Hearing Research 344 (2017) 295–303

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Informational masking and the effects of differences in fundamental frequency and fundamental-frequency contour on phonetic integration in a formant ensemble

Robert J. Summers [a], Peter J. Bailey [b], Brian Roberts [a, *]

[a] Psychology, School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK
[b] Department of Psychology, University of York, Heslington, York YO10 5DD, UK

ARTICLE INFO

Article history:
Available online 1 November 2016

Keywords:
Speech intelligibility
Formant integration
Informational masking
Competition
Speech segregation
Target-masker similarity

ABSTRACT

This study explored the effects on speech intelligibility of across-formant differences in fundamental frequency (ΔF0) and F0 contour. Sentence-length speech analogues were presented dichotically (left = F1+F3; right = F2), either alone or—because competition usually reveals grouping cues most clearly—accompanied in the left ear by a competitor for F2 (F2C) that listeners must reject to optimize recognition. F2C was created by inverting the F2 frequency contour. In experiment 1, all left-ear formants shared the same constant F0 and $\Delta F0_{F2}$ was 0 or ±4 semitones. In experiment 2, all left-ear formants shared the natural F0 contour and that for F2 was natural, constant, exaggerated, or inverted. Adding F2C lowered keyword scores, presumably because of informational masking. The results for experiment 1 were complicated by effects associated with the direction of $\Delta F0_{F2}$; this problem was avoided in experiment 2 because all four F0 contours had the same geometric mean frequency. When the target formants were presented alone, scores were relatively high and did not depend on the $F0_{F2}$ contour. F2C impact was greater when F2 had a different F0 contour from the other formants. This effect was a direct consequence of the associated ΔF0; the $F0_{F2}$ contour *per se* did not influence competitor impact.

1. Introduction

When more than one talker is speaking at once, successful communication depends on the ability of the listener to separate the formant ensemble reaching their ears into a figure (target) and background (interferer). There are a number of ways in which the interferer may lower the intelligibility of the target speech; these can be categorized broadly into energetic masking, in which the auditory-nerve response to the target is swamped by the response to the masker, modulation masking, in which masker amplitude variation lowers sensitivity to similar rates of variation in the target (e.g., Stone and Moore, 2014; Stone and Canavan, 2016), and informational masking, which is of central origin and may be considered as encompassing all other forms of interference (e.g., Durlach et al., 2003; Kidd et al., 2008). The study reported here is concerned with informational masking, in which the interference may arise from the disruption of auditory object formation or selection, or from an increase in the cognitive load on the listener (see, e.g., Shinn-Cunningham, 2008; Mattys et al., 2012).

The voices of two talkers speaking at the same time usually differ in fundamental frequency (F0) and in F0 contour; these differences provide acoustic cues for voice segregation that may assist listeners trying to understand what is being said. In the context of the integration of acoustic-phonetic information across formants, it is known that a difference in fundamental frequency (ΔF0) between formants influences their grouping and segregation (Darwin, 1981; Gardner et al., 1989; Bird and Darwin, 1998; Summers et al., 2010). The focus of these studies differs from the many that have explored the effect of ΔF0 on the ability to separate a mixture of two voices within the same ear (e.g., Brokx and Nooteboom, 1982; Binns and Culling, 2007; Deroche et al., 2014) in that performance is limited mainly by the ability to group acoustic elements correctly across frequency regions rather than to separate overlapping harmonics (for a review, see Summers et al., 2010). Studies of the perceptual organization of a formant ensemble indicate that imposing a ΔF0

Abbreviations: CV, consonant-vowel; F0, fundamental frequency; F1, first formant; F2, second formant; F2C, second formant competitor; F3, third formant; % pts, percentage points

[*] Corresponding author.
E-mail address: b.roberts@aston.ac.uk (B. Roberts).

http://dx.doi.org/10.1016/j.heares.2016.10.026
0378-5955/

on one formant in the ensemble can reduce its phonetic contribution to the speech percept, but suggest that this reduction occurs solely or mainly 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) explored the effect of differences in F0 on across-formant grouping and segregation using sentence-length speech analogues and the second-formant competitor (F2C) paradigm (e.g., Remez et al., 1994; Roberts et al., 2010). This paradigm involves the dichotic presentation of two versions of F2, for which 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, a single extraneous formant intended to act as a competitor to F2 (F2C). Hence, the listener must reject the competitor to optimize recognition of the utterance. The version of the F2C paradigm used by Summers et al. (2010) involved presenting the target formants on a monotonic F0 of 150 Hz to separate ears (left ear = F1± F2C+F3; right ear = F2). The inclusion of the competitor lowered intelligibility and applying a ΔF0 to F2C relative to the target formants led to a significant but relatively modest fall in interference, which was attributed to grouping by common F0. The dichotic configuration allowed competition between the two versions of F2 in a context where any interference must have arisen primarily through informational masking. Note that any contribution of energetic masking to competitor impact arising from adding F2C in the same ear as F1+F3 must have been small or negligible for two reasons. First, F1 was lower in frequency and more intense than F2C. Second, competitor impact remained the same when the possibility of upward spread of masking from F2C to F3 was eliminated by moving F3 to the opposite ear (Summers et al., 2010; cf. Rand, 1974).

There are two limitations of the study by Summers et al. (2010) that merit further investigation. First, it did not explore the effect of applying a ΔF0 to the target F2 (rather than to the competitor); second, it did not explore the role of natural F0 contours in the integration of acoustic-phonetic information across formants. The first limitation is important because the target F2 is spatially isolated from the others in the stimulus configuration used and so may be particularly susceptible to perceptual exclusion on the basis of primitive grouping cues (Bregman, 1990; Darwin, 2008). The second limitation is important because the Gestalt principle of good continuation suggests that the smooth and continuous change characteristic of a natural F0 contour might assist in binding together all acoustic elements following that contour, and yet almost no attention has been paid to whether across-formant differences in F0 contour *per se* influence the grouping and segregation of formants. Specifically, are there any direct effects of differences in F0 contour between formants, over and above those arising from the ΔF0 that inevitably results from any mismatch in F0 contour?

