis indicates the source characteristics of F2 and F2C—harmonic (H) or tonal (T). For ease of reference, condition numbers are included immediately above the bottom axis.

(Snedecor and Cochran, 1967). The scores for the frame±F2C conditions (C1 and C2) differed from those for all other conditions, including each other (range: $p = 0.021 - p < 0.001$). The difference between the mean scores for C1 and C2 indicates that the addition of F2C tended to reduce further the limited intelligibility supported by F1+F3 alone, indicating the integration of misleading phonetic information carried by F2C. Performance was best when the three target formants were presented without competitors. Notably, the addition of the target F2 to the F1+F3 frame substantially increased intelligibility, irrespective of whether F2 matched the source properties of the frame (C1 vs C7: mean difference = 35.0 percentage points) or not (C1 vs C8: mean difference = 37.2 percentage points). This outcome indicates the effective integration of phonetic information across the target formants even when they had different source characteristics.

The effects of differences in the acoustic form of the target speech (all-harmonic or hybrid) and of competitor presence and form (F2C status: harmonic, tonal, or absent) were explored using a two-way ANOVA restricted to the experimental conditions (C3–C8). This analysis revealed a highly significant main effect of F2C status [$F(2,46) = 38.19$, $p < 0.001$, $\eta_p^2 = 0.62$], but there was no effect of whether the target speech was all-harmonic or hybrid [$F(1,23) = 0.005$, $p = 0.94$] and the two factors did not interact [$F(2,46) = 0.996$, $p = 0.38$]. All pairwise comparisons within the F2C status factor were significant (harmonic vs absent, $p < 0.001$; harmonic vs tonal, $p < 0.001$; tonal vs absent, $p = 0.04$). Adding to the target formants an F2C created using the time-reversed frequency and amplitude contours of F2 typically reduced intelligibility, but the extent of competitor impact depended on the source characteristics of F2C. Notably, the intelligibility cost of adding a harmonic F2C to either form of target speech was substantial (mean = 24.2 percentage points), whereas the cost of adding a tonal F2C was relatively modest (mean = 6.8 percentage points).

C. Discussion

There are two main findings from this experiment. First, in the absence of a competitor, there was no intelligibility cost of a mismatch in source properties between the F1+F3 frame and the target F2 when they were presented in the same ear. This indicates that the phonetic information carried by the formants comprising a hybrid target must be integrated effectively, despite their acoustic dissimilarities. Clearly, a sine-wave analogue of F2 is capable of conveying phonetic information in a way that can be combined with that carried by the buzz-excited formants. This outcome confirms the interpretation of one of our earlier studies, using similar monaural targets, for which the intelligibility cost of a mismatch across formants in source properties in the absence of F2C was small (~5 percentage points; Roberts *et al.*, 2015). A limitation of that study was that the main experiment did not include a frame-only condition and so it was only possible to estimate the intelligibility gain from adding the target F2 with reference to pilot data collected using F1+F3 stimuli. Without clear evidence that the target F2 made a substantial contribution to overall intelligibility, it is hard to interpret the small intelligibility cost of the across-formant mismatch in acoustic form. As it turns out, the estimates provided by Roberts *et al.* (2015) of the mean score for the frame-only case (~20%), and of the increase in keyword scores when the target F2 is added (~35 percentage points), are both similar to the values observed here.

Adding an F2C to the target formants typically reduced intelligibility, presumably through informational masking; this result is consistent with evidence from previous studies that listeners are often unable to ignore contralateral interferers (e.g., Brungart *et al.*, 2005; Gallun *et al.*, 2007; Roberts and Summers, 2015). The other main finding of this experiment is that the extent of competitor impact depended on the source characteristics of F2C. Specifically, adding the harmonic version of F2C in the contralateral ear had considerably more impact on intelligibility than adding the tonal version, but the greater impact of H2C did not depend on whether the target F2 was harmonic (matched to F1+F3 frame) or tonal (mismatched). The first of these outcomes is in accord with our earlier findings for otherwise comparable dichotic-target stimuli, but the absence of an interaction between F2C source properties and whether the target F2 matches the frame is not. The results for the dichotic targets were rather different; in that case, the intelligibility cost of competition from H2C was much greater for T2 than for H2. In terms of changes in keyword scores, the effect of adding

H2C was ~2.1 times greater for hybrid than for all-harmonic dichotic targets (Roberts *et al.*, 2015).

