Informational masking of speech by acoustically similar intelligible and unintelligible interferers

Robert J. Summers, and Brian Roberts

Citation: The Journal of the Acoustical Society of America **147**, 1113 (2020); doi: 10.1121/10.0000688

View online: https://doi.org/10.1121/10.0000688

View Table of Contents: https://asa.scitation.org/toc/jas/147/2

Published by the Acoustical Society of America

JASA ARTICLE

Informational masking of speech by acoustically similar intelligible and unintelligible interferers

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

ABSTRACT:
Masking experienced when target speech is accompanied by a single interfering voice is often primarily informational masking (IM). IM is generally greater when the interferer is intelligible than when it is not (e.g., speech from an unfamiliar language), but the relative contributions of acoustic-phonetic and linguistic interference are often difficult to assess owing to acoustic differences between interferers (e.g., different talkers). Three-formant analogues ($F1+F2+F3$) of natural sentences were used as targets and interferers. Targets were presented monaurally either alone or accompanied contralaterally by interferers from another sentence ($F0 = 4$ semitones higher); a target-to-masker ratio (TMR) between ears of 0, 6, or 12 dB was used. Interferers were either intelligible or rendered unintelligible by delaying $F2$ and advancing $F3$ by 150 ms relative to $F1$, a manipulation designed to minimize spectro-temporal differences between corresponding interferers. Target-sentence intelligibility (keywords correct) was 67% when presented alone, but fell considerably when an unintelligible interferer was present (49%) and significantly further when the interferer was intelligible (41%). Changes in TMR produced neither a significant main effect nor an interaction with interferer type. Interference with acoustic-phonetic processing of the target can explain much of the impact on intelligibility, but linguistic factors—particularly interferer intrusions—also make an important contribution to IM.
https://doi.org/10.1121/10.0000688

(Received 17 July 2019; revised 6 December 2019; accepted 19 January 2020; published online 14 February 2020)
[Editor: Hari M. Bharadwaj] Pages: 1113–1125

I. INTRODUCTION

Speech is a sparse signal in a frequency × time representation. Consequently, when we listen to a talker in the presence of one or two interfering voices, there is often relatively little energetic masking (EM) of the target speech unless the level of the interfering speech is high. Rather, the masking we experience is often primarily informational (Brungart *et al.*, 2006; see also Darwin, 2008). Informational masking (IM) arises in the central auditory system from three broad and overlapping causes—failures of object formation owing to incomplete perceptual segregation of target and interferer, failures of selective attention to properties of the target, and the use of limited central resources to process the interferer that would otherwise be available to process the target (see, e.g., Bregman, 1990; Shinn-Cunningham, 2008). These aspects of IM can have a considerable effect on the success of spoken communication even in circumstances where the properties of the target speech are well represented in the responses of the peripheral auditory system, and their effect is likely to be even greater when the peripheral representation of the target is degraded by EM or by sensorineural hearing loss (see, e.g., Moore, 1998).

Determining the circumstances in which there is release from IM is important for understanding speech intelligibility in adverse listening conditions. For example, it is known that the release of target speech from IM is facilitated when target-masker similarity is reduced—e.g., when the interfering speech is spatially distinct from the target (Freyman *et al.*, 2001; Arbogast *et al.*, 2002) or is spoken by a different-sex talker (Brungart *et al.*, 2001; Kidd *et al.*, 2016). Many studies have also reported that intelligible interferers cause more IM than broadly comparable unintelligible interferers (e.g., Freyman *et al.*, 2001; Calandruccio *et al.*, 2010; Brouwer *et al.*, 2012). In most studies, the unintelligible interferers used are either created by time-reversing intelligible speech or are similar utterances drawn from an unfamiliar language. However, estimating the extent of the contribution of linguistic factors is almost always complicated by the remaining acoustic differences between these interferers or difficulties in isolating the informational from the energetic components of speech-on-speech masking. The aim of the study reported here is to illuminate further the relative contributions of acoustic-phonetic interference and linguistic interference to speech-on-speech masking in a context where the informational component of masking is isolated effectively and acoustic differences between corresponding intelligible and unintelligible interferers are minimized. We define acoustic-phonetic interference as those aspects of IM that hinder the extraction or integration of information about speech articulation carried by the time-varying formant-frequency contours, and linguistic interference as those aspects of IM that occur after lexical objects have been formed, such as the intrusion of words from an interfering sentence into the percept of the target sentence.

[a]Electronic mail: r.j.summers@aston.ac.uk, ORCID: 0000-0003-4857-7354.
[b]ORCID: 0000-0002-4232-9459.

Unfortunately isolating the IM components of speech-on-speech masking is difficult, because the intelligibility of clear natural speech presented monaurally is usually unaffected by contralateral interfering speech (e.g., Cherry, 1953), owing to strong spatial release from IM in speech perception (e.g., Freyman *et al.*, 2001). One way to tackle this problem is to retain a dichotic configuration but to degrade the monaural target speech before adding the contralateral masker—e.g., by adding a fixed-level ipsilateral noise masker to the target speech (Brungart and Simpson, 2002). This approach has been applied successfully in several studies of IM, but to our knowledge—with one exception, discussed below (Gallun *et al.*, 2007)—whenever it has been used to compare the IM generated by intelligible and unintelligible interferers, their acoustic properties arguably have not been matched sufficiently closely. Another approach, more often employed in studies attempting to compare the IM produced by intelligible and unintelligible interferers, is to allow within-ear mixing of the target and interferer being manipulated (usually binaural presentation) and either to attempt to change the properties of the interferer in a way that is anticipated not to change the EM it produces (e.g., Cullington and Zeng, 2008) or to estimate and take into account differences in the EM caused by the different interferers used (e.g., Kidd *et al.*, 2016). The success of this approach is ultimately limited by the extent to which differences between maskers in the EM produced can be either eliminated or factored out.

When there is within-ear mixing of target and interferer, creating intelligible and unintelligible interferers that produce precisely matched levels of EM is difficult. For example, time-reversing intelligible speech to render it unintelligible is an appealing manipulation because it does not change the spectral content of the stimulus; it does, however, change its attack and decay characteristics. Given that the shape of the temporal envelope of speech is typically dominated by plosive sounds, and these sounds are characterized by rapid onsets and slow decays (Rosen, 1992), reversal of the speech signal results in envelopes with abrupt offsets. This has been shown to increase the forward-masking component of the EM produced by ∼2–3 dB (Rhebergen *et al.*, 2005). When it is anticipated that changing the properties of the interferer is likely to change both the IM and EM of the target, an effective method of isolating the EM component is required. Kidd *et al.* (2016) used ideal time-frequency segregation (ITFS) processing (Wang, 2005; Brungart *et al.*, 2006) to estimate the relative levels of EM arising from time-forward (intelligible) and time-reversed (unintelligible) two- or four-talker speech maskers by comparing their effects on the intelligibility of target speech in the case where all spectro-temporal information was preserved (natural condition) with the case where only those time-frequency elements dominated by the target were preserved (glimpsed condition). Target intelligibility was higher for the time-reversed interferers in the natural case but similar levels of intelligibility were found for both types of interferer in the glimpsed cases. Kidd *et al.* (2016) concluded that, since these maskers generated similar amounts of EM, the differences in speech reception threshold between the masking conditions in the natural cases (∼17 dB for the two-talker masker) were due to IM.