To our knowledge, only one experiment has examined the effect of introducing time-varying (as well as static) ΔF0s between formants in an ensemble, in this case one constituting a consonant-vowel (CV) syllable. In their second experiment, Gardner et al. (1989) manipulated a synthetic four-formant ensemble that could be perceived as /ru/ or /li/. When presented alone, formants 1, 2, and 3 elicited /ru/ percepts and formants 1, 3, and 4 elicited /li/ percepts. When all four formants were presented together on the same F0, almost all responses indicated /ru/ percepts. However, when a ΔF0 was applied to formant 2, the syllable could be heard as /ru/ or /li/ (or as both) depending on whether or not the phonetic information carried by formant 2 was integrated into the percept. In addition to static F0 differences between formant 2 and the rest, the effects of coherent and incoherent sinusoidal modulation of F0 between the two sets were compared (rate = 6 Hz or 12 Hz; depth = ±3% or ±8%; phase difference = 0° or 90°). There was no evidence that the coherence of the motion of F0 had any additional effect on the perceptual grouping of the formants over and above the effect of a static ΔF0. Nonetheless, it would be premature to generalize from this finding obtained for synthetic CV syllables and sinusoidal F0 contours and to assume that there is no additional role for F0 contour in the grouping and segregation of formants for sentence-length utterances synthesized using the natural pattern of F0 variation.

Investigations of the influence of variations in voice pitch on speech intelligibility have generally been restricted to cases where all the formants share the same F0 contour. A number of studies have shown that changing the F0 contour from the natural pattern of variation usually lowers the intelligibility of sentence-length utterances. Such effects have been found even when high-quality speech is heard in quiet, but the impact of such change tends to become more pronounced in more adverse listening conditions, such as low-pass filtering (Hillenbrand, 2003) or the presence of background noise (Miller et al., 2010) or a competing talker (Binns and Culling, 2007). The most common manipulation is to flatten the F0 contour to a monotone, removing any prosodic information carried by the natural pattern of F0 variation (Wingfield et al., 1984; Laures and Weismer, 1999; Hillenbrand, 2003; Binns and Culling, 2007; Miller et al., 2010; Deroche et al., 2014). Under otherwise similar listening conditions, the impact on intelligibility is greater when the prosodic information provided by F0 variation is not simply removed but is instead made misleading by inverting the natural pattern of variation. For example, Miller et al. (2010) found that flattening the F0 contour of speech presented in noise lowered keyword scores by ~13 percentage points (% pts) relative to the natural contour, whereas inverting the F0 contour lowered performance by ~23% pts. Their study also included a condition in which the natural F0 variation was exaggerated by × 1.75, for which the effect was similar to flattening the contour (~13% pts reduction). Presumably, exaggeration had less effect than inversion because the variations were in the same direction moment-to-moment as for the natural contour. In contrast to these studies, which were designed primarily to explore the prosodic properties of F0 contours, the current study used natural, constant, exaggerated, and inverted contours to introduce time-varying differences in F0 between one formant (F2) and the others.

The two experiments reported here addressed the limitations of Summers et al. (2010) by comparing the effects of applying differences either in constant F0 or in F0 contour to the target F2, in the presence and absence of F2C. When F2C was present, its F0 contour always matched that of F1+F3. For this stimulus configuration, note that there are two grouping cues (ear of presentation and common F0) favouring the fusion of the extraneous formant with the other target formants. Whilst the primary goal of this study was to use speech acoustics to extend our understanding of the role of F0 as an auditory grouping cue, these experiments also cast further light on the nature of acoustic-phonetic integration in speech perception.

2. Experiment 1

In this experiment, the F0 of F2 ($F0_{F2}$) could be the same as, or different from, that of the other formants. The purpose of the experiment was to measure the extent to which the intelligibility of dichotic target speech (F1+F3; F2) was dependent on the difference in F0 between the isolated target F2 and the other formants, in the presence and absence of a competitor (F2C) that shared a common F0 and ear of presentation with the F1+F3 “frame”. Note that the presence of the competitor is challenging for the listener, as maximizing intelligibility involves discarding the acoustic-phonetic information carried by a misarticulated but seemingly genuine second formant that accompanies F1 and F3. This experiment used monotonous F0 contours throughout.

2.1. Method

2.1.1. 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 Procedure). All listeners completed the training successfully, but five did not meet the additional criterion of a mean score of ≥20% keywords correct in the main experiment when collapsed across conditions, and so were subsequently replaced. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. Twenty-four listeners (three males) successfully completed the experiment (mean age = 20.5 years, range = 18.4–43.8). All of these listeners had previously taken part in at least one speech perception experiment in our laboratory but, to our knowledge, none of them had heard any of the sentences used in the main 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.

2.1.2. Stimuli and conditions

The stimuli for the main experiment were derived from recordings of 48 sentences with almost continuous voicing, spoken by a British male talker of “Received Pronunciation” English. Speech with almost continuous voicing was used to optimize the measurement of the effect of ΔF0 on across-formant grouping (Bird and Darwin, 1998; Summers et al., 2010). The sentences used were taken from two sources (Binns and Culling, 2007; Bird and Darwin, 1998). Given the requirement for almost continuously voiced speech, most of the sentences were semantically unusual (e.g., “Moles are lowly rural vermin” and “The new royals rule evilly over the realm”). A set of keywords was chosen for each sentence; most designated keywords were content words. The stimuli for the training session (see Procedure) were derived from 50 sentences spoken by a different talker and taken from commercially available recordings of the Harvard sentence lists (IEEE, 1969). Each of the selected sentences contained ≤25% phonemes involving closures or unvoiced frication.

For each sentence, the F0 contour and 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 F0 contour and the three formant frequencies were hand-corrected using a graphics tablet; artifacts are not uncommon and manual post-processing 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.

The frequency and amplitude contours of the target formants were used to control three parallel buzz-excited second-order resonators. The type of excitation source used was a monotonous periodic train of pulses modeled on the glottal waveform, shown by Rosenberg (1971) to be capable of producing synthetic speech of good quality. When a competitor was required, the frequency and amplitude contours of F2C (see below) were used to control a fourth resonator receiving the same excitation source as the F1+F3 frame. The 3-dB bandwidths of the resonators corresponding to F1, F2/F2C, and F3 were set to constant values of 50, 70, and 90 Hz, respectively. All stimuli were presented in a dichotic configuration (F1±F2C+F3; F2). Following Klatt (1980), the outputs of the resonators corresponding to F1, F2C, and F3 were summed using alternating signs (+, −, +) to minimize spectral notches between adjacent formants in the same ear; for consistency, the same output sign (−) was used for the isolated target F2. For each sentence, the competitor was generated using a frequency contour created by inverting that of the corresponding target F2 on a log scale; this manipulation preserves the rate and depth of frequency variation found in F2 but changes its pattern. The amplitude contour used was the same as for the target F2. When present, F2C was always delivered in the same ear as the F1+F3 frame. Stimuli were selected such that the centre frequency of F2C was always ≥80 Hz from F1 and F3.

