Dichotic integration of acoustic-phonetic information: Competition from extraneous formants increases the effect of second-formant attenuation on intelligibility

Brian Roberts, and Robert J. Summers

Citation: The Journal of the Acoustical Society of America **145**, 1230 (2019); doi: 10.1121/1.5091443

View online: https://doi.org/10.1121/1.5091443

View Table of Contents: https://asa.scitation.org/toc/jas/145/3

Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Perceptual contributions of vowels and consonant-vowel transitions in simulated electric-acoustic hearing
The Journal of the Acoustical Society of America **145**, EL197 (2019); https://doi.org/10.1121/1.5093451

Age effects on the contributions of envelope and periodicity cues to recognition of interrupted speech in quiet and with a competing talker
The Journal of the Acoustical Society of America **145**, EL173 (2019); https://doi.org/10.1121/1.5091664

Hemispheric specializations affect interhemispheric speech sound integration during duplex perception
The Journal of the Acoustical Society of America **145**, EL190 (2019); https://doi.org/10.1121/1.5092829

Error patterns of native and non-native listeners' perception of speech in noise
The Journal of the Acoustical Society of America **145**, EL129 (2019); https://doi.org/10.1121/1.5087271

Segregation of voices with single or double fundamental frequencies
The Journal of the Acoustical Society of America **145**, 847 (2019); https://doi.org/10.1121/1.5090107

Influence of working memory and attention on sound-quality ratings
The Journal of the Acoustical Society of America **145**, 1283 (2019); https://doi.org/10.1121/1.5092808

Dichotic integration of acoustic-phonetic information: Competition from extraneous formants increases the effect of second-formant attenuation on intelligibility

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

(Received 27 November 2018; revised 1 February 2019; accepted 1 February 2019; published online 6 March 2019)

Differences in ear of presentation and level do not prevent effective integration of concurrent speech cues such as formant frequencies. For example, presenting the higher formants of a consonant-vowel syllable in the opposite ear to the first formant protects them from upward spread of masking, allowing them to remain effective speech cues even after substantial attenuation. This study used three-formant (F1+F2+F3) analogues of natural sentences and extended the approach to include competitive conditions. Target formants were presented dichotically (F1+F3; F2), either alone or accompanied by an extraneous competitor for F2 (i.e., F1±F2C+F3; F2) that listeners must reject to optimize recognition. F2C was created by inverting the F2 frequency contour and using the F2 amplitude contour without attenuation. In experiment 1, F2C was always absent and intelligibility was unaffected until F2 attenuation exceeded 30 dB; F2 still provided useful information at 48-dB attenuation. In experiment 2, attenuating F2 by 24 dB caused considerable loss of intelligibility when F2C was present, but had no effect in its absence. Factors likely to contribute to this interaction include informational masking from F2C acting to swamp the acoustic-phonetic information carried by F2, and interaural inhibition from F2C acting to reduce the effective level of F2.

https://doi.org/10.1121/1.5091443

[AKCL] Pages: 1230–1240

I. INTRODUCTION

Speech perception depends on the integration of acoustic-phonetic information that is distributed across frequency and time and, in some circumstances, across ears (e.g., Broadbent and Ladefoged, 1957; Carlson *et al.*, 1975; Roberts *et al.*, 2010) or modes of stimulation (acoustic and electroacoustic hearing—e.g., Turner *et al.*, 2004; see also Verschuur *et al.*, 2013). Indeed, all these forms of integration are needed for successful speech perception by a listener with a cochlear implant in one ear and residual hearing in the other, experiencing a mixed-mode listening scenario in which the higher formants of a speech stimulus are represented in the electrical signal delivered to the implanted ear and the first formant is represented in the low-frequency acoustic signal delivered to the other ear. In most everyday situations, however, we listen to speech in the presence of extraneous sounds—including the speech of other talkers (e.g., Cherry, 1953)—and this can pose a substantial challenge, even for listeners with normal hearing. One aspect of this challenge is energetic masking, in which some of the features of the target speech are partially obscured by the extraneous sounds, but in many circumstances a greater challenge arises from the perceptual allocation of detected features to the appropriate sound sources (e.g., Darwin, 2008) and from the additional processing load required to ignore the irrelevant sounds (e.g., Mattys *et al.*, 2012). These components of the perceptual challenge are examples of informational masking (see, e.g., Brungart *et al.*, 2006; Kidd *et al.*, 2008).

Some aspects of acoustic-phonetic integration remain poorly understood, but it has long been known that this integration can occur quite readily—and in perhaps surprising circumstances—when successful identification of the target speech requires the listener to put together all the acoustical elements presented. For example, listeners are usually able to understand sine-wave analogues of speech, a highly unnatural stimulus in which the lowest few formants of the speech signal are each replaced by a time-varying sinusoid tracking the frequency and amplitude contour of that formant (Bailey *et al.*, 1977; Remez *et al.*, 1981). Listeners are also capable of combining the acoustic-phonetic information carried by formants with different excitation source properties, such as differences in fundamental frequency (F0; e.g., Cutting, 1976) or stimuli for which some formants are rendered as buzz-excited resonances and others as sine-wave analogues (Roberts *et al.*, 2015; Summers *et al.*, 2016).

The intelligibility cost of presenting formants dichotically is usually modest (e.g., Carlson *et al.*, 1975), albeit with occasional changes in consonant identity (e.g., Ainsworth, 1978, 1979). Similar findings have been reported when natural speech is filtered through two narrowband spectral slits centered at 370 Hz and 6000 Hz (Warren *et al.*, 1995). Each band was fairly unintelligible when heard alone

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

(keywords correct: low band = 23%, high band = 24%), but intelligibility was much higher when the bands were presented together, regardless of whether the combination was diotic (78%) or dichotic (76%). Speech perception can also be surprisingly unaffected by changes in the spectral tilt of the speech signal, despite the consequent changes in relative amplitude for the constituent formants (e.g., Ainsworth and Millar, 1972). However, rather less is known about the factors governing the integration of acoustic-phonetic information carried by different formants when success involves putting together some of the formants in the stimulus ensemble but rejecting others. The experiments reported here concern the impact on speech intelligibility of changes in the relative level of the second formant (F2) under dichotic presentation in the presence and absence of an extraneous formant acting primarily as an informational masker.

Previous research on the perceptual effects of changing the relative levels of different formants has generally used isolated synthetic vowels or consonant-vowel (CV) syllables, often with the aim of exploring the status of the formant as a perceptual entity. Several studies using front vowels—for which the frequencies of F1 and F2 are relatively far apart—have shown that their perceived identity is typically maintained over a wide range of relative formant amplitudes (e.g., Lindqvist and Pauli, 1968; Carlson *et al.*, 1970; Ainsworth and Millar, 1972). For example, one study found that the perceived identity of a two-formant analogue of a front vowel remained unchanged until the level of F2 was 28 dB or more below the level of F1 (Ainsworth and Millar, 1972). This outcome indicates some form of feature integration that is relatively insensitive to level differences, presumably one based on formant frequencies (e.g., Klatt, 1985; see also Darwin, 2008). Rather less attenuation of F2 can be tolerated for back vowels because in their case the relative proximity of F1 and F2 soon leads to energetic masking of F2.