The release from IM reported by Kidd *et al.* (2016) is relatively large compared with that reported in many other studies (e.g., 5–7 dB by Cullington and Zeng, 2008; ∼5 dB by Freyman *et al.*, 2001), but there are a number of caveats to the ITFS approach that merit comment. First, this approach uses a simple energy-based model that does not take into account factors such as forward masking between time-frequency elements (see Brungart *et al.*, 2006); nonetheless, it should be acknowledged that failure to account for the greater forward masking associated with time-reversed than time-forward speech (Rhebergen *et al.*, 2005) in the context of the study by Kidd *et al.* (2016) would be expected to reduce, rather than to increase, the estimated difference in IM between the two conditions. Second, time reversal of two-talker maskers typically leads to greater release from IM than for single-talker maskers (e.g., Kidd *et al.*, 2010). Third, listeners may not process a stimulus containing actual gaps (deleted time-frequency elements; glimpsed condition) in the same way as a stimulus in which the presence of a masker indicates regions of "missing evidence" about the target (see Bregman, 1990). Indeed, it has been shown that filling the gaps in ITFS-processed speech with unmodulated broadband noise can improve its intelligibility (Cao *et al.*, 2011). Finally, there appears to be more IM—and hence greater scope for release from IM—when the target stimuli and maskers are drawn from the same closed set. For one-talker maskers, masking release for closed-set stimuli is ∼5 dB (e.g., Kidd *et al.*, 2010) and for open-set stimuli it is ∼0–4 dB (Rhebergen *et al.*, 2005; Cullington and Zeng, 2008); for two-talker maskers, masking release for closed-set stimuli is ∼10–17 dB (Marrone *et al.*, 2008; Kidd *et al.*, 2010, 2016), and for open-set stimuli it is ∼5–7 dB (Freyman *et al.*, 2001; Cullington and Zeng, 2008).

As noted above, the role of the linguistic properties and content of interfering speech in IM of target speech has received much attention. Masking of target speech by interferers spoken in an unfamiliar language tends to be lower in comparison with interferers spoken in the same language as the target (e.g., Freyman *et al.*, 2001; Van Engen and Bradlow, 2007). However, some caution is required when interpreting these results in terms of linguistic contributions to IM because differences in EM may also occur owing to acoustic differences between speech materials drawn from different languages (e.g., Calandruccio *et al.*, 2013). To our knowledge, no study comparing IM produced by interfering speech drawn from familiar or unfamiliar languages has controlled for EM by using a dichotic stimulus configuration or by applying ITFS processing; rather the EM component is usually assumed to remain fairly constant across interferer type. Furthermore, the use of different talkers for the familiar- and unfamiliar-language interferers is common and inevitably leads to spectro-temporal differences between the interferers (a notable exception is the study by Freyman et al., 2001). For example, although Calandruccio et al. (2013) found for English monolinguals that release from masking increased with increasing language dissimilarity of the interferer (English, Dutch, or Mandarin), they acknowledged that (in addition to the use of different talkers) some of the masking effects may have been accounted for by differences in long-term-average spectrum between the different-language interferers used.

Notwithstanding the difficulties in obtaining appreciable IM of natural speech when using dichotic presentation of target and masker, Dai et al. (2017) avoided altogether the issue of spectral dissimilarity between corresponding intelligible and unintelligible maskers. They tested the intelligibility of monaural natural speech in the presence of contralateral 2- or 4-band noise-vocoded speech maskers (NV2 or NV4) at various target-to-masker ratios (TMRs). NV2 speech was almost completely unintelligible; NV4 speech was fairly unintelligible but its intelligibility increased after training. Listeners were tested both before and after training on NV4 speech. Overall, perhaps unsurprisingly, the presence in the contralateral ear of low-resolution vocoded interferers had relatively little impact on target intelligibility, even at negative TMRs, but nonetheless there was a 2%–3% fall in target intelligibility in the presence of NV4 interfering speech after training on the latter. This small difference was attributed to pure linguistic interference. The implication of the study by Dai et al. (2017) is that there is a genuine linguistic component of IM, but that it may be relatively modest in size. Note, however, that target-masker similarity was low in their experiment because the target speech was natural and the interfering speech was noise-vocoded.

Usually, pure IM of natural speech can only be demonstrated if the monaural target speech is degraded in some way (e.g., Brungart and Simpson, 2002). In previous work (Roberts and Summers, 2015, 2018; Summers et al., 2016) we have shown that, in conditions of spatial uncertainty, monaural three-formant buzz-excited synthetic speech can be masked, often substantially, by presenting a single extraneous formant derived from $F2$ (termed an $F2$ competitor, or $F2$C) in the other ear. The properties of these extraneous formants were created in various ways—e.g., by time reversal or inversion of the target's $F2$ formant frequency contour. Interference is minimal if $F2$C has constant frequency, and increases until the range of frequency variation in $F2$C is around 150% of that in the natural $F2$ contour for stimuli derived from clearly enunciated speech (Roberts and Summers, 2015). This effect did not depend on whether the pattern of formant-frequency variation in the competitor was speech-like (inverted $F2$ frequency contour) or not (contour derived from a periodic triangle wave; Roberts et al., 2014). Three-formant interferers (time-reversed $F1$, $F2$, and $F3$ contours) have an even greater effect than a single-formant interferer derived from $F1$ (Roberts and Summers, 2018), and this difference cannot be attributed to the increase in total energy (typically <1 dB, as $F1$ contains most of the energy). These findings suggest that, whatever the contribution of linguistic factors, interference with acoustic-phonetic processing—which is heavily dependent on the extraction and integration of information carried by formant-frequency change—also plays a major role in speech-on-speech IM.

In order to assess the linguistic component of speech-on-speech IM, we need to generate corresponding intelligible and unintelligible three-formant interferers that are as acoustically similar as possible. As noted earlier, time reversal of speech is known to change its forward-masking properties (Rhebergen et al., 2005) but, furthermore, it cannot be ruled out that time reversal may also affect the non-linguistic aspects of IM. An alternative approach to time reversal for rendering interferers unintelligible is suggested by a study that explored the effects of formant asynchrony on the perception of sine-wave speech. Intelligible sine-wave speech can be made unintelligible by introducing asynchrony between the formant tracks while preserving the time-forward frequency and amplitude properties of each individual track (Remez et al., 2008). Intelligibility fell to near floor once the asynchrony of the tonal analogue of $F2$ was at least 100 ms relative to the analogues of $F1$ and $F3$.

Two experiments are reported here. Experiment 1 assessed the effect on intelligibility of different extents of on-going formant asynchrony in more natural analogues of speech, using three-formant buzz-excited materials. Experiment 2 compared the impact on the intelligibility of monaural synthetic speech caused by interfering speech in the contralateral ear that was either intelligible or was acoustically similar but rendered unintelligible using a suitably large on-going formant asynchrony, based on the results of experiment 1.

II. EXPERIMENT 1

This experiment explored the effect on intelligibility of introducing and manipulating the duration of an on-going asynchrony between $F1$, $F2$, and $F3$ in synthetic versions of sentence-length utterances. The aim was to identify the extent of formant asynchrony needed to render otherwise intelligible buzz-excited interferers largely unintelligible. Since $F1$ is the most intense formant, it remained unchanged and was used as the reference case. $F2$ was delayed and $F3$ was advanced with respect to $F1$ over the range 0 to 200 ms (cf. Remez et al., 2008).