The F0 of the excitation source used to generate the F1+F3 frame and F2C was set to a constant value of 140 Hz; this value is similar to the mean F0 for the set of utterances spoken by this talker. Three versions of the target F2 were created by setting the excitation source to one of the following constant values: $F0_{F2}$ = 111.1, 140.0, or 176.4 Hz. These values were chosen to create a set of ΔF0s for F2 relative to the other formants of −4, 0, and 4 semitones, respectively. The RMS power of the ±4-semitone versions was set to the same value as for the matched-F0 version (i.e., the 0-semitone case). There were eight conditions in the main experiment (see Table 1). C1 and C2 were the F2-absent conditions. The stimuli for C1 comprised the F1+F3 frame alone; C2 differed only in that F2C was also present. The stimuli for C3-C5 comprised the competitor plus all three target formants, corresponding to ΔF0s on F2 of −4, 0, and 4 semitones, respectively. This range of mistuning was chosen to be large enough to provide a clear grouping cue (cf. Gardner et al., 1989) whilst limiting to some degree the opportunity for the effects of differences in absolute F0 to be manifested. The stimuli for the remaining conditions (C6-C8) differed only in that the target formants were unaccompanied. The 48 sentences were divided equally across conditions (i.e., six per condition), such that there were always 30 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.

2.1.3. 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 40–50 min to complete; listeners were free to take a break whenever they wished. In both parts of the experiment, stimuli were presented in a new quasi-random order for each listener.

The training session comprised 50 trials; stimuli were presented diotically, without competitors, and a new sentence was used for each trial. All were synthesized as three-formant analogues on their natural F0 contours. On each of the first ten trials, listeners heard presentations of the synthetic version (S) and the original (clear, C) recording (44.1 kHz sample rate) 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 presentation of the synthetic version of a sentence, which they were asked to transcribe using the keyboard. They were allowed to listen to the stimulus up to six times before typing in their transcription. After each transcription was entered, feedback

Table 1
Stimulus properties for the conditions used in experiment 1 (main session). The F0 frequency for F1, F2C, and F3 was always 140 Hz. The ΔF0 on the target F2 is relative to 140 Hz.

Condition	Stimulus configuration (left ear; right ear)	ΔF0 on target F2 (semitones)
C1	(F1+F3; −)	–
C2	(F1+F2C+F3; −)	–
C3	(F1+F2C+F3; F2)	−4
C4	(F1+F2C+F3; F2)	0
C5	(F1+F2C+F3; F2)	4
C6	(F1+F3; F2)	−4
C7	(F1+F3; F2)	0
C8	(F1+F3; F2)	4

was provided by playing the original recording followed by a repeat of the synthetic version. The same approach was used for the final ten training trials, except that the number of listens allowed was reduced to three. Davis et al. (2005) found the strategy of providing feedback by alternating presentations of the synthetic and original versions to be an efficient way of enhancing the perceptual learning of speech analogues. In the main experiment, listeners were again allowed to hear each stimulus up to three times before entering their transcription, but 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 & Kjaer, type 2209; Nærum, Denmark) coupled to the earphones by an artificial ear (type 4153). Stimuli were presented at a long-term reference level of 75 dB SPL; this describes the case where the left ear receives F1+F3. Given that F1 is far more intense than the higher formants, the presence or absence of F2C had little effect on the presentation level in the left ear. On average, the presentation level in the right ear (receiving F2) was ~10 dB lower. Owing to the use of diotic materials, the presentation level in the training session was lowered to 72 dB SPL, roughly to offset the increased loudness arising from binaural summation.

2.1.4. Data analysis

For each listener, the intelligibility of each stimulus was quantified in terms of the percentage of keywords identified correctly; obvious misspellings were corrected and homonyms were accepted. The stimuli for each condition comprised six sentences. Given the variable number of keywords per sentence (4–7), 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, 2016), our principal measure involved classifying responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly. We used loose scoring as an additional measure, in which a response is scored as correct if the stem of the word is reported accurately – e.g., “type”, “types”, and “typed” would all be marked as correct for the keyword “typing” (see Foster et al., 1993). All values quoted are based on the tight scores unless otherwise stated. All statistical analyses reported here were computed using SPSS (SPSS statistics version 20, IBM Corp.). Given the low scores obtained for the control conditions, analysis of variance (ANOVA) was conducted using arcsine-transformed data ($Y' = 2 \arcsin(\sqrt{Y})$, where Y is the proportion correct score; see Keppel and Wickens, 2004); the measure of effect size reported is partial eta squared (η^2_p). Paired-samples comparisons (two-tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967).

2.2. Results and discussion

Fig. 1 shows the mean percentage scores (and inter-subject standard errors) across conditions for keyword identification. The black, grey, and white bars indicate the results for the frame±F2C (control), target-plus-competitor, and target-only conditions, respectively. Note that the relatively modest intelligibility of the target-only cases compared with natural speech is to be expected given the simple three-formant parallel vocal-tract model used to synthesize the sentences, the dichotic presentation of the target formants, and the semantically unusual nature of the sentences. A one-way within-subjects ANOVA across all conditions showed a highly significant effect of condition on intelligibility [$F(7,161) = 37.54$, $p < 0.001$, $\eta^2{}_p = 0.620$]. The control conditions show that intelligibility was low for the F1+F3 frame alone (C1) and near floor when the competitor was added in the absence of the target F2 (C2). Pairwise comparisons indicate that the mean score for C1 differed from those for all other conditions (range: $p = 0.017$ – $p < 0.001$) except C5 ($p = 0.205$). The mean score for C2 differed from those for all other conditions ($p < 0.001$, in all cases); the significant difference between C1 and C2 indicates that the addition of F2C tended to reduce further the limited intelligibility supported