In an investigation of intra-speech masking, Rand (1974) found that the identification of synthetic three-formant CV syllables—/*ba*/, /*da*/, and /*ga*/—was affected much less by the attenuation of F2+F3 if those formants were received in the opposite ear to F1 (dichotic presentation) rather than in the same ear (diotic presentation). In the diotic condition, identification remained near-perfect until F2+F3 attenuation approached 20 dB but rolled off steeply thereafter. In the dichotic condition, accuracy remained high until F2+F3 attenuation exceeded 30 dB and subsequent roll-off was shallow until attenuation exceeded 40 dB. Even for the largest F2+F3 attenuation tested (50 dB), dichotic performance remained above chance. Rand (1974) attributed this outcome to dichotic release of the higher formants from energetic masking by the more intense F1. These findings suggest that when energetic masking is limited or prevented—whether by within-ear spectral distance or presentation to opposite ears—the tendency to combine acoustic-phonetic information across formants over a wide range of levels may be a general characteristic of speech perception, at least when there is only one message present.

The generality of this suggestion is explored in the current study by addressing two important questions arising from limitations of previous research on the effects of changes in relative formant amplitude on speech perception. First, the study by Rand (1974) of intra-speech masking used a highly constrained set of stimulus parameters, involving only three response categories (/*ba*/, /*da*/, and /*ga*/). Under these circumstances, the improved performance associated with dichotic presentation may have been based on identifying the direction of the (unmasked) differentiating F2+F3 transitions (cf. Bailey and Herrmann, 1993), rather than on listeners experiencing an integrated percept of the syllables involving both ears. The second limitation is that, to our knowledge, all previous studies investigating the impact on intelligibility of changes in relative formant amplitude have only included formants belonging to the target speech.

It has often been noted that the effects of perceptual organization are usually best revealed in situations where competition arises (see, e.g., Bregman, 1990, pp. 165–172). For example, Barker and Cooke (1999) used diotic mixtures of pairs of sine-wave sentences to demonstrate that there were effects of low-level, non-speech-specific grouping cues (e.g., onset-time differences) on the perceptual organization of sine-wave speech that were not apparent when the sentences were heard in isolation. Note that these mixtures introduced more than one candidate for each of the lowest three formants, which could then compete with one another for inclusion in a particular perceptual organization. Furthermore, some of our recent studies of perceptual organization and informational masking using formant ensembles have shown that the presence or absence of a competitor formant can profoundly affect the extent to which the acoustic-phonetic information carried by a given formant is integrated into a coherent percept (Roberts *et al.*, 2015; Summers *et al.*, 2016). These findings—and their implications for speech perception under adverse listening conditions—are considered further in Sec. IV, in the context of the results of the current study.

The experiments reported here address two questions: (1) Does the relative immunity of intelligibility to changes in relative formant amplitude extend from closed-set isolated vowels and CV syllables to open-set sentence-length materials? (2) If so, does this resilience persist when optimal performance necessitates the integration of some of the formants present but the exclusion of others?

II. EXPERIMENT 1

This experiment explored the extent to which the resilience of speech perception to changes in relative formant amplitude when there is dichotic protection from masking generalizes from a small set of CV syllables to open-set sentence-length stimuli. The dichotic configuration used by Rand (1974) was of the form (left ear = F1; right ear = F2+ F3). In accord with many of our previous studies (Roberts *et al.*, 2010, 2014, 2015; Summers *et al.*, 2010, 2017), we used a variant of this configuration in which F2 is isolated from the other formants (i.e., F1+F3; F2). Note that the second formant—typically associated with the front cavity (see, e.g., Stevens, 1998)—on average carries the most acoustic-phonetic information of all the formants. For example—at least for three-formant sine-wave analogues of speech—removal of F2 typically lowers intelligibility the most and removal of F3 the least; also, for single formant stimuli, F2 is typically the most intelligible and F3 is the least (Han and Chen, 2017). The primary purpose of experiment 1 was to identify, for sentence-length stimuli, the range over which F2 can be attenuated without significant loss of intelligibility when it is protected from energetic masking caused by the other formants. Hence, this experiment did not seek to quantify the benefit of dichotic protection directly, which would have required the inclusion of diotic (or monaural) conditions.

A. Method

1. Listeners

Most listeners were students or members of staff at Aston University and received either course credit 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. 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); around two-thirds of these listeners passed the training and took part in the main experiment. Thirty-six listeners (five males) successfully completed the experiment (mean age = 22.6 yr, range = 18.1–47.3 yr). To our knowledge, none of the listeners had heard any of the sentences used in the main experiment in any previous study or assessment of their speech perception. All were native speakers of English (mostly British) and gave informed consent. The research was approved by the Aston University Ethics Committee.

2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of a collection of short sentences spoken by a British male talker of "Received Pronunciation" English. The text for these recordings 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% or fewer phonemes involving vocal tract closures or unvoiced frication. A set of keywords was chosen for each sentence; most designated keywords were content words. The stimuli for the training session were derived from 50 sentences spoken by a different talker and taken from commercially available recordings of the Harvard sentence lists (IEEE, 1969). These sentences were also selected to contain 25% or fewer phonemes involving closures or unvoiced frication.

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

Synthetic-formant analogues of each sentence were created using the corrected frequency and amplitude contours to control three parallel second-order resonators whose outputs were combined. Following Klatt (1980), alternating signs (+,–,+) were applied to the outputs of the resonators corresponding to F1, F2, and F3 in order to minimize spectral notches whenever adjacent formants were summed in the same ear. Although this situation did not occur in the current experiment, it did arise in experiment 2 (see Sec. III A). A monotonous source with an F0 of 140 Hz was used to synthesize all stimuli for the training and main experiment. The excitation source was a periodic train of simple excitation pulses modeled on the glottal waveform, which Rosenberg (1971) has shown to be capable of producing synthetic speech of good quality. The 3-dB bandwidths of the resonators corresponding to F1, F2, and F3 were set to constant values of 50, 70, and 90 Hz, respectively.

There were 12 conditions in the main experiment (see Table I), including 2 control conditions used to assess performance for F2 alone (C1) and F1+F3 alone (C2). The remaining conditions (C3-C12) comprised all three formants presented in a dichotic configuration (F1+F3; F2), for which the attenuation applied to F2 relative to its baseline level ranged from 48 dB to −6 dB (i.e., a 6-dB boost) in 6-dB steps. The main focus of the experiment was on the effect of F2 attenuation, but the boosted-F2 case was included to test whether raising the relative level of F2 would have a deleterious effect on intelligibility. This is because pilot observations had suggested that the consequent greater prominence of F2 might draw attention away from F1+F3 in the left ear. High-quality reproduction of the F2 stimulus was maintained across the range of attenuations tested by using programmable attenuators (see below) to set the desired output levels independently at each ear. The stimuli are illustrated in Fig. 1 using the wideband spectrogram of a synthetic analogue of an example sentence. The left- and right-hand panels show the effect of attenuating F2 when presented in the same ear as F1+F3 or in the opposite ear, respectively; the ipsilateral configuration was not tested but is included to allow comparison. For each listener, the 48 sentences were divided equally across conditions (four per condition) such that there were 13–14 keywords per condition. Allocation of sentences was counterbalanced by rotation across each set of 12 listeners tested. Hence, the total number of listeners needed to produce a balanced dataset was a multiple of 12.