A. Method

1. Listeners

All 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 training designed to improve the intelligibility of the speech analogues used (see

Sec. II A 3). About two-thirds of these listeners completed the training successfully and took part in the main experiment. All of them met the additional criterion of a mean score of ≥20% keywords correct in the main experiment, when collapsed across conditions. This nominally low criterion was chosen to take into account the poor intelligibility expected for some of the stimulus materials used. Twelve listeners (all female) successfully completed the experiment (mean age = 19.4 yr, range = 18.6–21.9). 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 36 sentences used were selected to contain ≤25% phonemes involving vocal tract closures or unvoiced frication. A set of keywords was chosen for each sentence; most designated keywords were content words. The stimuli for the training session were derived from 50 sentences spoken by a different talker and taken from commercially available recordings of the Harvard sentence lists (IEEE, 1969). These sentences were also selected to contain ≤25% phonemes involving closures or unvoiced frication.

For each sentence, the frequency contours of the first three formants were estimated from the waveform automatically every 1 ms from a 25-ms-long Gaussian window, using custom scripts in Praat (Boersma and Weenink, 2017). 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 digital second-order resonators in parallel whose outputs were summed. Following Klatt (1980), the outputs of the resonators corresponding to $F1$, $F2$, and $F3$ were summed using alternating signs (+, −, +) to minimize spectral notches between adjacent formants in the same ear. A monotonous periodic source with a fundamental frequency ($F0$) of 140 Hz was used in the synthesis of all stimuli for the training and main experiment; note that there was no noise source and so all phonetic segments in these analogues were rendered fully as voiced, regardless of their original source characteristics. The excitation source was a periodic train of simple excitation pulses modeled on the glottal waveform, which Rosenberg (1971) has shown to be capable of producing synthetic speech of good quality. The 3-dB bandwidths of the resonators corresponding to $F1$, $F2$, and $F3$ were set to constant values of 50, 70, and 90 Hz, respectively.

Stimuli for the different conditions were created by manipulating the asynchrony applied jointly to the frequency and amplitude contours of $F2$ (delayed) and $F3$ (advanced) with respect to $F1$. A delay in $F2$ was achieved by removing a section equivalent to the duration of the desired asynchrony from the end of the formant track and inserting it at the beginning; an advance in $F3$ involved removing a section from the beginning of the track and inserting it at the end. A 25-ms half-cycle of a cosine function was used to smooth the join in the spliced formant frequency contour; note that this tactic was purely precautionary because the join corresponded to the beginning and end of the original contour and hence the formant amplitude around this point was close to zero. Examples of the stimuli from this experiment can be found in the supplementary material.[1]

There were six conditions in this experiment, corresponding to $F2$ and $F3$ asynchronies of ±0, 25, 50, 100, 150, and 200 ms with respect to $F1$. The stimuli are illustrated in Fig. 1 using the wideband spectrogram of a synthetic analogue of an example sentence and its waveform (top row) and after processing by the five $F2$ and $F3$ asynchronies used (remaining rows). For each listener, the 36 sentences were divided equally across conditions (i.e., six per condition), such that there were 19 keywords in each condition. Allocation of sentences to conditions was counterbalanced by rotation across each set of six listeners tested. Hence, the total number needed to produce a balanced dataset was a multiple of six listeners.

3. Procedure

During testing, listeners were seated in front of a computer screen and a keyboard in a single-walled sound-attenuating chamber (Industrial Acoustics 401 A; Winchester, United Kingdom) housed within a quiet room. The experiment consisted of training 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, diotic presentation was used and the stimuli were presented in a new quasi-random order for each listener.

The training session comprised 50 trials; stimuli were presented without interferers and a new sentence was used for each trial. On each of the first ten trials, listeners heard 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

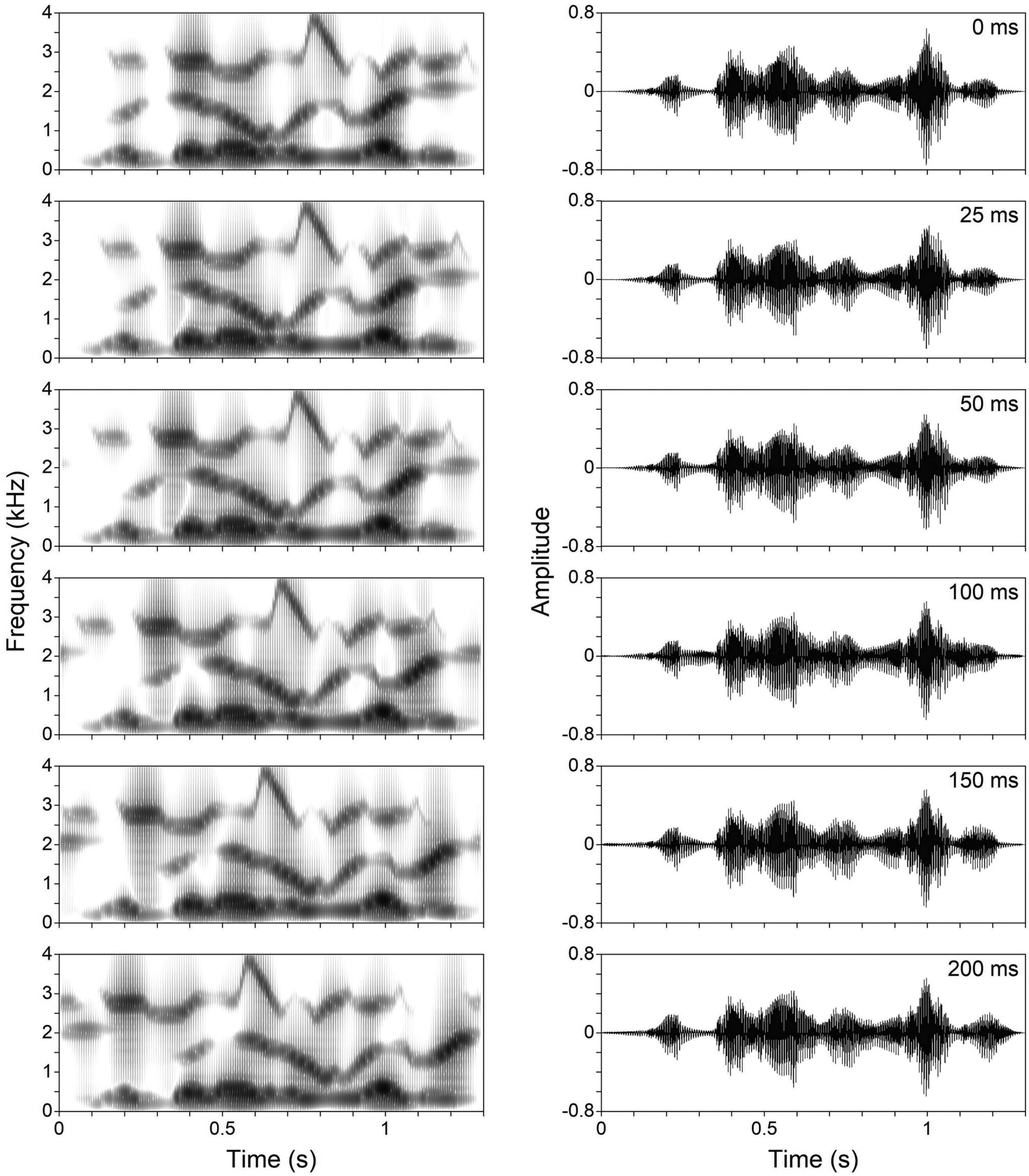

FIG. 1. Stimuli for experiment 1—wideband spectrograms (left column) and waveforms (right column) for a three-formant analogue of the sentence "The dinner was ready," for increasing durations of on-going formant asynchrony (descending rows). $F2$ was delayed and $F3$ was advanced with respect to $F1$ using asynchronies of ±0, 25, 50, 100, 150, and 200 ms. Note that the waveform is relatively insensitive to these changes in formant asynchrony.