III. EXPERIMENT 2

F1 and F3 were always rendered as sine-wave analogues in this experiment; these stimuli have lower baseline intelligibility than their harmonic counterparts. The target F2 was either generated in the same way or by passing a monotonous periodic source through a second-order resonator. As for experiment 1, F2C (when present) could take either form. The results of experiment 1 indicate a tendency for a tonal analogue to be less effective when competing with a harmonic analogue, irrespective of which one corresponded to the target F2. They also indicate that, unlike their dichotic counterparts, there is no evidence of an interaction between the effects of the source properties of F2 and F2C for monaural targets. The change to a tonal F1+F3 frame allowed an assessment of whether these outcomes generalize to cases where at least two of the target formants are tonal. Note also that if the advantage of harmonic analogues found in experiment 1 were a consequence of grouping by similarity in source characteristics, one would expect the harmonic analogue to be less effective in the context of a tonal F1+F3 frame. This is not what happens for stimulus configurations involving dichotic presentation of the target formants (Roberts *et al.*, 2015).

A. Method

Except where described, the same method was used as for experiment 1. Twenty-four listeners (nine males) passed the training and successfully completed the experiment (mean age = 25.3 years, range = 18.9–48.8); this includes replacements for listeners who did not meet the additional criterion of an overall mean score of ≥20% keywords correct in the main session. All listeners had already successfully completed at least one speech perception experiment in our laboratory, but none using stimuli derived from the sentences used in the main session. The stimuli for the training session differed from those used in experiment 1 only in that half of the sentences were rendered as sine-wave speech (T1+T2+T3) and for the rest F2 was instead rendered as the output of a buzz-excited resonator (i.e., T1+H2+T3).

The stimuli for the main experiment were derived from a set of 48 BKB sentences, again allocated such that there were always 19 or 20 keywords per condition. These sentences did not overlap with those used in experiment 1, but the sentences and designated keywords were identical to those used in the corresponding experiment reported by Roberts *et al.* (2015). The harmonic frame shared by all conditions in experiment 1 was replaced here with a tonal frame (T1+T3), raised to match the RMS power of its harmonic counterpart. As before, F2 and F2C could be rendered either as the output of a buzz-excited ($F0 = 140$ Hz) resonator (H2, H2C) or as an RMS-matched time-varying sinusoid (T2, T2C). To be consistent with experiment 1, the sign of the outputs of the resonators corresponding to H2 and H2C was inverted (–). There were eight conditions in the main session (see Table II), arranged in an analogous pattern to that used in experiment 1; C7 constitutes the all-tonal reference case. Figure 3 illustrates both versions of the three-formant monaural target (left panels: all-tonal = top, hybrid = bottom) and both versions of the competitor (right panels: tonal = top, harmonic = bottom).

TABLE II. Stimulus properties for the conditions used in experiment 2 (main session). T and H denote tonal and harmonic formant analogues, respectively. The F1+F3 frame was tonal in all conditions. Instances where F2 and/or F2C were rendered using different source characteristics from the frame are shown in bold. The F0 frequency of the Rosenberg source for the harmonic analogues of F2 and F2C was always 140 Hz.

Condition	Stimulus configuration (target ear)	Stimulus configuration (other ear)
C1	T1+T3	—
C2	T1+T3	T2C
C3	T1+T2+T3	T2C
C4	T1+T2+T3	**H2C**
C5	T1+**H2**+T3	T2C
C6	T1+**H2**+T3	**H2C**
C7	T1+T2+T3	—
C8	T1+**H2**+T3	—

B. Results

Figure 4 shows the mean percentage scores (and intersubject standard errors) across conditions for keyword identification. The black, white, and gray bars indicate the results for the frame±F2C, all-tonal target, and hybrid-target conditions, respectively. A one-way ANOVA showed a highly significant effect of condition on intelligibility [$F(7,161) = 33.97$, $p < 0.001$, $\eta_p^2 = 0.60$]. Intelligibility was near floor for T1+T3 alone and was not reduced further by the addition of T2C (C1 vs C2, $p = 0.85$). Pairwise comparisons showed that the scores for T1+T3 alone differed from those for all experimental conditions (range: $p = 0.005 - p < 0.001$) other

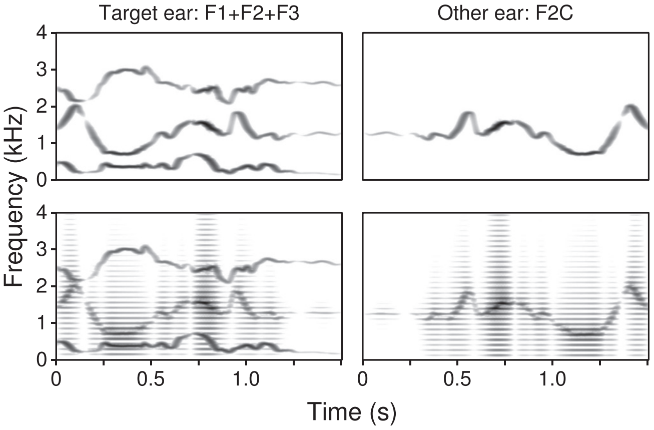

FIG. 3. Stimuli for experiment 2—narrowband spectrograms of the example sentence "They called an ambulance" (left panels) accompanied by a competitor for F2 (F2C) in the contralateral ear (right panels). The formants constituting the F1+F3 frame were rendered as tonal (sine-wave) analogues in all conditions. F2 and F2C were rendered either as harmonic (H) or tonal (T) analogues; the F0 frequency of the harmonic analogues was 140 Hz. On the left, the upper and lower panels illustrate the all-tonal target (T1+T2+T3) and hybrid target (T1+H2+T3) cases, respectively. On the right, the upper and lower panels illustrate the tonal (T2C) and harmonic (H2C) competitor cases, respectively.