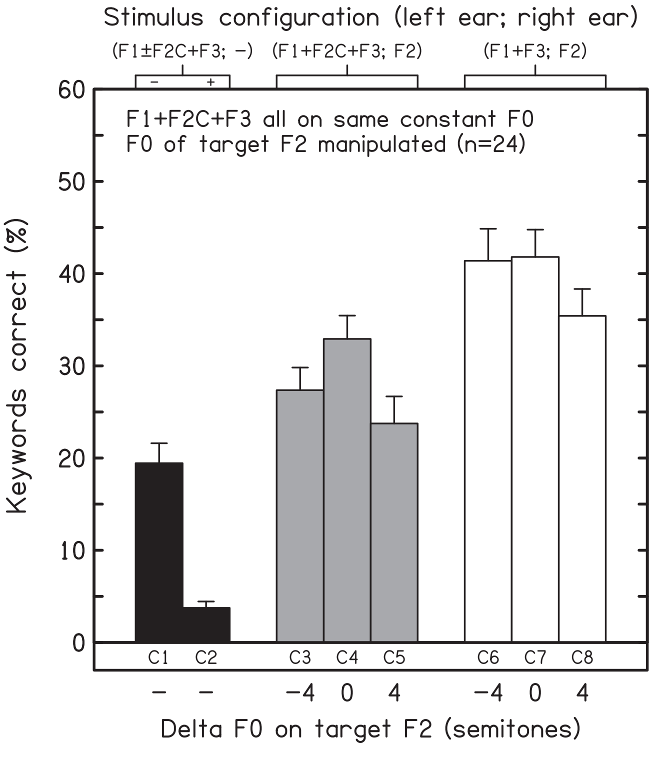

Fig. 1. Results for experiment 1 – effect of applying a ΔF0 to the target F2 on the intelligibility of analogues of sentences spoken with almost continuous voicing in the presence and absence of a competitor (F2C). Mean keyword scores and inter-subject standard errors (n = 24) are shown for tight scoring in the F2-absent conditions (black bars), the target-plus-competitor conditions (grey bars), and the target-only conditions (white bars). The top axis indicates which formants were presented to each ear; the bottom axis indicates the ΔF0 for F2 (when present). For ease of reference, condition numbers are included above the bottom axis. The corresponding means for loose scoring across conditions are 21.1% (C1), 3.9% (C2), 30.0% (C3), 34.4% (C4), 24.7% (C5), 43.8% (C6), 43.6% (C7), and 38.1% (C8).

by F1+F3 alone.

It was predicted that performance would be best when the three target formants were presented alone and that imposing a ΔF0 on F2 would lower intelligibility solely (or mainly) when the competitor was present. Visual inspection of Fig. 1 appears to support the first prediction, but not necessarily the second. The effects of adding a competitor to the target speech (F2C = present or absent) and of introducing a difference in F0 between the target F2 and the other formants (ΔF0 = −4, 0, or 4 semitones) were explored using a two-way within-subjects ANOVA restricted to the experimental conditions (C3-C8). This analysis revealed a significant main effect on keyword scores of adding a competitor [mean difference = 11.5% pts; $F(1,23) = 22.89$, $p < 0.001$, $\eta^2_p = 0.499$]. Intelligibility was lowered when the target formants were accompanied by an F2C created using the inverted F2 frequency contour, presumably as a result of informational masking. There was also a significant main effect of $F0_{F2}$ [mean differences for cases ΔF0 = 0 vs. −4 and 0 vs. 4 semitones = 3.0 and 7.8% pts, respectively; $F(2,46) = 5.801$, $p = 0.006$, $\eta^2_p = 0.201$]; pairwise comparisons showed that the fall in keyword scores arising from a 4-semitone rise in $F0_{F2}$ was significant ($p = 0.001$) but that the effect of a 4-semitone fall in $F0_{F2}$ was not ($p = 0.235$). Despite the suggestion of an asymmetry between the effects of raising and lowering $F0_{F2}$, a pairwise comparison showed that this difference did not quite reach significance (ΔF0 = −4 vs. 4 semitones; $p = 0.067$).

When expressed as a difference score with respect to the corresponding target-only case, F2C impact for the three versions of the target F2 (ΔF0 = −4, 0, or 4 semitones) was 14.0, 8.9, and 11.6% pts, respectively. Although this pattern is consistent with the notion that competition accentuates the effect of a mismatch in F0 between F2 and F1+F3, the two factors in the ANOVA did not interact [$F(2,46) = 0.539$, $p = 0.587$]. The outcomes for the supplementary analyses (loose scores) were fully consistent with those for the main analyses.

The results for the conditions in which the three target formants were accompanied by a competitor (C3-C5) are complementary to those of the analogous conditions in the study reported by Summers et al. (2010). In that study, the impact of F2C *decreased* as its F0 was mistuned relative to that of the target formants, whereas here the impact of F2C *increased* as the F0 of the target F2 was mistuned relative to that of the other formants. Note that, in addition to the primary effect of mistuning, there is a suggestion of an asymmetry in both studies arising from a secondary effect of absolute F0. In our earlier study, competitor impact was greater when F2C was mistuned upwards and so had a higher F0 than the target formants (Summers et al., 2010). It was suggested that this was due to the progressive change in the excitation of F2C towards fewer, more intense, and better-resolved harmonics as F0 was increased. To the extent that an asymmetry is apparent here, it is in the opposite direction − i.e., intelligibility falls more when the target F2 is mistuned upwards and so has a higher F0 than the other formants. This effect may arise because the precision of the representation of the target F2 frequency declines as the harmonics exciting F2 become sparser. Deroche et al. (2014) reported an asymmetry in the same direction for diotic presentation of a target voice and a speech-shaped harmonic complex or babble. Specifically, the masker was considerably less effective when its F0 was 11 semitones above that of the target F0 than when its F0 was 11 semitones below. Therefore, before rejecting the idea that competition accentuates the effect of F0 differences between formants, it would be prudent to explore the consequences of applying a ΔF0 on F2 without introducing a difference in mean F0 between F2 and the other formants. This is possible to achieve only if time-varying F0 contours are used.

3. Experiment 2

In this experiment, the F0 contour extracted from the natural utterance was used to generate all the formants received by the left ear (F1±F2C+F3); the F0 contour for the contralateral F2 could be the same as, or different from, that of the other formants. The purpose of this experiment was twofold. First, it was to measure the extent to which the intelligibility of dichotic target speech (F1+F3; F2) was dependent on the difference in F0 between the isolated target F2 and the other formants, in the presence and absence of a competitor, under circumstances where mismatches in F0 contour did not lead to differences in mean F0. Second, it was to assess whether or not differences in F0 contour between F2 and the other formants influenced target intelligibility directly, rather than indirectly through the associated ΔF0 that inevitably arises when the two contours are mismatched.

3.1. Method

Except where described, the same method was used as for experiment 1. Forty listeners (ten males) passed the training and successfully completed the experiment (mean age = 28.6 years, range = 18.5−52.4); this includes three replacements for listeners who did not meet the additional criterion of an overall mean score of ≥20% keywords correct in the main session. The stimuli for the main experiment were derived from a set of 60 sentences spoken with almost continuous voicing; these sentences were divided equally across conditions such that there were always 29 or 30 keywords per condition. Given that the sentences overlapped with those used in experiment 1, a different set of listeners took part in this experiment. The procedure used differed from that used in experiment 1 in only one respect − during the final ten training trials and the main experiment, listeners were allowed to hear each stimulus only once before entering their transcription. This change was made in response to pilot work demonstrating higher overall intelligibility of materials synthesized using their natural F0 contours. Together, the training session and main experiment took about 50−60 min.