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 6 times before typing in their transcription. After each transcription was entered, feedback was provided by playing the original recording (44.1 kHz sample rate) followed by a repeat of the synthetic version. Davis *et al.* (2005) found that the strategy of providing feedback using an alternating presentation 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 needing 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.

All speech analogues were synthesized using MITSYN (Henke, 2005) at a sample rate of 40 kHz and with 10-ms raised-cosine onset and offset ramps. They were played at 16-bit resolution over Sennheiser HD 480–13II earphones (Hannover, Germany) via a Sound Blaster X-Fi HD sound card (Creative Technology Ltd., Singapore), programmable attenuators (Tucker-Davis Technologies, TDT PA5, Alachua, FL), and a headphone buffer (Tucker-Davis Technologies, TDT HB7, Alachua, FL). 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, Nærum, Denmark). All target sentences were presented at a long-term average of 72 dB sound pressure level.

4. Data analysis

The stimuli for each condition comprised six sentences. Given the variable number of keywords per sentence (2–4), the mean score for each listener in each condition was computed as the percentage of keywords reported correctly giving equal weight to all the keywords used. As in our previous studies (e.g., Roberts *et al.*, 2010; Roberts and Summers, 2015, 2019), we classified responses using tight scoring, in which a response is scored as correct only if it matches the keyword exactly; homonyms were accepted. Except where stated otherwise, the values and statistics reported here are based on these tight keyword scores. All statistical analyses reported here were computed using R 3.5.3 (R Core Team, 2019) 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).

B. Results and discussion

Figure 2 shows the mean percentage scores (and intersubject standard errors) across conditions for keywords correctly identified as a function of formant asynchrony. Intelligibility was relatively good in the reference condition (0 ms: ∼58% keywords correct), despite the simple source properties and three-formant parallel vocal-tract model used to synthesize the sentences, but fell progressively as formant asynchrony increased. Performance was at floor for the longest formant asynchronies tested (150 and 200 ms: ∼3% and ∼2% keywords correct, respectively). A one-way within-subjects analysis of variance (ANOVA) of the keyword scores across all six conditions showed that the effect of formant asynchrony on intelligibility was highly significant [$F(5,55) = 89.555$, $p < 0.001$, $\eta^2 = 0.891$].[2] Pairwise comparisons of the keyword scores between neighboring test values of formant asynchrony (e.g., 0 vs 25 ms, 25 vs 50 ms, etc.) revealed that there were significant differences between all of them (range: $p = 0.008 - p < 0.001$), except for the cases 0 vs 25 ms [mean difference = 10.1 percentage points (% pts); $p = 0.061$] and 150 vs 200 ms (mean difference = 1.3% pts; $p = 0.339$).

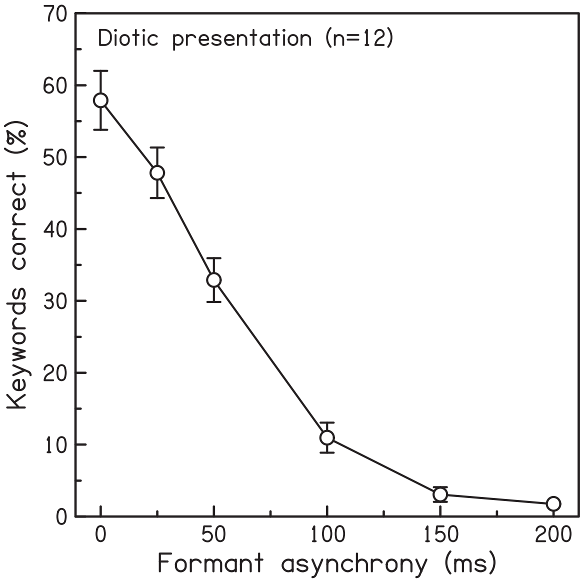

FIG. 2. Results for experiment 1—effect of formant asynchrony on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and intersubject standard errors ($n = 12$) are shown for the six asynchronies tested (±0, 25, 50, 100, 150, and 200 ms). For each asynchrony tested, $F2$ was delayed and $F3$ was advanced relative to $F1$, which was unchanged.

The results of this experiment merit comparison with those reported by Remez *et al.* (2008) for sine-wave speech. In their experiment, the timing of the tonal analogues of $F1$ and $F3$ both remained unchanged; rather, the asynchrony of the tonal analogue of $F2$ was manipulated either to lead or lag the other formants (without wrap-around). Performance in their experiment was measured in terms of syllables correct and showed an apparently more marked fall in performance between asynchronies of 0 and 50 ms (72% vs 32% syllables correct) than was found here for the corresponding comparison (58% vs 33% keywords correct). Also, Remez *et al.* (2008) found for sine-wave speech that performance reached floor for an $F2$ asynchrony of 100 ms (∼5% syllables correct) but a formant asynchrony of 150 ms (∼3% keywords correct) was necessary for the buzz-excited analogues used here. Bearing in mind that, in the experiment reported here, all three formants were made asynchronous relative to one another and that the asynchrony between $F2$ and $F3$ was twice that of their asynchrony with $F1$, it would appear that buzz-excited speech is more robust than sine-wave speech to the effects of formant asynchrony.

III. EXPERIMENT 2

This experiment compared directly the effects of intelligible and unintelligible three-formant contralateral interferers on target intelligibility in a context where acoustic differences between corresponding pairs of intelligible and unintelligible interferers were minimized. The set of unintelligible interferers was derived from the set of intelligible interferers by delaying $F2$ and advancing $F3$ by 150 ms with respect to $F1$; the results of experiment 1 indicated that further increases in formant asynchrony would lead to little further reduction in their intelligibility. In addition to manipulating formant asynchrony, the effects of TMR were assessed at 0, 6, and 12 dB. Gallun *et al.* (2007) showed that there are circumstances in which the IM of speech can be altered substantially by changes in TMR across ears of 10 dB.

A. Method

Except where described, the same method was used as for experiment 1. There were seven conditions in experiment 2; hence, the number of listeners required to produce a balanced dataset was a multiple of seven. Twenty-eight listeners (eight males) passed the training and successfully completed the experiment (mean age = 25.3 yr, range = 18.1–47.9); none of these listeners took part in experiment 1. The training session was nearly identical to that for experiment 1, with the exception that the last ten sentences were presented monaurally, rather than diotically, and with random selection of ear of presentation on each trial. The stimuli for the main experiment were derived from recordings of 60 sentences drawn from the same set of materials as those used in experiment 1 and spoken by the same talker. Forty-two of the sentences were designated target sentences and were allocated to different conditions in the same way as for experiment 1 (18–19 keywords per condition). The remaining 18 sentences were used to create 36 interferers; half were intelligible (0-ms formant asynchrony) and the other half were unintelligible (150-ms formant asynchrony).