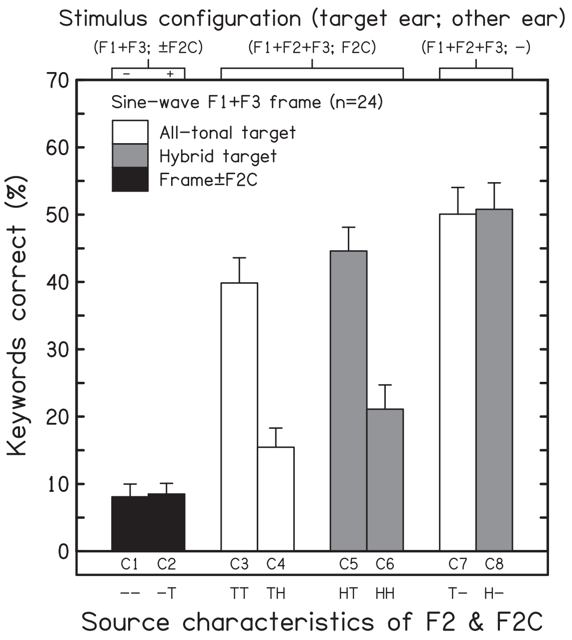

FIG. 4. Results for experiment 2—effects of source characteristics and competitors (F2Cs) on the intelligibility of sentence analogues when the F1+F3 frame was tonal. Mean scores and inter-subject standard errors (n = 24) are shown for the F2-absent conditions (black bars), the conditions for which all target formants were tonal analogues (matched, white bars), and the conditions for which the target speech was a mixed-source hybrid comprising the tonal F1+F3 frame and a harmonic analogue of F2 (mismatched, gray bars). The top axis indicates which formants were presented to each ear; the bottom axis indicates the source characteristics of F2 and F2C—tonal (T) or harmonic (H). For ease of reference, condition numbers are included immediately above the bottom axis.

than C4 ($p = 0.064$). As for experiment 1, performance was best when the target formants were presented without competitors. Once again, adding the target F2 to the F1+F3 frame substantially increased intelligibility, irrespective of whether F2 matched the source properties of the frame (C1 vs C7: mean difference = 42.0 percentage points) or not (C1 vs C8: mean difference = 42.7 percentage points). This outcome provides further evidence that phonetic information was integrated effectively across the target formants, even when they had different source characteristics.

The effects of differences in the acoustic form of the target speech (all-tonal or hybrid) and of the presence and form of F2C (tonal, harmonic, or absent) were explored using a two-way ANOVA restricted to the experimental conditions (C3–C8). This analysis revealed a similar pattern to that observed in experiment 1. There was a highly significant main effect of F2C status [$F(2,46) = 36.28$, $p < 0.001$, $\eta_p^2 = 0.61$], but there was no effect of whether the target speech was all-tonal or hybrid [$F(1,23) = 1.94$, $p = 0.18$] and the two factors did not interact [$F(2,46) = 0.357$, $p = 0.70$]. As for experiment 1, the intelligibility cost of adding a harmonic F2C to either form of target speech was substantial (mean = 32.2 percentage points), whereas the cost of adding a tonal F2C was relatively modest (mean = 8.2 percentage points). Indeed, pairwise comparisons within the F2C status factor indicated that adding H2C to the target formants significantly reduced intelligibility ($p < 0.001$), but that here the impact of T2C did not quite reach significance ($p = 0.068$); the difference in impact between the two types of competitor was significant ($p < 0.001$). This outcome is notable because of the mismatch in source characteristics between H2C and the T1+T3 frame.

C. Discussion

The overall intelligibility of the materials used here was lower than for experiment 1, as would be expected given the higher proportion of target formants presented here as sine-wave analogues. Nonetheless, the main findings are in accord with those of experiment 1. First, keyword scores increased from <10% for T1+T3 alone to ~50% irrespective of whether T2 or H2 was added to the frame. This suggests that, in the absence of competition, listeners fully integrate the phonetic information carried by the harmonic analogue of F2 with that carried by the tonal F1+F3 frame, despite the mismatch in acoustic form. Second, the impact of the competitor was greater when it was rendered as a buzz-excited formant, not when it matched the source characteristics of the tonal F1+F3 frame, which suggests that acoustic similarity does not play a major role in formant grouping and segregation. Furthermore, the greater impact of H2C was independent of whether the target F2 was tonal (matched to F1+F3 frame) or harmonic (mismatched). This outcome contrasts with that for dichotic targets, for which the impact on keyword scores of adding H2C was ~2.6 times greater for all-tonal than for hybrid targets (Roberts *et al.*, 2015).