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

TABLE I. Stimulus properties for the conditions used in experiment 1 (main session).

Condition	Stimulus configuration (left ear; right ear)	F2 attenuation (dB)
C1	(—; F2)	0
C2	(F1+F3; —)	∞
C3	(F1+F3; F2)	48
C4	(F1+F3; F2)	42
C5	(F1+F3; F2)	36
C6	(F1+F3; F2)	30
C7	(F1+F3; F2)	24
C8	(F1+F3; F2)	18
C9	(F1+F3; F2)	12
C10	(F1+F3; F2)	6
C11	(F1+F3; F2)	0
C12	(F1+F3; F2)	−6

experiment consisted of a training session followed by the main session and typically took about 45 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; all stimuli were presented diotically, without competitors, and a new sentence was used for each trial. On each of the first ten 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 the 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 that the strategy of providing feedback using alternating presentations of the synthetic and original versions was an efficient way of enhancing the perceptual learning of speech-like stimuli. The final ten trials of the training differed in that listeners heard the stimulus only once before entering their transcription; they continued to receive feedback. Listeners progressed to the main experiment if they met either or both of two criteria: (1) ≥50% keywords correct across all 40 trials requiring a transcription (30 with repeat listening; 10 without); (2) ≥50% keywords correct for the final 15 trials with repeat listening. In the main experiment, listeners were allowed to hear each stimulus only once before entering their transcription and no feedback was given. An additional criterion for inclusion in the final dataset was obtaining a mean score of ≥20% keywords correct in the main experiment when collapsed across conditions. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. All listeners who passed the training also met this criterion.

All speech analogues were synthesized using MITSYN (Henke, 2005) at a sample rate of 22.05 kHz and with 10-ms raised-cosine onset and offset ramps. They were played at 16-bit resolution over Sennheiser HD 480-13II earphones (Hannover, Germany) via a Sound Blaster X-Fi HD sound card (Creative Technology Ltd., Singapore), a pair of programmable attenuators (Tucker-Davis Technologies, TDT

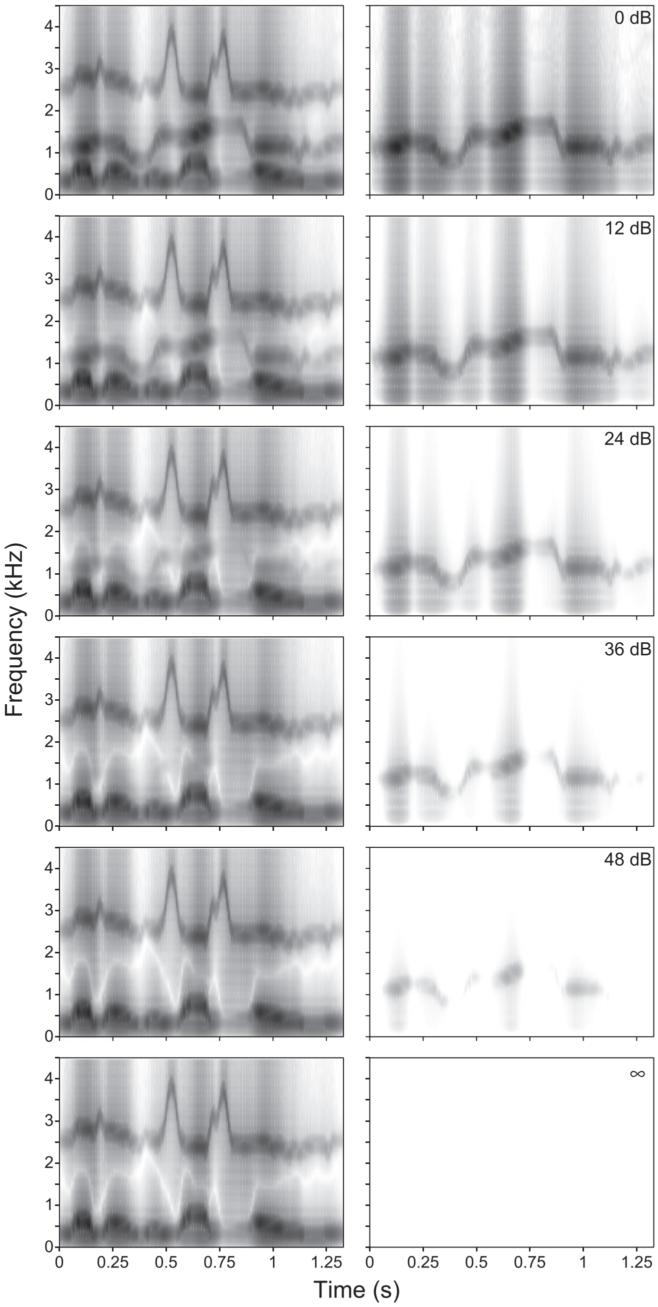

FIG. 1. Stimuli for experiment 1—wideband spectrograms illustrating the effect of attenuating F2 in two different contexts for the example sentence "Mother was at home." The right-hand panels show the effect of attenuating F2 when it is presented in the opposite ear to F1+F3, and is therefore protected from energetic masking. For comparison (not tested in the experiment), the left-hand panels show the effect of attenuating F2 when it is presented in the same ear as F1+F3, for which F2 is subject to energetic masking. In descending order from top to bottom panels, the F2 attenuations shown are 0 dB, 12 dB, 24 dB, 36 dB, 48 dB, and ∞, respectively. The gray scale used is set such that any frequency-time region exceeding 20 dB SPL is visible in the spectrogram. When F2 is presented in isolation (right-hand panels), it is sufficiently intense that some parts of its trajectory are visible even when attenuated by 48 dB and hence, in principle, F2 remains able to provide useful acoustic-phonetic information. When F2 is accompanied by F1+F3 (left-hand panels), it is hard to discern any part of its trajectory when it is attenuated by more than 24 dB.

PA5; Alachua, FL), each controlling the output to one ear, and a headphone buffer (Tucker-Davis Technologies, 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 (Brüel and Kjaer, type 4153). Stimuli in the main experiment were presented at a reference level (long term average) of 75 dB sound pressure level (SPL); this describes the case where the left ear received F1 (the most intense formant) and F3. There was inevitably some variation in the presentation level of F2 in the right ear owing to natural variation between sentences (mean without attenuation ~67 dB SPL) but most of the variation was a consequence of the range of attenuations used (mean for 6-dB boost ~73 dB SPL; mean for 48-dB attenuation ~19 dB SPL). In the training session, the presentation level of the diotic materials used was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation.