As before, F1+F3 and (when present) F2C always shared the same F0, but here this was the natural F0 contour as extracted from each original recording by Praat (with manual post-processing where necessary). In this experiment, $F0_{F2}$ was set to one of the following contours: natural (N), constant at the geometric mean frequency of the natural F0 contour (C), twice the depth of variation about the geometric mean F0 (exaggerated, E), or inverted (I) in which the sign of the variation of the F0 contour about its geometric mean was reversed. The RMS power of the mismatched-F0 versions was set equal to that of the matched-F0 version (i.e., the natural contour). An example stimulus is illustrated in Fig. 2, showing spectrograms and associated F0 contours for F1+F2C+F3 (left ear, natural F0 contour) and for the four versions of the target F2 (right ear, $F0_{F2}$ = N, C, E, or I). Averaged across all 60 sentences, the mean absolute ΔF0 between the other formants (F0 = N) and the target F2 was: 0 ($F0_{F2}$ = N), 1.65 ($F0_{F2}$ = C or E), and 3.29 semitones ($F0_{F2}$ = I), respectively. The choice of a scaling factor of × 2 for the exaggerated contour (rather than × 1.75, as used by Miller et al., 2010) ensured that the mean absolute ΔF0 between N and E was the same as between N and C. Note that all four F0 contours had the same mean value, which allowed the effects of ΔF0 to be isolated from those of absolute F0. There were ten conditions in the main session (see Table 2). Hence, the total number of listeners required to produce a balanced dataset was a multiple of ten. As before, C1 and C2 were the F2-absent control conditions. The stimuli for C3-C6 comprised the frame-plus-competitor accompanied by each of the four versions of the target F2; the stimuli for the remaining

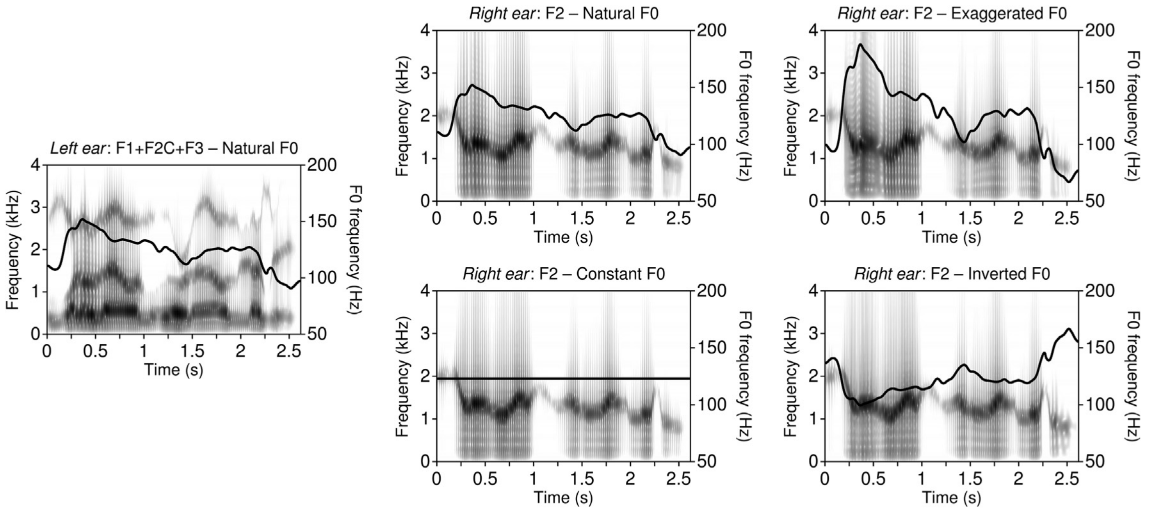

Fig. 2. Stimuli for experiment 2 – dichotic stimulus configuration, illustrated using wideband spectrograms and F0 contours (solid black lines) for analogues of the example sentence “The yellow lion wore an iron muzzle.” The left ear received F1, F2C, and F3 (left panel); the right ear received one of four variants of the target F2 (other panels). The left-ear formants always shared the natural F0 contour; the F0 contour of the target F2 could be natural (N, upper centre panel), exaggerated (E, upper right panel), constant (C, lower centre panel), or inverted (I, lower right panel). The formant-frequency contour of F2C (when present) was inverted about the geometric mean frequency with respect to the target F2.

conditions (C7-C10) differed only in that the competitor was absent.

3.2. Results and discussion

Fig. 3 shows the mean percentage scores (and inter-subject standard errors) across conditions for keyword identification. The black, grey, and white bars indicate the results for the frame±F2C (control), target-plus-competitor, and target-only conditions, respectively. A one-way ANOVA across all conditions showed a highly significant effect of condition on intelligibility [F(9,351) = 88.01, $p < 0.001$, $\eta^2_p = 0.693$]. As before, the control conditions show that intelligibility was modest for the F1+F3 frame alone (C1) and near floor when the competitor was added in the absence of the target F2 (C2). Pairwise comparisons indicate that the mean scores for C1 and C2 differed from those for all conditions ($p < 0.001$), including each other, except for C1 vs. C3 ($p = 0.069$).

It was predicted that performance would be best when the three target formants were presented alone and that imposing a different F0 contour on F2 would lower intelligibility solely (or mainly) when the competitor was present. Visual inspection of Fig. 3 appears to support both predictions. In particular, there was little or no intelligibility cost of applying a different F0 contour on F2 in the

Table 2
Stimulus properties for the conditions used in experiment 2 (main session). The F0 contour for F1, F2C, and F3 was always the natural contour (N). The F0 contour for the target F2 could be inverted (I), constant (C), exaggerated (E), or natural (N).