Taken together, experiments 1 and 2 support the main conclusion of Roberts *et al.* (2015) that acoustic source characteristics, rather than across-formant similarity, govern the phonetic contribution made by a particular formant. However, there is one important difference in outcome between the two studies. Specifically, the finding for dichotic targets that the greater impact of H2C than T2C on intelligibility (irrespective of frame type) is magnified when the target F2 is tonal rather than harmonic does not occur for monaural targets. The basis for this difference in outcome is considered further in Sec. V.

IV. EXPERIMENT 3

This experiment examined whether the tendency for harmonic competitors to have a greater impact on intelligibility than their tonal counterparts, irrespective of which one shares common source characteristics with the F1+F3 frame, is maintained when H2C and T2C are set to the same loudness as T2 and H2, respectively, rather than to the same RMS power. This merits checking because wideband harmonic analogues are heard as louder than narrowband tonal analogues of equal intensity, which in principle might boost their relative effectiveness. All target sentences in this experiment were rendered as hybrid stimuli—i.e., the source characteristics of the F1+F3 frame and the target F2 were different. As for the equivalent conditions in experiments 1 and 2, tonal analogues of the target F2 and F1+F3 frame were set to the same RMS power as the corresponding harmonic versions.

When the alternative versions of the second formant shared the same source properties, F2C was set to the same RMS power as F2, as in all our previous experiments. When

F2C and F2 had different source properties, the level adjustments required for loudness matching were made to the competitors. Relative to matching F2C and F2 for equal RMS power, H2C level was *lowered* to match the loudness of T2 for the harmonic-frame cases, and T2C level was *raised* to match the loudness of H2 for the tonal-frame cases. Since the competitor was presented contralaterally to the target and F1+F3 frame, any changes in competitor impact resulting from these adjustments could not occur through changes in energetic masking. Given the nature of our stimuli, the magnitudes of the level adjustments required were computed using a model of loudness applicable to time-varying sounds known as the TVL model (Glasberg and Moore, 2002). This model is well established and has been evaluated by other researchers (see, e.g., Rennies *et al.*, 2010; Zorila *et al.*, 2016). The TVL model was also used to compare the loudness of RMS-matched targets and competitors with the same source properties; it was anticipated that these stimuli would be similar in loudness.

TABLE III. Stimulus properties for the conditions used in experiment 3 (main session). H and T denote harmonic and tonal formant analogues, respectively. The F1+F3 frame was either harmonic (C1-C4) or tonal (C5-C8). Instances where F2 and/or F2C were rendered using different source characteristics from the F1+F3 frame are shown in bold. Note that all cases involving all three target formants used hybrid stimuli (i.e., source mismatch between F1+F3 frame and target F2). Downward- and upward-pointing arrows indicate the direction of level adjustment (relative to equal RMS power) required to match the loudness of F2C to the target F2, according to the loudness model for time-varying stimuli provided by Glasberg and Moore (2002).

Condition	Stimulus configuration (target ear)	Stimulus configuration (other ear)
C1	H1+H3	—
C2	H1+**T2**+H3	↓H2C↓
C3	H1+**T2**+H3	**T2C**
C4	H1+**T2**+H3	—
C5	T1+T3	—
C6	T1+**H2**+T3	**H2C**
C7	T1+**H2**+T3	↑T2C↑
C8	T1+**H2**+T3	—

A. Method

Except where described, the same method was used as for experiments 1 and 2. Sixteen listeners (three males) passed the training and successfully completed the experiment (mean age = 20.8 years, range = 18.2–37.9); no replacements were required based on the additional criterion of an overall mean score of ≥20% keywords correct in the main session. All listeners had experience of previous speech perception experiments in our laboratory, but none involving stimuli derived from the sentences used in the main session. The 48 BKB sentences used comprised two sets; the first set corresponded to the 24 most intelligible sentences from the all-harmonic reference condition in experiment 1, and the second to the 24 most intelligible sentences from the all-tonal reference condition in experiment 2. All stimuli used in the main experiment were synthesized and played back at a higher sample rate (32 kHz), owing to the requirements of the software used to compute estimates of their loudness (Glasberg and Moore, 2002).