4. Data analysis

For each listener, the intelligibility of each stimulus was quantified using keyword scoring as the main 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; homonyms were accepted. Responses were classified using tight scoring, in which a response is scored as correct only if it matches the keyword exactly (see, e.g., Foster *et al.*, 1993; Roberts *et al.*, 2010). Following Roberts *et al.* (2014), phonemic scoring was used as an additional measure of intelligibility. Typed responses were converted automatically into phonemic representations using eSpeak (Duddington, 2014), which generates phonemic representations of the input text using a pronunciation dictionary and a set of generic pronunciation rules for English orthography. The mean percentage of phonemes correctly identified across all words in the sentences was computed using an algorithm that finds an optimal alignment between the sequence of phonemes for the original sentence and its transcription through 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 reported here were computed using R 3.5.1 (R Core Team, 2018) and the *ez* analysis package (Lawrence, 2016). The measures of effect size reported here are eta squared (η^2) and partial eta squared (η_p^2). All *a posteriori* pairwise comparisons (two-tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967; Keppel and Wickens, 2004). Unless otherwise stated, all statistics presented here were computed using keyword scores; statistics computed using phoneme scores are presented only on occasions where 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 (upper panel) and phonemes (lower panel) correctly identified. In each panel, the open circles, open diamond, and

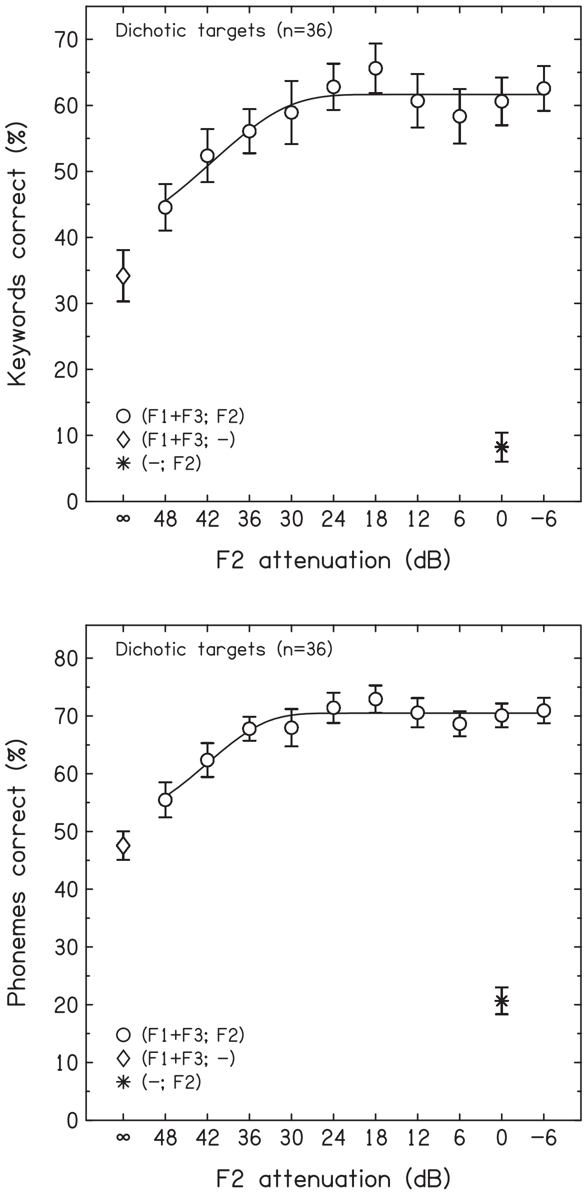

FIG. 2. Results for experiment 1—effect of F2 attenuation on the intelligibility of three-formant analogues of the target sentences under dichotic presentation. The top panel displays the mean keyword scores and intersubject standard errors ($n = 36$) for the experimental conditions (open circles), F1+F3 control condition (open diamond), and F2 control condition (asterisk). The inset indicates which formants were presented to each ear; the bottom axis indicates the attenuation applied to F2. The bottom panel displays the corresponding means and standard errors for the phoneme scores. In each panel, the set of mean scores for F2 attenuations from 48 dB to −6 dB (i.e., 6-dB boost) has been fitted using a Weibull function (solid line) for which the equation is $\Psi(x) = \gamma + (1 - \gamma - \lambda)(1 - \exp(-(x/\alpha)^{\beta}))$, where $\Psi(x)$ is the proportion correct score and x is the attenuation in linear units. For the fit to the keyword scores, the guess-rate, γ, was set to 0.341 (the score for F1+F3 alone) and the remaining parameter values were $\lambda = 0.383$ (lapse error rate), $\alpha = 0.00859$ (point of inflection), and $\beta = 0.812$ (slope). The corresponding parameter values for the fit to the phoneme scores were: $\gamma = 0.476$, $\lambda = 0.295$, $\alpha = 0.00818$, and $\beta = 1.068$. These fits were good: $r^2(8) = 0.891$ (keywords) and 0.934 (phonemes).

asterisk indicate the results for the experimental conditions (F2 attenuation), the F1+F3 control, and the F2-only control, respectively. For the experimental conditions, each set of scores has been fitted using a Weibull function (Wichmann and Hill, 2001) to give a psychometric function describing the influence of F2 attenuation on the intelligibility of three-formant analogues of the target sentences. As would be expected, the two functions were similar in form but the mean phoneme scores were consistently higher than their keyword counterparts. Note that intelligibility was relatively good (∼60% keywords correct) in the reference condition [C11: (F1+F3; $F2_{0dB}$)] despite the dichotic presentation of the target formants and the simple source properties and three-formant parallel vocal-tract model used to synthesize the sentences.

A one-way within-subjects analysis of variance (ANOVA) across all 12 conditions showed a highly significant effect of condition on intelligibility [$F(11,385) = 29.822$, $p < 0.001$, $\eta^2 = 0.460$].[1] Keyword scores for the control conditions showed that intelligibility was fairly low for F1+F3 alone (C2) and near floor for F2 alone (C1). Pairwise comparisons indicated that the mean scores for C1 and C2 differed from those for all other conditions, including each other (range: $p = 0.019 - p < 0.001$). In particular, note that accompanying F1+F3 with F2 in the other ear greatly improved keyword scores [C2 vs C11 = 26.4 percentage points (% pts); $t(35) = 6.454$, $p < 0.001$] and that, although reduced, the benefit remained considerable even when F2 was attenuated by 48 dB [C2 vs C3 = 10.4% pts; $t(35) = 2.455$, $p = 0.019$]. The effect of attenuating F2 on its contribution to overall intelligibility was explored using a one-way ANOVA restricted to the conditions containing all three target formants (C3–C12). The effect of F2 attenuation on keyword scores was significant overall but the effect size was modest [$F(9,315) = 3.824$, $p < 0.001$, $\eta^2 = 0.098$], and pairwise comparisons revealed that the fall in keyword scores was significant only when F2 was attenuated by 48 dB [C3 vs C11 = 16.1% pts; $t(35) = 4.563$, $p < 0.001$]. Although the effect of attenuating F2 by 42 dB did not quite reach significance for the keyword scores [C4 vs C11 = 8.2% pts; $t(35) = 1.856$, $p = 0.072$], the effect on the (less variable) phoneme scores was significant [C4 vs C11 = 7.7% pts; $t(35) = 2.226$, $p = 0.033$]. None of the other levels of attenuation tested led to a significant change in either the keyword or phoneme scores relative to those for the reference condition (C11).