All stimuli were generated using the same excitation source, resonator bandwidths, and synthesizer configuration as for experiment 1. The $F0$ frequencies of the target speech and the interfering speech were 120.3 and 150.5 Hz, respectively. These $F0$s correspond to 135 Hz ± 2 semitones; a 4-semitone difference was used to distinguish clearly between the target speech and the interfering speech. The target speech was presented to one ear only, selected randomly on each trial to create spatial uncertainty and hence to increase IM (see, e.g., Kidd *et al.*, 2008). When present, the interferer was received in the contralateral ear. Listeners were asked to ignore the sounds on the higher pitch and only to transcribe the words on the lower pitch. For each stimulus including an interferer, the durations of the target and interferer were matched to their mean duration by linear interpolation of the formant frequency and amplitude contours prior to re-synthesis; this method did not affect the properties of the excitation source. In order to minimize the degree of duration rescaling necessary, the stimuli for the targets and interferers were sorted by duration and paired up; if a pairing resulted in the target and interferer sharing more than one keyword then the next-nearest interferer in duration was chosen. This pairing and rescaling was done separately for each rotation of the stimuli across conditions. Over all seven rotations of the target stimuli across conditions, only 24 out of the 252 target/interferer pairs shared a keyword.

Table I illustrates the seven conditions in the main experiment; for ease of reference, the formants of the masker are labelled $M1$, $M2$, and $M3$. Six conditions (C1–C6) contained an interferer and one (C7) was a reference condition, comprising only the target speech. For three of the conditions including interferers (C1–C3), the interferer had a formant asynchrony of 0 ms (i.e., it was intelligible) and for the remaining three interferer conditions (C4–C6) the formant asynchrony used was 150 ms (i.e., the interferer was unintelligible). Stimuli were selected such that the frequency of the target $F2$ was always at least 80 Hz away from the frequencies of $F1$ and $F3$ at any one moment, irrespective of the formant asynchrony applied. Hence, there were no approaches between formant tracks close enough to cause audible interactions between corresponding harmonics exciting adjacent formants. The across-ear TMR was 12 dB (C1 and C4), 6 dB (C2 and C5), or 0 dB (C3 and C6). Target level was unchanged; attenuation of the masker was achieved without loss of resolution by reducing the output of the appropriate channel using one of the programmable attenuators. A follow-up experiment to assess the intelligibility of the interferers from the main experiment was carried out immediately afterwards. These interferers were presented diotically on the same $F0$ (150.5 Hz), once each and without feedback, in a quasi-random order.

B. Results and discussion

Figure 3 shows the mean percentage keyword scores (and intersubject standard errors) for the target sentences, either separately for each condition (top panel) or averaged across TMR for the two types of interferer tested (bottom panel). A one-way within-subjects ANOVA over all seven conditions showed a highly significant effect of condition

TABLE I. Stimulus properties for the conditions in experiment 2 (main session). Three-formant maskers ($M1$+$M2$+$M3$) were used as interferers. Interferers were either intelligible (formant asynchrony = 0 ms) or rendered unintelligible by delaying $F2$ and advancing $F3$ with respect to $F1$ (asynchrony = ±150 ms; see text for details). Stimuli were presented at three different TMRs across ears (12, 6, or 0 dB) by attenuating the interferer as required.

Condition	Stimulus configuration (target ear; other ear)	Asynchrony of $M2$ (+) and $M3$ (−) with respect to $M1$ (ms)	TMR (dB)
C1	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	0	12
C2	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	0	6
C3	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	0	0
C4	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	±150	12
C5	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	±150	6
C6	($F1$+$F2$+$F3$; $M1$+$M2$+$M3$)	±150	0
C7	($F1$+$F2$+$F3$; —)	—	—

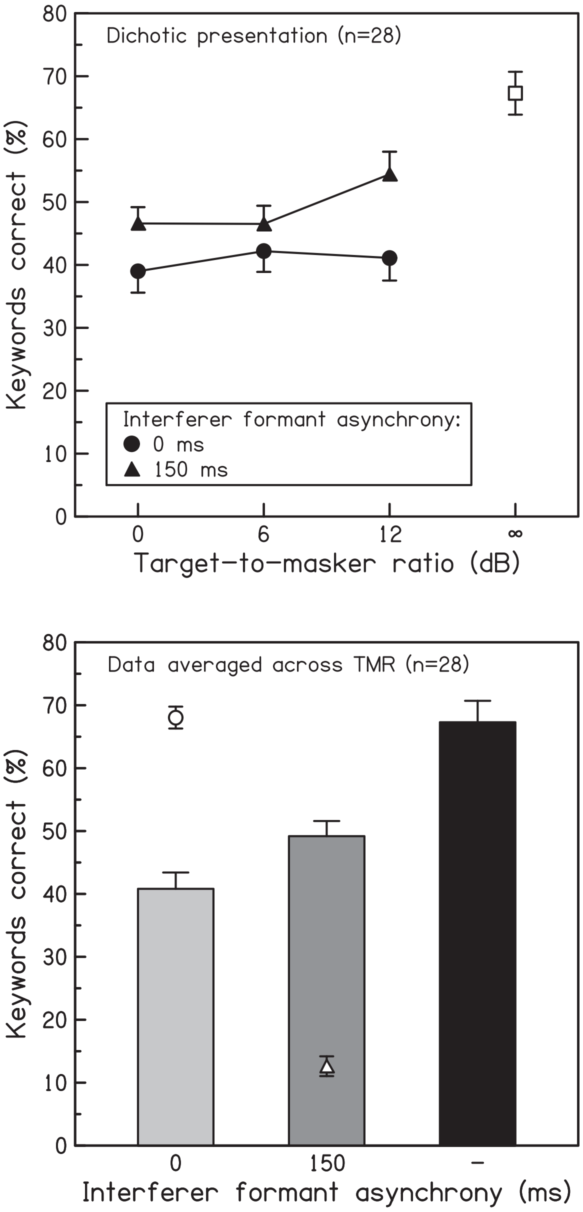

FIG. 3. Results for experiment 2—effect of target-to-masker ratio (0, 6, or 12 dB) and of the formant asynchrony applied to interfering speech (0 or 150 ms) on the intelligibility of three-formant analogues of the target sentences. Mean keyword scores and intersubject standard errors ($n = 28$) are shown. The top panel shows performance in the absence of interfering speech (square) and for the three TMRs tested in the presence of unintelligible interferers (150 ms asynchrony, triangles) or intelligible interferers (0 ms asynchrony, circles). The bottom panel shows mean performance when collapsed across TMR for the intelligible interferers (light gray bar), the unintelligible interferers (dark gray bar), and for no interferer (black bar). The symbols shown, circle and triangle, indicate performance for diotic presentation of the interferers alone with 0 and 150 ms asynchronies, respectively, in the follow-up experiment.

on target intelligibility [$F(6,162) = 14.843$, $p < 0.001$, $\eta^2 = 0.355$]. Performance was best (~67% keywords correct) when the three target formants were presented alone (C7). Pairwise comparisons showed that intelligibility was significantly lowered, often substantially, when the target speech was accompanied by any of the interferers (C7 vs C1–C6, overall mean difference = 22.4% pts, $p < 0.001$ in all cases). This substantial decrease in target intelligibility occurred even though the interferer was presented in the ear contralateral to the target and on a different $F0$ ($\Delta F0 = 4$ semitones).