There were eight conditions in the main session (see Table III). The stimuli for C1–C4 (harmonic frame) and C5–C8 (tonal frame) were derived from the first and second sets of sentences, respectively. The stimuli for C1 and C5 correspond to the frame-only control cases. The stimuli for C2–C4 allow comparison of the impact of H2C and T2C on the intelligibility of hybrid targets (H1+T2+H3) when the level of H2C has been lowered from that required to match the RMS power of T2 to that required to match the estimated loudness of T2. The stimuli for C6–C8 allow comparison of the impact of H2C and T2C on the intelligibility of hybrid targets (T1+H2+T3) when the level of T2C has been raised from that required to match the RMS power of H2 to that required to match the estimated loudness of H2. Given that the two sets of sentences were non-overlapping, counterbalancing by rotation only required a multiple of four listeners. Across the two sets, no sentence shared more than one keyword with any other sentence. The training session was analogous to those used for experiments 1 and 2, consisting of an equal number of harmonic-frame and tonal-frame stimuli.

For time-varying signals like speech, listeners can judge the short-term loudness of the stimulus (e.g., the loudness of a particular syllable) or the overall impression of loudness for a relatively long segment (e.g., the long-term loudness of a sentence). In this experiment, using sentence-length materials, our aim was to match the overall loudness of stimuli with different source properties. The TVL model (Glasberg and Moore, 2002) uses the time waveform of the signal as its input and has seven stages. First, a finite impulse response filter simulates signal transfer through the outer and middle ear. Second, the short-term spectrum is computed using the fast Fourier transform (FFT); to obtain sufficient spectral resolution at low frequencies and temporal resolution at high frequencies, longer and shorter signal segments are used for low and high frequencies, respectively. Third, an excitation pattern is computed from the physical spectrum. Fourth, the excitation pattern is transformed into a specific loudness pattern. Fifth, the area under the specific loudness pattern is taken as the value for the "instantaneous" loudness of the signal. Sixth, the short-term perceived loudness of the signal is computed from the instantaneous loudness using an averaging mechanism similar to an automatic gain control system. Finally, the overall impression of loudness for longer signals is computed from successive short-term loudness estimates using a similar averaging mechanism, but with longer attack and release times.

When F2C and F2 had different source properties, we used the TVL model to adjust the level of the competitors to match the loudness of their target counterparts. To facilitate comparison with F2C, loudness estimates for the target F2 were computed when it was presented in isolation at a level corresponding to that for F2 in the behavioral experiments (for which F2 was accompanied by the F1+F3 frame). To allow sufficient time for the algorithm to stabilize, the overall loudness of each stimulus was computed using short-term loudness estimates $\geq$250 ms from the start of the signal; sentence duration was typically in the range 1–2 s. An iterative process was used to match the loudness of H2C to T2 and T2C to H2. First, the overall loudness of each stimulus was computed based on the mean value of the long-term loudness. This established that on average H2 and H2C had loudness levels $\sim$9 phon above those of their tonal counterparts. Second, assuming approximate equivalence between the phon and dB scales, the levels of the harmonic and tonal competitors were changed for all sentences by -9 and $+9$ dB, respectively, to achieve approximate similarity in loudness for F2C and F2 when they had different source properties. Third, the overall loudness levels of H2C and T2C were computed for each sentence after the coarse adjustment, and individual dB corrections were applied to H2C and T2C, equal to the difference in phon from the corresponding target F2. As a result, final matches in loudness between harmonic and tonal counterparts were close (H2C vs T2: mean difference = 0.03 phon, SD = $\pm$0.02; T2C vs H2: mean difference = -0.12 phon, SD = $\pm$0.08). We also used the TVL model to compare the loudness of each competitor with its RMS-matched target counterpart when the two candidates shared a common source. As anticipated, these stimuli were similar in loudness (H2C vs H2: mean difference = -0.20 phon, SD = $\pm$0.32; T2C vs T2: mean difference = -0.30 phon, SD = $\pm$0.43). Given that loudness matching was used when F2C and F2 had different source properties, and RMS matching was a good surrogate for loudness matching when F2C and F2 had common source properties, corresponding H2C–T2C pairs were also similar in loudness.

B. Results

Figure 5 shows the mean keyword scores (and inter-subject standard errors) for the harmonic-frame (top panel) and tonal-frame (bottom panel) conditions, respectively. The black bars indicate the results for the frame-only conditions; the gray bars indicate the results for the hybrid-target conditions in the presence and absence of loudness-matched harmonic and tonal competitors. The results for the two types of frame were analysed separately using one-way ANOVAs. In each case, the analysis revealed a highly significant effect of condition on intelligibility irrespective of whether the frame-only condition was included (harmonic frame: [$F(3,45) = 27.67$, $p < 0.001$, $\eta_p^2 = 0.65$]; tonal frame: [$F(3,45) = 40.64$, $p < 0.001$, $\eta_p^2 = 0.73$]) or not (harmonic frame: [$F(2,30) = 11.53$, $p < 0.001$, $\eta_p^2 = 0.44$]; tonal frame: [$F(2,30) = 16.08$, $p < 0.001$, $\eta_p^2 = 0.52$]). Pairwise comparisons showed that the scores for the frame-only cases differed from those for all the experimental conditions (harmonic frame: $p = 0.004 - p < 0.001$; tonal frame: $p < 0.001$ in all cases).