Condition	Stimulus configuration (left ear; right ear)	F0 contour of target F2
C1	(F1+F3; −)	−
C2	(F1+F2C+F3; −)	−
C3	(F1+F2C+F3; F2)	I
C4	(F1+F2C+F3; F2)	C
C5	(F1+F2C+F3; F2)	E
C6	(F1+F2C+F3; F2)	N
C7	(F1+F3; F2)	I
C8	(F1+F3; F2)	C
C9	(F1+F3; F2)	E
C10	(F1+F3; F2)	N

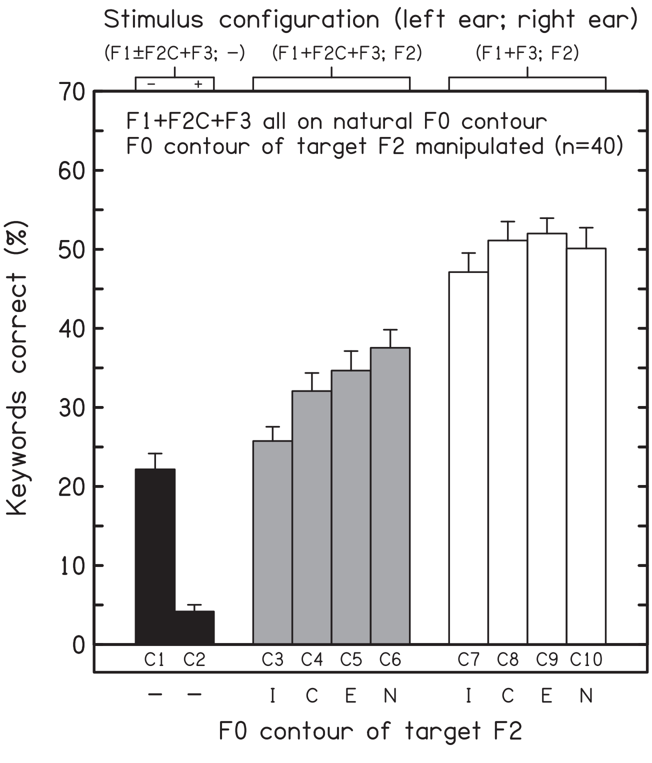

Fig. 3. Results for experiment 2 – effect of applying different F0 contours to the target F2 on the intelligibility of analogues of sentences spoken with almost continuous voicing in the presence and absence of a competitor (F2C). Mean keyword scores and inter-subject standard errors ($n = 40$) are shown for tight scoring in the F2-absent conditions (black bars), the target-plus-competitor conditions (grey bars), and the target-only conditions (white bars). The top axis indicates which formants were presented to each ear; the bottom axis indicates the F0 contour for F2 (when present). The F0 contour for F2 could be inverted (I), constant (C), exaggerated (E), or natural (N); The F0 contour for F1, F2C, and F3 was always the natural contour (N). For ease of reference, condition numbers are included above the bottom axis. The corresponding means for loose scoring across conditions are 23.8% (C1), 4.8% (C2), 27.3% (C3), 34.1% (C4), 36.6% (C5), 39.6% (C6), 50.2% (C7), 53.6% (C8), 54.0% (C9), and 52.8% (C10).

absence of F2C, despite the consequent ΔF0 between F2 and the F1+F3 frame. The effects of adding a competitor to the target speech (F2C = present or absent) and of introducing a difference in F0 contour between F2 and the other formants – F0 contour for F2 = natural (N), exaggerated (E), constant (C), or inverted (I) – were explored using a two-way within-subjects ANOVA restricted to the experimental conditions (C3-C10). This analysis revealed a significant main effect on keyword scores of adding a competitor, presumably arising from informational masking [mean difference = 17.0% pts; $F(1,39) = 141.36$, $p < 0.001$, $\eta^2_p = 0.784$]. There was also a significant main effect of $F0_{F2}$ contour [$F(3,117) = 5.030$, $p = 0.003$, $\eta^2_p = 0.114$]; pairwise comparisons showed significant differences for cases N vs. I ($p = 0.004$), I vs. C ($p = 0.036$), and I vs. E ($p = 0.005$).

When expressed as a difference score with respect to the corresponding target-only case, competitor impact for the $F0_{F2}$ contours tested was 21.5 (I), 18.5 (E), 17.3 (C), and 10.7% pts (N), respectively. This pattern, in which the impact of F2C on intelligibility was about twice as large for the $F0_{F2}$ = I vs. N cases, is in accord with the notion that competition caused or accentuated the effect of a mismatch in F0 contour between F2 and F1+F3. Although the interaction between the two factors narrowly missed significance in the main analysis using tight scoring [$F(3,117) = 2.666$, $p = 0.051$], the outcome for the supplementary analysis using loose scoring was significant [$F(3,117) = 3.123$, $p = 0.029$, $\eta^2_p = 0.074$]. Note also that the need to rotate the allocation of sentences across conditions in this design inevitably increased the extent of uncontrolled variance and therefore reduced overall sensitivity. On balance, it seems reasonable to conclude that the effect of introducing a difference in F0 contour between F2 and the other formants was greater when the F1+F3 frame was accompanied by an F0-matched competitor.

The other aim of this experiment was to explore whether introducing a difference in F0 contour between F2 and the other formants had any effects of intelligibility other than those arising from the associated ΔF0. There are two ways in which such effects might arise. First, it is already established that the overall intelligibility of a synthetic or resynthesized sentence is typically highest when it is generated using the F0 contour extracted from the corresponding natural utterance (e.g., Laures and Weismer, 1999; Binns and Culling, 2007; Miller et al., 2010). Hence, changing the $F0_{F2}$ contour from natural variation to any of the other versions might contribute directly to a fall in intelligibility, in addition to any indirect effect of the associated ΔF0. Any effect of this kind should occur whether or not F2C is present, which is not consistent with the pattern observed here. Indeed, although the mean keyword score in the absence of F2C was nominally lowest for the inverted $F0_{F2}$ contour (C7), it was only 3.0% pts lower than for the natural $F0_{F2}$ contour (C10); the corresponding difference in the presence of F2C was much larger (C3 vs. C6 = 11.8% pts).

Second, using an alternative to the natural $F0_{F2}$ contour might directly disadvantage the integration of the phonetic information carried by F2 in the presence of a competitor sharing the natural F0 contour with F1+F3. The signature of an effect of this kind would be a greater impact of competition from F2C than expected based on the ΔF0 arising from the mismatch between the two F0 contours. In contrast with this hypothesis, the results obtained can be accounted for purely in terms of ΔF0; there was no indication that the shape of the $F0_{F2}$ contour *per se* affected F2C impact. This outcome is consistent with the findings of Gardner et al. (1989) using the /ru/-/li/ paradigm, for which the coherence across formants of sinusoidal frequency modulation (FM) applied to F0 had no additional effect to that of ΔF0. Our basis for drawing this conclusion is as follows.