The effect of the experimental manipulations of the interfering formants was explored further using a two-way ANOVA restricted to the target-plus-interferer conditions (C1–C6). The two factors were formant asynchrony applied to the interferer (two levels: 0 or 150 ms) and across-ear TMR (three levels: 0, 6, or 12 dB). This analysis revealed a significant main effect of formant asynchrony [$F(1,27) = 27.897$, $p < 0.001$, $\eta_p^2 = 0.508$], but there was no main effect of TMR [$F(2,54) = 1.527$, $p = 0.226$, $\eta_p^2 = 0.054$] and no interaction between the two factors [$F(2,54) = 1.994$, $p = 0.146$, $\eta_p^2 = 0.069$]. Note, however, that visual inspection of Fig. 3 suggests at least the possibility of an interaction between TMR and formant asynchrony, implying that listeners may have used the level cue in combination with the lack of interferer intelligibility to direct their attention quickly to the target ear. Given that repeated-measures ANOVA does not take into account random effects arising from differences in the intelligibility of different targets and interferers, as a precautionary measure we performed a further analysis using a linear mixed effects model that does take them into account. Following the approach of Luke (2017), and using the package *lmerTest* (Kuznetsova *et al.*, 2017) for its implementation of the Satterthwaite approximation for estimating the degrees of freedom of the denominator term in the F statistic, the analysis confirmed the results of the original ANOVA. Specifically, there was a significant main effect of formant asynchrony [$F(1,26.38) = 15.360$, $p < 0.001$], but there was no main effect of TMR [$F(2,929.11) = 2.344$, $p = 0.096$] and no interaction between the two factors [$F(2,929.11) = 2.379$, $p = 0.093$] for the dichotic stimulus configuration used here. Although it seems likely that using a larger range of TMRs or a greater number of test materials may have revealed a significant main effect or interaction, the effect of masker attenuation was clearly substantially less than that of formant asynchrony. This outcome suggests that the masking effects observed here may be largely obligatory and so relatively unmodulated by attentional focus. On average, the addition of an unintelligible interferer (150-ms asynchrony) lowered scores by 18.2% pts. Scores were lowered by a further 8.4% pts (i.e., by 26.6% pts) if the interferer was intelligible (0-ms asynchrony).

The results for the follow-up experiment (symbols in bottom panel of Fig. 3) confirmed that, heard in isolation, the speech analogues used as interferers were around as intelligible as the target speech when the formant asynchrony was 0 ms (~68% keywords correct) but intelligibility was low when the asynchrony was 150 ms (~13% keywords correct). Albeit that different listeners took part in this experiment, the keyword scores in the follow-up were somewhat higher than for the corresponding cases in experiment 1 (all 18 sentences used as interferers here were tested in experiment 1). This is probably a consequence of exposure to the interferers shortly beforehand in the main experiment; note also that hearing the intelligible version of a given interferer before the asynchronous version is likely to enhance the latter's intelligibility owing to the close acoustic similarity between them. At this point, it merits comment that our measure of intelligibility for the asynchronous stimuli was with reference to the designated keywords for the original sentences. Hence, at least in principle, it is possible that listeners may have perceived as many words for the asynchronous stimuli, but different ones from those keywords. This was not the case because, on average, listeners transcribed 5.0 words/stimulus for the synchronous interferers when presented alone in the follow-up experiment but only 3.3 words/stimulus for the asynchronous interferers.

The errors made by listeners when trying to report the target sentence in the target-plus-interferer conditions can also provide insight into the nature of the interference experienced. We began by pooling the results across TMR and identifying for each target sentence any errors corresponding to keywords present exclusively in the interfering sentences; these error counts were then expressed as a percentage of the total number of interferer keywords. On average, these scores were 12.6% and 0.8% for the intelligible and (notionally) unintelligible interferers, respectively. This outcome suggests that an appreciable proportion of the fall in target scores associated with intelligible interferers involved substitution of target keywords with words from the interferer.

The mistakes that listeners made can, in principle, be classified into those arising from reporting keywords from the interfering sentence (intrusion errors) and those arising from reporting words that were not present in either the target or the interferer (other errors). We classified all reported words that would usually be considered to be keywords into these two categories; occasions where the interferer and target shared a keyword that was reported by the listener were not counted as errors. These error counts should be considered against the baseline of 54–55 target keywords overall for each level of formant asynchrony (0 vs 150 ms) when pooled across the three levels of TMR. The average number of intrusion errors was higher for the intelligible (7.5) than for the (notionally) unintelligible interferers (0.5), whereas the average number of other errors was similar for the intelligible (17.6) and unintelligible interferers (18.6). This indicates that the primary difference in errors between the intelligible and unintelligible interferer conditions arises from intrusions. Furthermore, considering other errors as a proportion of total errors for the intelligible interferers implies that as much as 70% [17.6/(17.6 + 7.5)] of the IM caused by an intelligible interferer arises from acoustic-phonetic rather than from linguistic interference. Making the same calculation using loose scores yields only a small decrease in this estimate to 67% [16.9/(16.9 + 8.3)].

The significant additional effect on target speech intelligibility observed here when intelligible interferers were used is in accord with the results of experiments where target and masking speech are mixed together in the same ear and the masker is spoken in either the same language or a different (unfamiliar) language to the target speech (e.g., Brouwer *et al.*, 2012). Here, however, the dichotic configuration ensured that masking by the interferer was purely informational and the method used to control interferer intelligibility minimized the acoustic differences between corresponding intelligible and unintelligible interferers. Overall, the results demonstrate that speech-on-speech IM comprises an acoustic-phonetic component, which makes a substantial contribution to the masking of target speech, and a linguistic component, which can make a considerable additional contribution.

IV. GENERAL DISCUSSION

In experiment 1, introducing an increasingly large asynchrony between formants in buzz-excited three-formant analogues of speech led to a progressive fall in intelligibility that approached floor for asynchronies ≥150 ms (cf. ≥100 ms for sine-wave speech; Remez *et al.*, 2008). In experiment 2, where masking of the target speech by a single-talker interferer was purely informational, the impact on target intelligibility was substantial even though the target and interfering voices were on different $F0$s ($\Delta F0 = 4$ semitones) and presented in different ears. Moreover, intelligible interferers caused more masking than that caused by acoustically similar interferers rendered largely unintelligible by applying an ongoing asynchrony of ±150 ms to $F2$ and $F3$ relative to $F1$ (cf. the similar but small effect observed by Dai *et al.*, 2017, using NV4 speech maskers). This finding confirms the results of previous studies (e.g., Freyman *et al.*, 2001; Van Engen and Bradlow, 2007; Calandruccio *et al.*, 2013) that unintelligible interferers cause less IM than intelligible interferers, but avoids the potentially confounding issues of attempting to partition the energetic and informational components of masking when target and interfering speech are mixed in the same ear (e.g., Kidd *et al.*, 2016), and of the various acoustic differences often present between corresponding intelligible and unintelligible interferers (e.g., Rhebergen *et al.*, 2005). Also, the considerable effect of interferer intelligibility observed here occurred despite our use of open-set materials (cf. Iyer *et al.*, 2010); the use of the same closed-set materials for target and interferer tends to increase the amount of IM observed (e.g., Marrone *et al.*, 2008; Kidd *et al.*, 2010, 2016).