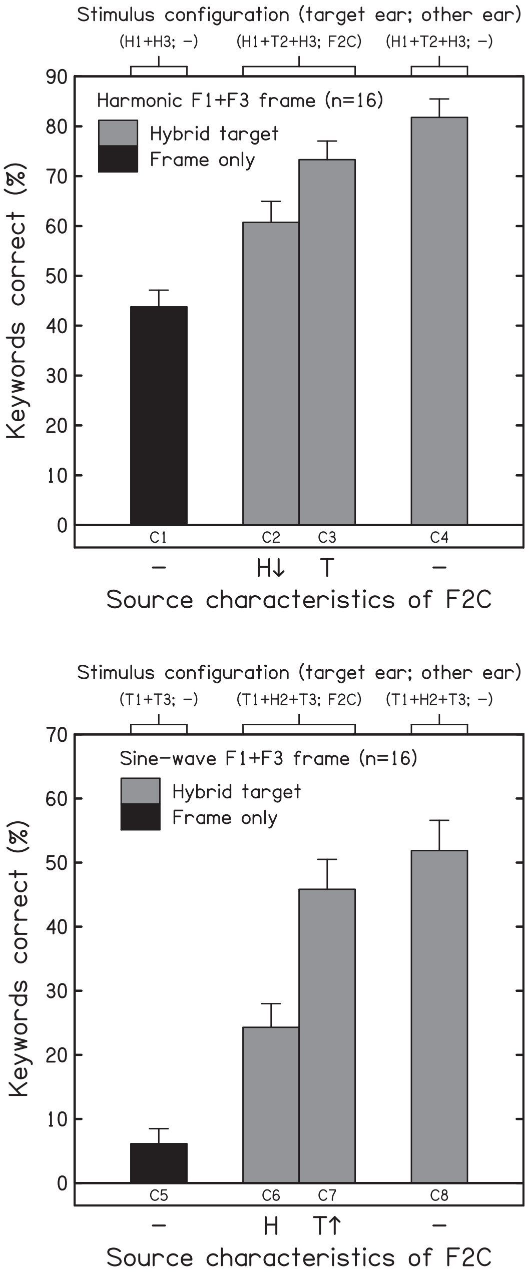

FIG. 5. Results for experiment 3—effects of source characteristics and competitors (F2Cs) on sentence intelligibility when the level of F2C was matched to the loudness of F2, using the model of Glasberg and Moore (2002). The top and bottom panels indicate the results for the conditions where the F1+F3 frame was harmonic and tonal, respectively. In each panel, the mean scores and inter-subject standard errors (n = 16) are shown for the frame-only condition (black bar), and the conditions for which the target speech was a hybrid with different source characteristics for the F1+F3 frame and for F2 (mismatched, gray bars). In each panel, the top axis indicates which formants were presented to each ear and their source characteristics. The bottom axis indicates the source characteristics of F2C—harmonic (H) or tonal (T); an arrow is used to indicate the direction of level adjustment for F2C needed to change from a match with F2 in RMS power to a match in loudness when F2 and F2C have different source properties. For ease of reference, condition numbers are included immediately above the bottom axis.

As for experiments 1 and 2, performance was best when the target formants were presented without competitors. Adding a mismatched target F2 to either type of frame substantially increased intelligibility, indicating the integration of phonetic information across formants despite the difference in source properties between them (C1 vs C4 and C5 vs C8: mean differences = 38.0 and 45.7 percentage points, respectively). Pairwise comparisons were used to assess the effects of adding each of the two types of loudness-matched competitor to each of the two configurations of hybrid-target speech (H1+T2+H3 and T1+H2+T3).

Regardless of whether a harmonic or tonal frame was used, the intelligibility cost of adding H2C was substantial and significant (C2 vs C4: mean difference = 21.1 percentage points, $p < 0.001$; C6 vs C8: mean difference = 27.6 percentage points, $p < 0.001$), whereas the cost of adding a tonal F2C was small and non-significant (C3 vs C4: mean = 8.5 percentage points, $p = 0.091$; C7 vs C8: mean difference = 6.0 percentage points, $p = 0.15$). Furthermore, the difference in impact between the two types of F2C was significant in the context of both frames (C2 vs C3: mean = 12.6 percentage points, $p = 0.021$; C6 vs C7: mean difference = 21.5 percentage points, $p = 0.001$). Even when the two types of F2C were closely matched in loudness to their target counterparts, H2C was a far more effective competitor than T2C, irrespective of whether or not it matched the source characteristics of the F1+F3 frame.