A more conservative approach is to estimate how much attenuation of F2 is necessary to have any discernible impact on intelligibility. This estimate can be obtained from the point of maximum rate of change in the gradient of the Weibull function. According to this measure, the smallest attenuations needed on average to lower the keyword and phoneme scores were 31.0 dB and 33.9 dB, respectively. However, perhaps the most striking outcome of this experiment is that F2 made a moderate contribution to intelligibility even when it was attenuated by 48 dB. As noted earlier, on average this corresponds to a presentation level of ∼19 dB SPL, for which we can be confident that only parts of the trajectory of F2 would have been audible to our listeners. On occasions when F2 was so greatly attenuated, acoustic-phonetic information about the front cavity will have been best preserved for the vocalic nuclei.

Although this experiment did not test directly the hypothesis put forward by Rand (1974), the results are clearly in accord with it. This is because it is hard to envisage any way in which such a highly attenuated F2 could contribute useful acoustic-phonetic information without the dichotic protection from masking offered by the stimulus configuration used (see Fig. 1). This suggests that the findings reported by Rand (1974) are not restricted to a small closed set of synthetic CV syllables, but extend to open-set sentence-length stimuli. Clearly, the auditory system is capable of combining information carried by different target formants over a wide range of relative levels.

III. EXPERIMENT 2

In experiment 1, and the study reported by Rand (1974), optimal performance required the listener to integrate across ears the acoustic-phonetic information carried by all the formants presented. Experiment 2 explored whether the same tolerance of F2 attenuation occurs when an extraneous formant in the stimulus ensemble provides an alternative candidate for the second formant, in the opposite ear to F2, referred to as the second-formant competitor (F2C). The properties of F2C were chosen such that it carried misleading acoustic-phonetic information that would impair intelligibility unless the competitor was excluded from the percept of the target sentence (cf. Remez *et al.*, 1994; Roberts *et al.*, 2010). The perceptual challenge to the listener was set high by presenting F2C without attenuation and in the same ear as F1 and F3, thus encouraging its fusion with them. Note that by keeping the target F2 isolated in the right ear, the impact on intelligibility of the competitor—particularly any interaction with the effects of attenuating F2—cannot be attributed to energetic masking, but can instead be attributed to informational masking. The stimulus configuration used and the task requirements for listeners are illustrated in Fig. 3.

A. Method

Except where described, the same method was used as for experiment 1. Thirty listeners (four males) passed the training and successfully completed the main experiment (mean age = 23.6 yr, range = 18.3–41.2 yr); there were no exclusions based on the additional criterion of a mean overall score of ≥20% keywords correct in the main experiment. The training session was identical to that used in experiment 1; no competitor formants were presented. The stimuli for the main experiment were derived from the same collection of recordings as were used in experiment 1; 47 of the 50 sentences were the same as those used in experiment 1.

The target stimuli were created in the same way as before. In addition, some conditions included—in the same ear as F1+F3—a competitor for F2 (F2C) created by inverting the F2 frequency contour (about its geometric mean) and using the F2 amplitude contour. Several studies have shown that single-formant competitors with time-varying frequency contours are effective informational maskers (e.g., Roberts

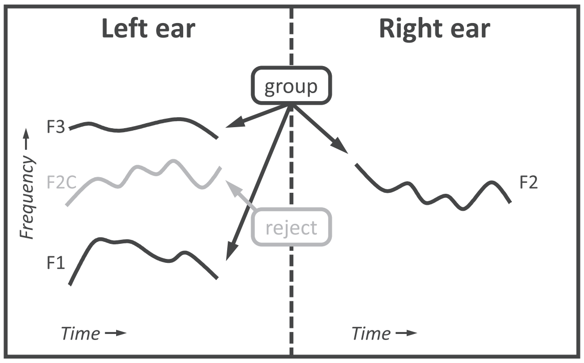

FIG. 3. Stimuli for experiment 2—schematic illustrating the dichotic configuration of formants used and the task requirements for listeners. For the target sentence, the left ear receives F1+F3 and the right ear receives F2. When present, the second-formant competitor (F2C) is received in the same ear as F1+F3. To optimize intelligibility, listeners must integrate the target formants across ears but reject F2C when it is present.

et al., 2010, 2014; Roberts and Summers, 2015, 2018; see also Summers *et al.*, 2012). All F2Cs were generated using the same excitation source (Rosenberg pulses), F0 (140 Hz), 3-dB bandwidth (70 Hz), and output sign (–) as used to synthesize F2; the root-mean-square (RMS) level of F2C was always set to match that of the target F2 at 0-dB attenuation. Stimuli were selected such that the frequency of F2C was always at least 80 Hz away from the frequencies of F1 and F3 at any one moment. Hence, there were no approaches between formant tracks close enough to cause audible interactions between corresponding harmonics exciting adjacent formants. Note that the addition of F2C to the left ear had little effect on overall presentation level (always <1 dB) owing to the spectral tilt of natural speech.

There were ten conditions in the main experiment (see Table II). The left ear received F1+F3 in five conditions (C1–C5) and F1+F2C+F3 in the other five (C6–C10). There were two control conditions in which the target F2 was absent (C1 and C6). The remainder (C2–C5 and C7–C10) included F2 in the right ear with attenuations ranging from 36 dB to 0 dB in 12-dB steps. For each listener, the 50 sentences used were divided equally across conditions (5 per condition); there were 16–17 keywords per condition. For a balanced dataset, allocation of sentences to the ten conditions was counterbalanced by rotation across each set of ten listeners.

TABLE II. Stimulus properties for the conditions used in experiment 2 (main session).