The constant and exaggerated cases shared a mean ΔF0 of 1.65 semitones with respect to the other formants, but these two cases arguably differ in the plausibility of their F0 contours in that the former is simply a scaled-up version of the natural variation, whereas the production of a constant F0 by a human talker is entirely implausible. Hence, one might have expected greater competition in the constant case, but in fact the mean difference scores (±F2C) for the constant and exaggerated cases were similar and the nominal difference (1.2% pts) was in the wrong direction. It should be acknowledged, however, that Miller et al. (2010) observed an equal fall in intelligibility when the natural F0 contour of sentence-length utterances was either flattened or exaggerated, which casts some doubt on the notion that plausibility is an important factor here. Nonetheless, there is clear evidence that the actively misleading prosodic information provided by the anticorrelated changes of an inverted F0 contour impairs intelligibility (Binns and Culling, 2007; Miller et al., 2010). It merits note, therefore, that changing the $F0_{F2}$ contour here from exaggerated to inverted only increased the mean difference score (±F2C) by a further 3.0% pts, despite the inverted case having arguably the least natural $F0_{F2}$ contour and having an associated mean ΔF0 twice as large (3.29 semitones). This outcome is broadly in accord with the effect of static F0 differences between formants in the /ru/-/li/ paradigm (Gardner et al., 1989). In that study (for which, as here, overlap of corresponding harmonics between different formants was not a major factor), about two thirds of the maximum available benefit for segregation was achieved once ΔF0 reached 2 semitones. Interpolating from their results across the range of ΔF0s they tested, one would expect relatively little extra benefit of doubling ΔF0 from 1.65 to 3.29 semitones.

4. General discussion

The results of the experiments reported here suggest that, once the influence of absolute F0 is controlled, the integration of acoustic-phonetic information across formants and ears is largely unaffected by a difference in F0 between F2 and F1+F3, unless F2C is also present. This integration occurs despite the fact that listeners typically hear more than one source under these circumstances (Broadbent and Ladefoged, 1957; Cutting, 1976). Keyword intelligibility is lowered when the target formants are accompanied by a competitor in the same ear and on the same F0 as F1+F3; the magnitude of this fall increases when the target F2 is mistuned relative to the other formants. This outcome is consistent with the notion that the factors influencing grouping and segregation are best revealed when there is competition between different perceptual organizations (Barker and Cooke, 1999). Nonetheless, it should be acknowledged that – unlike the effect of ΔF0 for a mixture of two voices under diotic presentation (e.g., Brokx and Nooteboom, 1982) – the effect of ΔF0 on the grouping of formants across frequency and ears is relatively modest, and the interaction with the presence of F2C is fairly marginal. In this regard, it is interesting to compare this outcome with the results of the /ru/-/li/ studies (Darwin, 1981; Gardner et al., 1989). In those studies, applying a ΔF0 to formant 2 in the ensemble reduced its phonetic contribution to the syllable. This effect was manifest as a change from almost all /ru/ responses to progressively more /li/ responses as the degree of mistuning was increased. However, this change plateaued at roughly half /li/ responses for ΔF0s above five semitones, indicating incomplete perceptual exclusion of formant 2 even for large ΔF0s.

Although it is well established that the F0 contour of a sentence can affect its overall intelligibility (e.g., Binns and Culling, 2007; Miller et al., 2010), the experiments reported here are consistent with the proposal that any effects of differences between formants in F0 contour are indirect, arising from the consequent ΔF0. There was no evidence to suggest that sharing a smooth and continuous F0 contour of a particular shape had any direct effect on across-formant grouping and segregation. Future research might explore whether this is also true for diotic mixtures of two voices, using a variant of the method described by Bird and Darwin (1998, experiment 2). In that experiment, each sentence was low- and high-pass filtered at 800 Hz, separating F1 from the higher formants. In the swapped-F0 condition, filtered stimuli were recombined so that the low-pass region of the target sentence shared the same F0 as the high-pass region of the interfering sentence, and the high-pass region of the target shared the same F0 as the low-pass region of the interferer (cf. Culling and Darwin, 1993). This manipulation impaired across-frequency grouping mechanisms by cueing inappropriate pairings of the F1 region with the region encompassing the higher formants. Bird and Darwin (1998) only used differences in constant F0 in the swapped-F0 condition, but in principle this approach could be extended to differences in time-varying F0 contour.

It is well established that informational masking can be reduced by primitive grouping cues (Bregman, 1990; Darwin, 2008) for the segregation of target and masker (e.g., Kidd et al., 1994). In particular, studies on the informational masking of non-speech stimuli indicate an important role for target-masker similarity in determining the extent of interference (e.g., Neff, 1995; Lee and Richards, 2011). The experiments reported here indicate that differences in F0 between one target formant and the others can influence their grouping and segregation, when accompanied by a competitor, in a manner consistent with grouping by target-masker similarity. This result contrasts with that obtained in recent studies involving more radical differences in source type (Roberts et al., 2015; Summers et al., 2016). Specifically, if some formants are rendered as tonal (sine-wave source; Bailey et al., 1977; Remez et al., 1981) and others as harmonic analogues (buzz source, as here), the tonal analogues always lose out to the harmonic analogues under competitive conditions, regardless of target-masker similarity.

What might be the basis for this difference in outcomes? Although changes in source characteristics between harmonic and tonal can have a large impact on intelligibility under competitive conditions, it has been proposed that these effects arise from differences in the effectiveness of carrying acoustic-phonetic information rather than from the extent of target-masker similarity (Roberts et al., 2015; Summers et al., 2016). It was conjectured that this difference in transmission efficiency was a consequence either of differences in bandwidth or in naturalness between harmonic and tonal analogues. Here, where the source types used were so similar, such differences in transmission efficiency are likely to have been small or absent, allowing any effect of target-masker similarity to be manifest. Hence, if we compare the current results for dichotic targets with those of Summers et al. (2010), the effect on intelligibility of applying ΔF0 to F2 or F2C appears to be broadly symmetrical (fall or rise), at least to the extent that it is possible to partial out the influence of absolute F0.

In conclusion, F0 differences between formants in an ensemble can influence the integration of acoustic-phonetic information across frequency and ears for sentence-length materials. However, these effects on grouping are often quite modest compared with others attributed to F0 cues (e.g., Bregman, 1990; Micheyl and Oxenham, 2010) and are usually apparent only when there is competition between alternative perceptual organizations. The effects of ΔF0 observed under competition are consistent with the notion that target-masker similarity governs the extent of interference. There is no evidence of any direct effect of differences in F0 contour between formants on the integration of acoustic-phonetic information, over and above the effect of the consequent ΔF0. It remains to be established whether or not this is also the case for diotic mixtures of two voices.