C. Discussion

The results confirm and extend those from experiments 1 and 2. Regardless of the acoustic source properties of the F1+F3 frame or the target F2, harmonic competitors remained significantly more effective than their tonal counterparts when they were similar in loudness. This outcome is striking for two reasons. First, on average, each adjusted H2C was ∼9 dB less intense than its tonal counterpart. Second, informal listening suggests that matching harmonic and tonal analogues using loudness estimates from the TVL model probably over-compensates for the difference in bandwidth between them. Most likely, this is because the model assumes full integration of loudness across channels, whereas human listeners may instead place greater weight on those channels closest to the formant peak.

Overall, we conclude that the difference in competitor effectiveness between H2C and T2C observed by Roberts *et al.* (2015) cannot be explained in terms of the greater loudness of harmonic than tonal analogues when matched for RMS power. Some other aspect of these stimuli, such as differences in naturalness or bandwidth *per se*, must be the critical factor. In particular, as noted by Roberts *et al.* (2015), widening the bandwidth of formant analogues can support higher intelligibility (e.g., Lewis and Carrell, 2007; Souza and Rosen, 2009), presumably because the spread of excitation across a greater number of channels makes these stimuli more effective at carrying phonetic information.

Another aspect of the results for experiment 3 that merits comment is the substantial difference in keyword scores between the H1+H3 case (C1, 43.5%) and the T1+T3 case (C5, 6.2%). The difference between the corresponding cases across experiments 1 and 2 was much smaller, albeit with the caveat that different listeners took part in the two experiments. Although not conclusive, it seems likely that this discrepancy is a consequence of selecting the most intelligible of the sentences used in experiments 1 and 2, given the greater overall spread in intelligibility for the harmonic targets. Indeed, for the tonal-frame conditions in experiment 3 relative to their counterparts in experiment 2, there was only a modest rise in scores for the T1+H2+T3 case and none at all for the tonal-frame case. Despite the high baseline performance for the harmonic-frame case in experiment 3, adding the (mismatched) target F2 nonetheless improved intelligibility considerably (from >40% to >80%), providing evidence of the integration of phonetic information across the target formants in the context of a harmonic (as well as a tonal) frame. Once again, it is clear that intelligible analogues of sentence-length utterances can be created by combining harmonic and tonal renditions of different target formants.

V. GENERAL DISCUSSION

The experiments reported here have explored the effects of source properties *per se*, and of differences in acoustic form between formants, on the integration of phonetic information across formants when listening to single presentations of monaural targets with unpredictable lateralization. To explore how competition modulates the effects of differences in source properties, these effects were compared in the presence and absence of single-formant interferers in the contralateral ear. The main outcomes of this study are as follows. First, in the absence of competition, the integration of phonetic information across formants was not affected by the introduction of radical differences between formants in their acoustic source characteristics (harmonic vs sine-wave analogues). Second, the impact of adding F2C was modest when it was tonal, but large when it was harmonic, regardless of whether the source for F2C matched that for F1+F3. This outcome suggests that harmonic analogues are more effective at carrying phonetic information than tonal ones, and provides further evidence against the idea that target-masker similarity is critical for grouping across formants and informational masking between formants. Third, H2C remained a more effective competitor than T2C when F2 and F2C were matched for loudness instead of RMS power. Fourth, an important difference from earlier results using dichotic targets (Roberts *et al.*, 2015) is that, for the monaural targets used here, H2C was no more effective at interfering with the phonetic contribution of T2 than with that of H2. This indicates that the particular advantage for harmonic analogues over tonal ones under competition in dichotic contexts found by Roberts *et al.* (2015) is specific to the additional need to integrate the target formants across ears.

The experiments reported here included conditions where alternative versions of a formant were placed in competition—a context in which differences between versions in the transmission efficiency of phonetic information are likely to be critical. Roberts *et al.* (2015) proposed that the dichotic presentation of a wideband (harmonic) and a narrowband (tonal) candidate for F2 leads to asymmetric informational masking, such that the information carried by the harmonic version tends to overwhelm that carried by the tonal version. Hence, intelligibility is typically highest when the two candidate formants are rendered as H2 and T2C, and lowest when they are T2 and H2C, irrespective of the source properties of the F1+F3 frame. The results obtained here for monaural targets qualify this account, indicating a role for spatial cues under competitive conditions. In particular, note that the distribution of formants across ears used in the previous study

(F1+F2C+F3; F2) provides spatial cues that favor the fusion of the competitor with the F1+F3 frame even when the competitor is mismatched (same ear) and act against fusion of the target F2 (opposite ear).

Taken together, the results from the two studies show that the impact of an interferer on intelligibility may depend on a number of interacting factors and constraints. In particular, the integration of phonetic information across formants with different source characteristics (or perhaps when signaled using different modes of stimulation) may be greatly affected not only by the presence of interferers, but also by the spatial configuration of formants in the ensemble. In particular, the informational masking produced by an interfering formant may be exacerbated under circumstances requiring the integration of target formants across ears. Such a situation may arise for cochlear-implant listeners with residual low-frequency hearing in the non-implanted ear who receive information about F2 and higher formants through the implanted ear and about F1 through the other ear.