Condition	Stimulus configuration (left ear; right ear)	F2 attenuation (dB)
C1	(F1+F3; —)	∞
C2	(F1+F3; F2)	36
C3	(F1+F3; F2)	24
C4	(F1+F3; F2)	12
C5	(F1+F3; F2)	0
C6	(F1+F2C+F3; —)	∞
C7	(F1+F2C+F3; F2)	36
C8	(F1+F2C+F3; F2)	24
C9	(F1+F2C+F3; F2)	12
C10	(F1+F2C+F3; F2)	0

B. Results and discussion

Figure 4 shows the mean percentage scores (and intersubject standard errors) across conditions for keywords (upper panel) and phonemes (lower panel). In each panel, the circles and diamonds indicate the results for the experimental conditions (F2 attenuation) and control conditions (F2 absent), respectively; the presence or absence of the competitor (F2C) was shown using filled and open symbols,

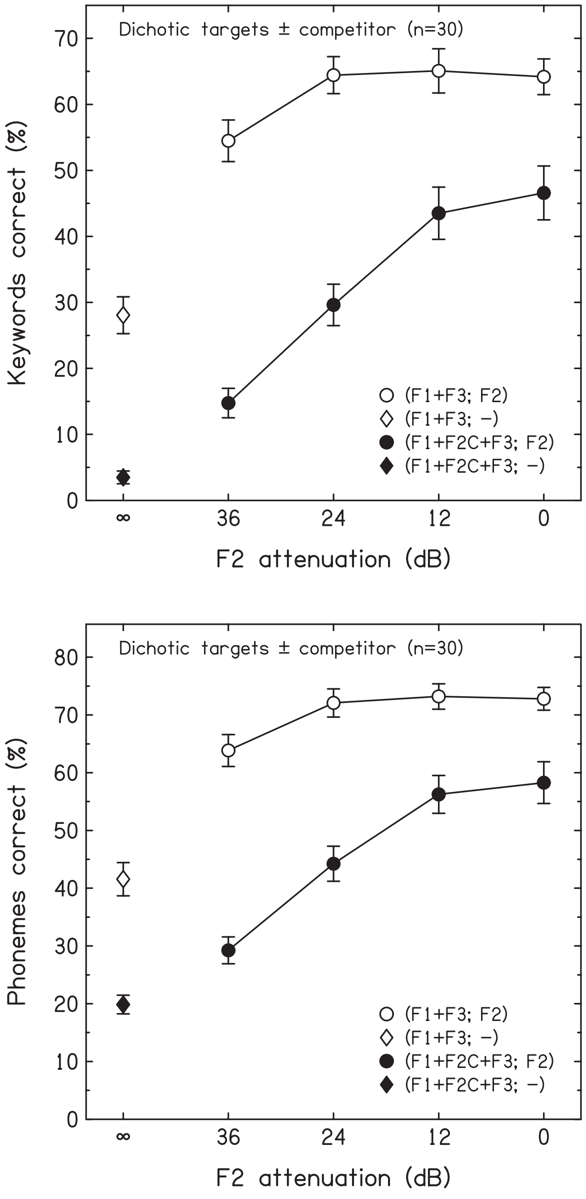

FIG. 4. Results for experiment 2—effect of F2 attenuation on the intelligibility of three-formant analogues of the target sentences under dichotic presentation in the presence and absence of an extraneous formant (F2C). The top panel displays the mean keyword scores and intersubject standard errors ($n = 30$) for the experimental conditions (circles) and the F1+F3 control conditions (diamonds); conditions for which F2C was present or absent are shown by filled and open symbols, respectively. The inset indicates which formants were presented to each ear; the bottom axis indicates the attenuation applied to F2. The bottom panel displays the corresponding means and standard errors for the phoneme scores.

respectively. Once again, intelligibility was fairly good (~65% keywords correct) in the reference condition [C5: (F1+F3; $F2_{0dB}$)] and the results for the phoneme scores showed a similar pattern to those for the keyword scores. Visual inspection of Fig. 4 suggests that including the competitor had two effects—it lowered overall intelligibility and also increased the impact of F2 attenuation. Therefore, the effects of the main stimulus manipulations (F2C inclusion/exclusion and F2 attenuation) were explored using a two-way within-subjects ANOVA restricted to the eight conditions in which the target F2 was present (C2–C5 and C7–C10). This analysis revealed significant main effects of F2C inclusion [mean difference = 28.4% pts; $F(1,29) = 83.409$, $p < 0.001$, $\eta_p^2 = 0.742$] and F2 attenuation [$F(3,87) = 19.357$, $p < 0.001$, $\eta_p^2 = 0.400$]. Moreover, the interaction between them was significant [$F(3,87) = 7.285$, $p < 0.001$, $\eta_p^2 = 0.201$], which is consistent with the observation that keyword scores fell much more steeply with F2 attenuation over the range tested (0–36 dB) when the competitor was present.

The effect of attenuation was explored further using separate one-way ANOVAs for C2–C5 (F2C absent) and C7–C10 (F2C present). Similar to the results for experiment 1, when the competitor was absent there was a significant effect of attenuation, but the effect size was modest [$F(3,87) = 3.007$, $p = 0.035$, $\eta_p^2 = 0.094$]. Pairwise comparisons indicated that keyword scores did not fall until F2 attenuation was increased to 36 dB [C2 vs C5 = 9.7% pts; $t(29) = 2.521$, $p = 0.017$]. When the competitor was present, the effect of attenuation was highly significant and the effect size was much larger [$F(3,87) = 24.831$, $p < 0.001$, $\eta_p^2 = 0.461$]. Pairwise comparisons indicated that there was a significant fall in keyword scores when F2 attenuation was 24 dB [C8 vs C10 = 17.0% pts; $t(29) = 4.320$, $p < 0.001$] or 36 dB [C7 vs C10 = 31.8% pts; $t(29) = 7.730$, $p < 0.001$], but not when it was 12 dB [C9 vs C10 = 3.1% pts; $t(29) = 0.713$, $p = 0.481$]. Keyword scores for the control conditions showed that intelligibility was fairly low for F1+F3 alone (C1) and near floor when F2C was also present (C6). Pairwise comparisons between the control and experimental conditions indicated that accompanying F1+F3 with F2 in the other ear improved keyword scores for all levels of attenuation tested ($p < 0.001$ in all cases) regardless of whether the competitor was absent (C1 vs C2–C5) or present (C6 vs C7–C10). This outcome indicates that, in either context, the target F2 continued to contribute some useful acoustic-phonetic information even when it was attenuated by 36 dB.

The most compelling aspect of these findings is the effect of adding F2C when the target F2 was attenuated by 24 dB. This level of attenuation caused no diminution in keyword scores in the absence of the competitor, but resulted in a substantial fall in intelligibility when the competitor was present, such that the mean difference scores between the corresponding pairs of conditions roughly doubled from 17.6% pts for the 0-dB case (C5 vs C10) to 34.8% pts for the 24-dB case (C3 vs C8). Hence, the steeper roll-off of scores when F2 was attenuated in the presence of F2C cannot be attributed to the lower overall intelligibility of the conditions in which the competitor was included. The results for the 12-dB case, for which the attenuation of F2 had little or no impact irrespective of whether F2C was present, indicate that the largest change in the modulating effect of the two stimulus contexts takes place somewhere in the range 12–24 dB of F2 attenuation. This finding has informed our discussion (see below) of the kinds of mechanism that might plausibly account for the effects of the competitor formant.