Acknowledgement

Correspondence concerning this article should be addressed to Brian Roberts, Psychology, School of Life and Health Sciences, Aston University, Birmingham, B4 7ET, UK, Email: b.roberts@aston.ac.uk, ORCID: 0000-0002-4232-9459. To access the research data underlying this publication, see http://dx.doi.org/10.17036/030af3e1-064c-4b80-b478-f17fa0e64842. This research was supported by Research Grant ES/K004905/1 from the Economic and Social Research Council (UK), awarded to Brian Roberts. We are grateful to Quentin Summerfield for enunciating the test sentences. Poster presentations on this research have been given at the Basic Auditory Science Meeting of the British Society of Audiology (Cardiff, UK, September 2015), and the 171st Meeting of the Acoustical Society of America (Salt Lake City, UT, May 2016).

References

Bailey, P.J., Summerfield, Q., Dorman, M., 1977. On the Identification of Sine-wave Analogues of Certain Speech Sounds. Haskins Lab. Status Rep. Speech Res. SR-51/52, pp. 1–25.
Barker, J., Cooke, M., 1999. Is the sine-wave speech cocktail party worth attending? Speech Commun. 27, 159–174.
Binns, C., Culling, J.F., 2007. The role of fundamental frequency contours in the perception of speech against interfering speech. J. Acoust. Soc. Am. 122, 1765–1776.
Bird, J., Darwin, C.J., 1998. Effects of a difference in fundamental frequency in separating two sentences. In: Palmer, A.R., Rees, A., Summerfield, A.Q., Meddis, R. (Eds.), Psychophysical and Physiological Advances in Hearing. Whurr, London, pp. 263–269.
Boersma, P., Weenink, D., 2010. PRAAT, a System for Doing Phonetics by Computer, Software Package, Version 5.1.28. Institute of Phonetic Sciences, University of Amsterdam, The Netherlands. Retrieved from. http://www.praat.org/.
Bregman, A.S., 1990. Auditory Scene Analysis: the Perceptual Organization of Sound. MIT Press, Cambridge, MA.
Broadbent, D.E., Ladefoged, P., 1957. On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Am. 29, 708–710.
Brokx, J.P.L., Nooteboom, S.G., 1982. Intonation and the perceptual separation of simultaneous voices. J. Phon. 10, 23–36.
Culling, J.F., Darwin, C.J., 1993. Perceptual separation of simultaneous vowels: within and across-formant grouping by F0. J. Acoust. Soc. Am. 93, 3454–3467.
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. Phil. Trans. R. Soc. B: Biol. Sci. 363, 1011–1021.
Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., 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.
Deroche, M.L.D., Culling, J.F., Chatterjee, M., Limb, C.J., 2014. Roles of the target and masker fundamental frequencies in voice segregation. J. Acoust. Soc. Am. 136, 1225–1236.
Durlach, N.I., Mason, C.R., Kidd, G., Arbogast, T.L., Colburn, H.S., Shinn-Cunningham, B.G., 2003. Note on informational masking. J. Acoust. Soc. Am. 113, 2984–2987.
Foster, J.R., Summerfield, A.Q., Marshall, D.H., Palmer, L., Ball, V., Rosen, S., 1993. Lip-reading the BKB sentence lists: corrections for list and practice effects. Br. J. Audiol. 27, 233–246.
Gardner, R.B., Gaskill, S.A., Darwin, C.J., 1989. Perceptual grouping of formants with static and dynamic differences in fundamental frequency. J. Acoust. Soc. Am. 85, 1329–1337.
Henke, W.L., 2005. MITSYN: a Coherent Family of High-level Languages for Time Signal Processing, Software Package. Belmont, MA.
Hillenbrand, J.M., 2003. Some effects of intonation contour on sentence intelligibility. J. Acoust. Soc. Am. 114, 2338 (abstract).
Institute of Electrical and Electronics Engineers (IEEE), 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust 225–246. AU-17.
Keppel, G., Wickens, T.D., 2004. Design and Analysis: a Researcher's Handbook, fourth ed. Pearson Prentice Hall, Englewood Cliffs, NJ.
Kidd, G., Mason, C.R., Deliwala, P.S., Woods, W.S., Colburn, H.S., 1994. Reducing informational masking by sound segregation. J. Acoust. Soc. Am. 95, 3475–3480.

Kidd, G., Mason, C.R., Richards, V.M., Gallun, F.J., Durlach, N.I., 2008. Informational masking. In: Yost, W.A., Fay, R.R. (Eds.), Auditory Perception of Sound Sources, Springer Handbook of Auditory Research, 29. Springer, Berlin, pp. 143–189.

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

Laures, J.S., Weismer, G., 1999. The effects of a flattened fundamental frequency on intelligibility at the sentence level. J. Speech Lang. Hear. Res. 42, 1148–1156.

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

Mattys, S.L., Davis, M.H., Bradlow, A.R., Scott, S.K., 2012. Speech recognition in adverse conditions: a review. Lang. Cogn. Proc. 27, 953–978.

Micheyl, C., Oxenham, A.J., 2010. Pitch, harmonicity and concurrent sound segregation: psychoacoustical and neurophysiological findings. Hear. Res. 266, 36–51.

Miller, S.E., Schlauch, R.S., Watson, P.J., 2010. The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. J. Acoust. Soc. Am. 128, 435–443.

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

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

Remez, R.E., Dubowski, K.R., Davids, M.L., Thomas, E.F., Paddu, N.U., Grossman, Y.S., 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., Lang, J.M., 1994. On the perceptual organization of speech. Psychol. Rev. 101, 129–156.

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

Roberts, B., 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., 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., 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., Bailey, P.J., 2015. Acoustic source characteristics, across-formant integration, and speech intelligibility under competitive conditions. J. Exp. Psychol. Hum. Percept. Perform. 41, 680–691.

Rosenberg, A.E., 1971. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am. 49, 583–590.

Shinn-Cunningham, B.G., 2008. Object-based auditory and visual attention. Trends Cogn. Sci. 12, 182–186.

Snedecor, G.W., Cochran, W.G., 1967. Statistical Methods, sixth ed. Iowa University Press; Ames, IA.

Stone, M.A., Canavan, S., 2016. The near non-existence of “pure” energetic masking release for speech: extension to spectro-temporal modulation and glimpsing. J. Acoust. Soc. Am. 140, 832–842.

Stone, M.A., Moore, B.C.J., 2014. On the near non-existence of “pure” energetic masking release for speech. J. Acoust. Soc. Am. 135, 967–1977.

Summers, R.J., Bailey, P.J., 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., 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., 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.

Wingfield, A., Lombardi, L., Sokol, S., 1984. Prosodic features and the intelligibility of accelerated speech: syntactic versus periodic segmentation. J. Speech Hear. Res. 27, 128–134.