In conclusion, the experiments reported here indicate that the effects of source characteristics on the phonetic contributions made by individual formants in an ensemble are governed by type, context, and spatial distribution, rather than by target-masker similarity. The results help to elucidate further how phonetic information is carried by formants and combined across them, particularly in circumstances where interfering formants are present and act mainly as informational maskers. These findings also suggest that there are clinically relevant situations in which listeners combining phonetic information across different modes of stimulation may be particularly susceptible to informational masking.

ACKNOWLEDGMENTS

This research was supported by Research Grant ES/K004905/1 from the Economic and Social Research Council (UK), awarded to B.R. To access the research data underlying this publication, see http://dx.doi.org/10.17036/832e5972-0344-4ae5-b38c-0ff066de3a5f. To access the results of our previous study (Roberts *et al.*, 2015), including the corresponding research data for dichotic targets, see http://dx.doi.org/10.17036/7428fbe0-d7fe-41f5-a53a-32e9726254cd. We are grateful to Quentin Summerfield for enunciating the test sentences. A poster presentation on this research was given at the 169th Meeting of the Acoustical Society of America (Pittsburgh, PA, May 2015).

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

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," software package, version 5.1.28. (Institute of Phonetic Sciences, University of Amsterdam, Amsterdam, the Netherlands), http://www.praat.org/ (Last viewed 9/29/2014).

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

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.

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.

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

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

Darwin, C. J. (**2008**). "Listening to speech in the presence of other sounds," Philos. Trans. R. Soc. B: Biol. Sci. **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.

Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., and Shinn-Cunningham, B. G. (**2003**). "Note on informational masking," J. Acoust. Soc. Am. **113**, 2984–2987.

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

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

Glasberg, B. R., and Moore, B. C. J. (**2002**). "A model of loudness applicable to time-varying sounds," J. Audio. Eng. Soc. **50**, 331–342; software for time-varying loudness (TVL) model, retrieved from http://hearing.psychol.cam.ac.uk/Demos/demos.html (Last viewed 10/15/2014).

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.

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (**2008**). "Informational masking," in *Auditory Perception of Sound Sources, Springer Handbook of Auditory Research*, edited by W. A. Yost and R. R. Fay (Springer, Berlin), Vol. 29, pp. 143–189.

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

Lee, T. Y., and Richards, V. M. (**2011**). "Evaluation of similarity effects in informational masking," J. Acoust. Soc. Am. **129**, EL280–EL285.

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

Neff, D. L. (**1995**). "Signal properties that reduce masking by simultaneous, random-frequency maskers," J. Acoust. Soc. Am. **98**, 1909–1920.

Qin, M. K., and Oxenham, A. J. (**2006**). "Effects of introducing unprocessed low-frequency information on the reception of envelope-vocoder processed speech," J. Acoust. Soc. Am. **199**, 2417–2426.

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.

Rennies, J., Verhey, J. L., and Fastl, H. (**2010**). "Comparison of loudness models for time-varying sounds," Acta Acust. Acust. **96**, 383–396.

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. (**2011**). "The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes," Proc. R. Soc. London B: Biol. Sci. **278**, 1595–1600.

Roberts, B., Summers, R. J., and Bailey, P. J. (**2014**). "Formant-frequency variation and informational masking of speech by extraneous formants:

Evidence against dynamic and speech-specific acoustical constraints," J. Exp. Psychol. Hum. Percept. Perform. **40**, 1507–1525.

Roberts, B., Summers, R. J., and Bailey, P. J. (**2015**). "Acoustic source characteristics, across-formant integration, and speech intelligibility under competitive conditions," J. Exp. Psychol. Hum. Percept. Perform. **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.

Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods*, 6th ed. (Iowa University Press, Ames, IA), pp. 1–310.

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

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.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., and Henry, B. A. (**2004**). "Speech recognition in noise for cochlear implant listeners: Benefits of residual acoustic hearing," J. Acoust. Soc. Am. **115**, 1729–1735.

Verschuur, C., Boland, C., Frost, E., and Constable, J. (**2013**). "The role of first formant information in simulated electro-acoustic hearing," J. Acoust. Soc. Am. **133**, 4279–4289.

Westermann, A., and Buchholz, J. M. (**2015**). "The influence of informational masking in reverberant, multi-talker environments," J. Acoust. Soc. Am. **138**, 584–593.

Zorila, T.-C., Stylianou, Y., Flanagan, S., and Moore, B. C. J. (**2016**). "Effectiveness of a loudness model for time-varying sounds in equating the loudness of sentences subjected to different forms of signal processing," J. Acoust. Soc. Am. **140**, 402–408.