IV. GENERAL DISCUSSION

Previous studies of the effects of changes in relative formant amplitude on the dichotic integration of acoustic-phonetic information have typically been restricted to small closed sets of isolated vowels or CV syllables, and have used identification tasks for which optimum performance required listeners to integrate all formants present in the stimulus ensemble. The experiments reported here have extended these investigations to include open-set sentence-length materials and competitive conditions in which optimum performance required listeners to integrate the target formants across ears while excluding an extraneous formant. The results of experiment 1 suggest that the benefit of providing dichotic protection for the higher formants from energetic masking by F1 extends to open-set sentence-length stimuli. In circumstances where the stimulus ensemble comprises only the target formants, F2 can be attenuated by at least 30 dB without any loss of intelligibility when it is presented in the opposite ear to F1+F3. Furthermore, F2 still carries some useful acoustic-phonetic information when it is attenuated by 48 dB (corresponding to a long-term average of ~19 dB SPL). Experiment 2 compared the effects of F2 attenuation in the presence and absence of an extraneous formant intended to provide an alternative perceptual possibility for the second formant. Decreasing the level of F2 was much more disruptive when the competitor (F2C) was included in the same ear as F1 and F3, such that an attenuation greater than 12 dB led to a substantial loss of intelligibility.

What do these results tell us about how the acoustic-phonetic information carried by different formants is integrated into a coherent speech percept? First of all, the finding that the intelligibility of open-set sentence-length materials—when heard in isolation and when intra-speech masking is controlled—is relatively immune to substantial changes in relative formant amplitude suggests that the propensity of the auditory system to combine acoustic-phonetic information across frequency and ears extends to across levels. It should be acknowledged, however, that there are other circumstances in which differences in presentation level can serve as a segregation cue. For example, the performance of listeners asked to attend selectively to one of two competing sentences—a situation in which speech-on-speech masking is primarily informational—sometimes improves when the signal-to-noise ratio is lowered from 0 dB to −9 dB (Brungart, 2001). Of course the failure of listeners in the current study to segregate the competitor from the target formants is unsurprising, given the primitive grouping cues for its inclusion that arise from presenting it in the same ear as F1+F3 and on the same F0 as all the target formants (see, e.g., Bregman, 1990; Darwin, 2008). However, its inclusion in a context where the relative level of the target F2 is manipulated can provide insights into how the useful acoustic-phonetic information carried by F2 interacts with the misleading information carried by F2C. The finding that the presence of F2C caused intelligibility to begin falling earlier and to fall more steeply as F2 attenuation was increased suggests that one or other (or both) of two processes are involved. These possibilities are considered in turn.

The first possibility is that some form of mandatory summation of acoustic-phonetic information takes place across ears, leading to progressive dilution of the information carried by the target F2 as its level is attenuated relative to that of the full-scale F2C. By this account, intelligibility falls steeply when F2 is attenuated under competitive conditions because the misleading information provided by the informational masker, F2C, soon becomes dominant in the weighted summation, swamping the useful acoustic-phonetic information carried by F2. The second possibility involves some form of interaural inhibition in which the more intense candidate for the second formant—the competitor—increasingly prevents either the extraction of the acoustic-phonetic information carried by the attenuated target F2 or its integration with the information carried by the contralateral target formants. A plausible mechanism for an account based on interaural inhibition is that the presence of the relatively intense competitor in the other ear reduces the effective internal level of the attenuated F2, such that less acoustic-phonetic information is available to be extracted from it. This account is consistent with the data of Scharf (1969) on the dichotic summation of loudness and the model of loudness developed by Moore and his colleagues (Moore and Glasberg, 2007). Anecdotally, informal listening to our stimuli suggests that the presence of F2C can influence one's perception of the trajectory of the attenuated F2 even when trying to focus attention only on the right ear.

Regarding the question of whether both these processes are relevant to the integration of acoustic-phonetic information across ears, it is perhaps worth noting that a computational model of vision incorporating interocular suppression (inhibition) prior to binocular integration (Baker *et al.*, 2007) has been very successful in accounting for the results of psychophysical binocular contrast discrimination and matching experiments. Indeed, this approach has recently been extended successfully to modeling thresholds for amplitude-modulation depth discrimination for various binaural stimulus configurations on the assumption that interaural inhibition is weaker than interocular inhibition (Baker *et al.*, 2018). Although there are clearly limits to the analogy that can be drawn with combining suprathreshold sources of acoustic-phonetic information—e.g., F2 and F2C occupy the same frequency region but their trajectories are very different—this general approach may offer a way of accounting for results like those reported here.

Although not conclusive, one aspect of our results that appears more consistent with an account based on interaural inhibition is the finding that attenuating F2 by 12 dB in the presence of F2C has little or no effect on intelligibility. If mandatory weighted spectral integration takes place across ears, one might expect the acoustic-phonetic information carried by F2 to begin to contribute less as soon as F2 has been attenuated by more than a few dB relative to F2C. If, however, an important effect of adding F2C is interaural inhibition that leads to a reduction in the effective level of the attenuated F2, then the effect of adding F2C can be regarded as similar to the effect of increasing F2 attenuation. Given that, in the absence of F2C, substantial F2 attenuation is required before any impact on intelligibility becomes apparent, one might expect the additional effect of interaural inhibition to be revealed only when a sufficient "baseline" attenuation has already been applied to F2. By this account, the results of experiment 2 suggest that the required baseline attenuation is ∼12 dB.

In a broader context, note that a full understanding of the perceptual organization of speech remains elusive. On the one hand, there is a wealth of evidence that primitive grouping principles are important in holding together the rapidly changing and acoustically diverse elements of speech, such as the continuity cues provided by formant transitions and the pitch contour (e.g., Cole and Scott, 1973; Darwin and Bethell-Fox, 1977; Stachurski *et al.*, 2015), and in separating the speech of one talker from another, such as differences in onset time and F0 (e.g., Darwin, 1981, 1984; Bird and Darwin, 1998; Barker and Cooke, 1999). On the other hand, there is a clear distinction between the factors influencing the spoken message heard and the number of voices heard. For example, Cutting (1976) found that introducing differences in F0 between formants in a dichotic ensemble typically led to listeners reporting more than one voice but a single message, indicating the integration of acoustic-phonetic information carried by all the formants presented despite the perception of multiple sources.

Even when a formant ensemble is configured such that the physical exclusion of one formant changes one intelligible stimulus (/ru/) into another (/li/), introducing a difference in F0 of nine semitones between that formant and the others is not sufficient to eliminate the /ru/ percept (Darwin, 1981; Gardner *et al.*, 1989). Furthermore, under competitive conditions, there are circumstances in which listeners fail to combine formants with shared acoustic source properties and to exclude extraneous formants with radically mismatched source properties—notably, when all the target formants are sine-wave analogues and the extraneous formant is rendered as a buzz-excited resonance (Roberts *et al.*, 2015; Summers *et al.*, 2016; see below). Some researchers have appealed to a speech-specific notion of phonetic coherence based on the plausibility of the articulatory gestures implied by the time-varying properties of formants in an ensemble rather than on general-purpose grouping cues (see, e.g., Liberman, 1982; Mann and Liberman, 1983; Remez *et al.*, 1994; Remez, 2001, 2003, 2005), but to our knowledge this concept has never been clearly defined acoustically. Moreover, Roberts *et al.* (2014) showed that the ability of a time-varying extraneous formant to impair the intelligibility of a target sentence did not depend on whether the pattern of frequency variation in the interferer was plausibly speech-like (inverted F2 frequency contour) or not (contour derived from a periodic triangle wave).