The difference in performance between conditions with intelligible and unintelligible interferers was primarily due to intrusion errors (i.e., reporting words from the interferer), rather than to other errors (i.e., reporting words that were not present in either the target or the interferer). Although not conclusive, this outcome may indicate that listeners sometimes had difficulty orienting their attention to the target rather than to the interferer, despite the clear difference in pitch between them arising from the one-third octave $\Delta F0$. Presumably, if this were the case, such a difficulty with selective attention would also have applied when the interferers were rendered unintelligible and, more generally, it would have facilitated not only intrusions but also acoustic-phonetic interference across ears. Of course, it is acknowledged that larger differences in pitch such as those typical of differences between adult male and female talkers may not have led to as many errors, and so the impact of increasing $\Delta F0$ on the number and proportion of the two types of error merits investigation in further research. Note also that the materials used here were short sentences, which limited the time available for reorienting attention, and so it may be the case that using longer materials would show clearer benefits of attentional cues such as differences in level or $F0$.

The lack of any significant effect of interferer attenuation is perhaps surprising, given that there are several studies in which changes in TMR within the range tested here have revealed fairly substantial effects on intelligibility (e.g., Brungart *et al.*, 2001, 2006; Gallun *et al.*, 2007; Thompson *et al.*, 2015). However, with the exception of Gallun *et al.* (2007), those studies used stimuli in which the target speech and the masker being manipulated were mixed in the same ear, such that changes in TMR would inevitably be expected to affect the extent of EM, irrespective of any possible effects on IM. Gallun *et al.* (2007) measured the intelligibility of monaural noise-vocoded target speech when mixed with a fixed-level ipsilateral masker and accompanied by a contralateral interferer whose level was varied systematically. They found that target intelligibility fell substantially over at least a 10-dB change in across-ear TMR. One might speculate that the discrepancy between our findings and theirs arises because of an interaction between the fixed ipsilateral masker and variable contralateral masker, perhaps due to an increased processing load (cf. Brungart and Simpson, 2007), that is not present in our experiment.

Our approach to rendering synthetic speech unintelligible has some similarities with that applied to natural speech by Carlile and Corkhill (2015). They decomposed the interfering speech into 22 bands, treated each band as a circular buffer, and then recombined the bands with random starting points. Their approach preserved the within-band spectro-temporal properties of the original signal but, unlike our approach, it did not preserve the coherent trajectories of the individual formants. Indeed, more generally, the relationship between an interferer's spectro-temporal coherence and the masking it generates has received little attention and remains an open question for research (see Roberts *et al.*, 2014, for a discussion). If the spectro-temporal coherence of an interferer is important for the IM it generates, then a possible modification to Carlile and Corkhill's method would be to filter natural speech into a small number of bands whose center frequencies and widths are matched to the overall ranges of the underlying formants, followed by recombining the bands after applying a constrained-random asynchrony to each one. These asynchronies may need to be relatively large to render natural speech unintelligible (see Arai and Greenberg, 1998). Note also that it is likely that the manipulation employed by Carlile and Corkhill (2015) would have substantially changed the overall amplitude envelope of the signal. In contrast, desynchronizing $F2$ and $F3$ relative to $F1$, as used here, caused relatively little change in the overall amplitude envelope of the interfering speech because it is governed primarily by the $F1$ amplitude envelope (see Fig. 1).

The impact of a time-varying interferer on target intelligibility observed here was considerable even when it was rendered unintelligible by introducing formant asynchrony (~18% pts fall in mean keyword scores). This outcome is broadly in accord with the results of our recent study using three-formant interferers made unintelligible by time reversal of their formant-frequency contours (Roberts and Summers, 2018), but the effect observed in that study was considerably larger (~39% pts fall). Both experiments used sentences drawn from the same set of BKB-like materials and it is likely that the more modest fall in keyword scores observed here is attributable mainly to the 4-semitone difference in $F0$ between the target and interferer. Specifically, the $\Delta F0$ may have facilitated attending to the target (whereas, in our previous study, all formants in the stimulus ensemble shared a common $F0$). Two other factors may also have contributed to the difference in impact of the interferers in the two experiments. First, different amplitude contours were used for the three-formant interferers; they were time-varying here but constant in the study by Roberts and Summers (2018). It has previously been shown for single-formant interferers presented in the ear contralateral to monaural target speech that constant-amplitude formants generate more IM than time-varying ones when matched for root-mean-square power (Roberts and Summers, 2015). The reason for this difference is unclear, but it may be because the formant-frequency variation in the interferer, which is known to be of primary importance for the IM generated (e.g., Roberts and Summers, 2015), is less clearly defined during the low-amplitude portions in the time-varying case. Second, each three-formant interferer used by Roberts and Summers (2018) was synthesized using the time-reversed formant-frequency contours of the corresponding target sentence, and so each target formant's counterpart in the interferer was exactly matched for its geometric mean frequency and frequency range; this was not the case in the current study.

Our assumption is that unintelligible interferers comprising single formants or formant ensembles interfere with basic acoustic-phonetic processing of the target speech, because this processing is heavily dependent on extracting and integrating information carried by the time-varying formant-frequency contours. The degree of interference caused by these maskers seems to be dependent on their spectro-temporal complexity. In particular, the greater the formant-frequency variation in the interferer, the greater the IM it produces (Roberts *et al.*, 2010, 2014; Roberts and Summers, 2015, 2018) and three-formant interferers ($F1$+$F2$+$F3$) cause more interference than that caused by an extraneous $F1$ alone (Roberts and Summers, 2018). However, as noted earlier, the extent to which the pattern of formant-frequency variation is plausibly speech-like does not appear to be important (Roberts *et al.*, 2014). Indeed, Roberts and Summers (2018) have pointed out that there are interesting parallels between the effect of formant-frequency variation in an interferer on the IM it produces and the irrelevant sound effect (ISE). The ISE demonstrates that task-irrelevant acoustic distractors involving frequency change cause significant cross-modal interference—e.g., to visual working memory (Jones and Macken, 1993). Furthermore, Dorsi *et al.* (2018) found that increasing the number of channels in noise-vocoded speech distractors increased the size of the ISE; this was also true in the case where two-thirds of the channels in the distractors were time reversed, rendering them largely unintelligible. Hence, the effect of frequency variation in the interferer on the ISE was separable from the effect of the interferer's intelligibility.

The importance of the non-linguistic component of speech-on-speech IM in our study is emphasized by the observation that the unintelligible interferers produced around two-thirds of the fall in keyword scores associated with the intelligible interferers. Indeed, the consequences of this component of IM for listeners with mild-to-moderate hearing loss may be even greater given the degraded peripheral representation of the target speech—in this regard, a useful aim for future research would be to establish the extent to which hearing-impaired listeners are susceptible to IM generated by formant-frequency variation in interfering speech-like stimuli (cf. Roberts and Summers, 2015). Nonetheless, it should be acknowledged that this assessment is likely to underestimate the contribution of the linguistic component of IM because it can only be demonstrated as *additional* masking to that produced by the non-linguistic component. Given that the stimulus configuration used in the current study is capable of demonstrating a reliable additional effect of interferer intelligibility on the perception of target speech presented in the contralateral ear, the question of how this approach might be used in future research merits consideration.