The experiments reported here were not intended to resolve these issues but the results help to constrain what is needed from a full account of the perceptual organization of speech. In particular, there is an interesting parallel between the results of the current study and some of our previous work exploring how listeners integrate and segregate formants in stimulus ensembles with mixed source properties (Roberts *et al.*, 2015; Summers *et al.*, 2016). Most notably, our ability to use the acoustic-phonetic information carried by a sine-wave analogue of F2 may be impaired greatly by the presence of an RMS-matched buzz-excited F2C in the other ear, but not when the source properties of F2 and F2C are reversed. This is the case even when the other target formants—F1 and F3—are also sine-wave analogues. These findings, along with the different effects of F2 attenuation in the presence and absence of F2C reported here, suggest that the acoustical (spectro-temporal) context within which each element of a stimulus ensemble is presented may play an important role in across-formant integration, even in circumstances where energetic masking is of little, if any, importance.

In terms of the clinical relevance of these findings, let us return to the case in which a listener is fitted with a cochlear implant in one ear and has residual hearing in the other. Such a listener may easily be able to integrate the acoustic-phonetic information received by the two ears when only the target formants are present—despite substantial differences in relative level and mode of stimulation—but may lose this ability in the presence of an extraneous formant-like sound acting primarily as an informational masker. To conclude, there is now a growing body of evidence that spectro-temporal context can be critical to our ability to integrate relevant acoustic-phonetic information, and to overcome interference, even in situations where that information is available and unmasked in the peripheral auditory system.

ACKNOWLEDGMENTS

This research was supported by Research Grant No. ES/N014383/1 from the Economic and Social Research Council (UK), awarded to B.R. To access the research data underlying this publication, see https://doi.org/10.17036/researchdata.aston.ac.uk.00000396. We thank Peter Bailey for his advice concerning this research, Brian Moore for pointing out the relevance to our findings of the literature on across-ear loudness summation, and both of them for their comments on earlier drafts of this manuscript. We are also grateful to Sheradan Miller for his assistance with data collection during the pilot phase of this research. Poster presentations on this research were given at the 175th Meeting of the Acoustical Society of America (Minneapolis, MN, USA, May 2018), and the annual Basic Auditory Science Meeting of the British Society of Audiology (Newcastle-upon-Tyne, UK, September 2018).

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

Ainsworth, W. A. (**1978**). "Perception of speech sounds with alternate formants presented to opposite ears," J. Acoust. Soc. Am. **63**, 1528–1534.

Ainsworth, W. A. (**1979**). "Perception of dichotically presented formants," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Öhman (Academic, London), pp. 135–142.

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

Bailey, P. J., and Herrmann, P. (**1993**). "A reexamination of duplex perception evoked by intensity differences," Percept. Psychophys. **54**, 20–32.

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. **51/52**, 1–25.

Baker, D. H., Meese, T. S., and Georgeson, M. A. (**2007**). "Binocular interaction: Contrast matching and contrast discrimination are predicted by the same model," Spat. Vis. **20**, 397–413.

Baker, D. H., Vilidaite, G., McClarnon, E., Valkova, E., and Millman, R. E. (**2018**). "Binaural fusion involves weak interaural suppression," bioRxiv on-line preprint, available at http://biorxiv.org/content/early/2018/03/07/278192.abstract (Last viewed 25 October 2018).

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

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

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

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, available at 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).

Broadbent, D. E., and Ladefoged, P. (**1957**). "On the fusion of sounds reaching different sense organs," J. Acoust. Soc. Am. **29**, 708–710.

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

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

Carlson, R., Fant, G., and Granström, B. (**1975**). "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 55–82.

Carlson, R., Granström, B., and Fant, G. (**1970**). "Some studies concerning perception of isolated vowels," Speech Trans. Lab. Q. Prog. Stat. Rep. **2-3**, 19–35.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and with two ears," J. Acoust. Soc. Am. **25**, 975–979.

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

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. (**1984**). "Perceiving vowels in the presence of another sound: Constraints on formant perception," J. Acoust. Soc. Am. **76**, 1636–1647.

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.

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

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.

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.

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

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

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

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.

Klatt, D. H. (**1985**). "The perceptual reality of a formant frequency," J. Acoust. Soc. Am. **78**, S81–S82.

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

Liberman, A. M. (**1982**). "On finding that speech is special," Am. Psychol. **37**, 148–167.

Lindqvist, J., and Pauli, S. (**1968**). "The role of relative spectrum levels in vowel perception," Speech Trans. Lab. Q. Prog. Stat. Rep. **2-3**, 12–15.

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

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Lang. Cognit. Process. **27**, 953–978.

Moore, B. C. J., and Glasberg, B. R. (**2007**). "Modeling binaural loudness," J. Acoust. Soc. Am. **121**, 1604–1612.

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.

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

R Core Team (**2018**). "The R project for statistical computing [software package]," The R Foundation, Vienna, Austria, available at https://www.R-project.org/ (Last viewed 30 July 2018).

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

Remez, R. E. (**2001**). "The interplay of phonology and perception considered from the perspective of perceptual organization," in *The Role of Speech Perception in Phonology*, edited by E. Hume and K. Johnson (Academic, San Diego), pp. 27–52.

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

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

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

Roberts, B., and Summers, R. J. (**2015**). "Informational masking of monaural target speech by a single contralateral formant," J. Acoust. Soc. Am. **137**, 2726–2736.

Roberts, B., 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., 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.

Scharf, B. (**1969**). "Dichotic summation of loudness," J. Acoust. Soc. Am. **45**, 1193–1205.

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

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).

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.

Summers, R. J., Bailey, P. J., and Roberts, B. (**2010**). "Effects of differences in fundamental frequency on across-formant grouping in speech perception," J. Acoust. Soc. Am. **128**, 3667–3677.

Summers, R. J., Bailey, P. J., and Roberts, B. (**2012**). "Effects of the rate of formant-frequency variation on the grouping of formants in speech perception," J. Assoc. Res. Otolaryngol. **13**, 269–280.

Summers, R. J., Bailey, P. J., and Roberts, B. (**2016**). "Across-formant integration and speech intelligibility: Effects of acoustic source properties in the presence and absence of a contralateral interferer," J. Acoust. Soc. Am. **140**, 1227–1238.

Summers, R. J., Bailey, P. J., and Roberts, B. (**2017**). "Informational masking and the effects of differences in fundamental frequency and fundamental-frequency contour on phonetic integration in a formant ensemble," Hear. Res. **344**, 295–303.

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.

Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (**1995**). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," Percept. Psychophys. **57**, 175–182.

Wichmann, F. A., and Hill, N. J. (**2001**). "The psychometric function: I. Fitting, sampling, and goodness of fit," Percept. Psychophys. **63**, 1293–1313.