Although some attention has been paid to the contributions made by the lexical, semantic, and syntactic properties of *interfering speech* to speech-on-speech IM, the results of these studies have been inconsistent and, with one exception, the experiments involved interferers comprising two talkers. For example, Brouwer *et al.* (2012) and Calandruccio *et al.* (2018) both found some circumstances in which the semantic coherence of two-talker maskers affected target speech intelligibility, but the direction of these effects was variable—e.g., across two experiments semantically anomalous sentences caused either more or less masking than semantically coherent sentences (Calandruccio *et al.*, 2018). Although Kidd *et al.* (2014) have demonstrated the role of syntax in maintaining a coherent stream of attended speech when listening to mixtures of concurrent speech, to our knowledge the effect of the syntactic status of a single interfering voice on target speech intelligibility has only been investigated by Newman *et al.* (2015). Their study found no effect of the syntax (in-order vs scrambled-order sentences) of single-talker interfering speech, but the interfering stimuli were constructed by concatenating words and so lacked the typical coarticulation of spoken sentences, limiting the generality of their result. In principle, questions about the role of the linguistic properties of interfering speech in the IM it generates might be addressed using an approach like that taken in the current study, so long as it is possible to identify suitable materials for comparison (e.g., syntactic vs non-syntactic) that, when rendered unintelligible by applying the formant asynchrony manipulation, generate a similar degree of IM. If this proves possible, then differences in the *additional* impact of the interfering speech when rendered intelligible by removing the formant asynchrony must arise mainly from whichever linguistic properties characterize those stimuli.

In conclusion, the extent of IM caused by interfering speech may be reduced considerably (cf. Kidd *et al.*, 2016) when it is rendered unintelligible by introducing an ongoing asynchrony between its formants, despite the limited effect of this manipulation on the overall spectro-temporal properties of the interferer. The greater impact of intelligible interferers observed here arose primarily from intrusion of words from the interfering sentence into the target percept. Nonetheless, the impact of an interfering voice on the intelligibility of target speech remained substantial when the interfering speech was rendered unintelligible, despite the protection from EM provided by the dichotic configuration used. Overall, the results suggest that central interference with acoustic-phonetic processing of the target can explain much of the interferer's impact on intelligibility, but that linguistic factors (for example, lexical access) also make an important contribution to speech-on-speech IM.

ACKNOWLEDGMENTS

This research was supported by Research Grant No. ES/N014383/1 from the Economic and Social Research Council (United Kingdom), awarded to B.R. To access the research data underlying this publication, see https://doi.org/10.17036/researchdata.aston.ac.uk.00000459. We are grateful to Peter Bailey for his advice concerning this research and for his comments on drafts of this paper. We also thank Rob Morse and Meghna Patel for providing the BKB-like sentences and Quentin Summerfield for enunciating them. Poster presentations on this research were given at the 177th Meeting of the Acoustical Society of America (Louisville, KY, May 2019) and the annual Basic Auditory Science Meeting of the British Society of Audiology (University College London, United Kingdom, September 2019).

[1] See supplementary material at https://doi.org/10.1121/10.0000688 for six example sentences, each presented with the six different $F2$/$F3$ asynchronies used in experiment 1.

[2] As a precaution, given the low scores obtained in some conditions, all ANOVAs 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.

Arai, T., and Greenberg, S. (**1998**). "Speech intelligibility in the presence of cross-channel spectral asynchrony," in *Proceedings of the 1998 IEEE International Conference on Acoustical Speech Signal Processing*, pp. 933–936.

Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (**2002**). "The effect of spatial separation on informational and energetic masking of speech," J. Acoust. Soc. Am. **112**, 2086–2098.

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

Boersma, P., and Weenink, D. (**2017**). "PRAAT, a system for doing phonetics by computer [software package]," Institute of Phonetic Sciences, University of Amsterdam, The Netherlands, available at http://www.praat.org/ (Last viewed December 6, 2019).

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

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

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

Brungart, D. S., and Simpson, B. D. (**2002**). "Within-ear and across-ear interference in a cocktail-party listening task," J. Acoust. Soc. Am. **112**, 2985–2995.

Brungart, D. S., and Simpson, B. D. (**2007**). "Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task," J. Acoust. Soc. Am. **122**, 1724–1734.

Brungart, D. S., Simpson, B. D., Ericson, M. A., and Scott, K. R. (**2001**). "Informational and energetic masking effects in the perception of multiple simultaneous talkers," J. Acoust. Soc. Am. **110**, 2527–2538.

Calandruccio, L., Brouwer, S., Van Engen, K. J., Dhar, S., and Bradlow, A. R. (**2013**). "Masking release due to linguistic and phonetic dissimilarity between the target and masker speech," Am. J. Audiol. **22**, 157–164.

Calandruccio, L., Buss, E., Bencheck, P., and Jett, B. (**2018**). "Does the semantic content or syntactic regularity of masker speech affect speech-on-speech recognition?," J. Acoust. Soc. Am. **144**, 3289–3302.

Calandruccio, L., Dhar, S., and Bradlow, A. R. (**2010**). "Speech-on-speech masking with variable access to the linguistic content of the masker speech," J. Acoust. Soc. Am. **128**, 860–869.

Cao, S., Li, L., and Wu, X. (**2011**). "Improvement of intelligibility of ideal binary-masked noisy speech by adding background noise," J. Acoust. Soc. Am. **129**, 2227–2236.

Carlile, S., and Corkhill, C. (**2015**). "Selective spatial attention modulates bottom-up informational masking of speech," Sci. Rep. **5**, 8662.

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

Cullington, H. E., and Zeng, F.-G. (**2008**). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," J. Acoust. Soc. Am. **123**, 450–461.

Dai, B., McQueen, J. M., Hagoort, P., and Kösem, A. (**2017**). "Pure linguistic interference during comprehension of competing speech signals," J. Acoust. Soc. Am. **141**, EL249–EL254.

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

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

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

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2001**). "Spatial release from informational masking in speech recognition," J. Acoust. Soc. Am. **109**, 2112–2122.

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

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

Iyer, N., Brungart, D. S., and Simpson, B. D. (**2010**). "Effects of target-masker contextual similarity on the multimasker penalty in a three-talker diotic listening task," J. Acoust. Soc. Am. **128**, 2998–3010.

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

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

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

Kidd, G., Jr., Mason, C. R., Best, V., and Marrone, N. (**2010**). "Stimulus factors influencing spatial release from speech-on-speech masking," J. Acoust. Soc. Am. **128**, 1965–1978.

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

Kidd, G., Jr., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.

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

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

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 December 6, 2019).

Luke, S. G. (**2017**). "Evaluating significance in linear mixed-effects models in R," Behav. Res. Meth. **49**, 1494–1502.

Marrone, N., Mason, C. R., and Kidd, G., Jr. (**2008**). "Tuning in the spatial dimension: Evidence from a masked speech identification task," J. Acoust. Soc. Am. **124**, 1146–1158.

Moore, B. C. J. (**1998**). *Cochlear Hearing Loss* (Whurr, London).

Newman, R. S., Morini, G., Ahsan, F., and Kidd, G., Jr. (**2015**). "Linguistically-based informational masking in preschool children," J. Acoust. Soc. Am. **138**, EL93–EL98.

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

R Core Team (**2019**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, available at http://www.r-project.org/ (Last viewed December 6, 2019).

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., Ferro, D. F., Wissig, S. C., and Landau, C. A. (**2008**). "Asynchrony tolerance in the perceptual organization of speech," Psychon. Bull. Rev. **15**, 861–865.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.

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

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

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

Roberts, B., Summers, R. J., and Bailey, P. J. (**2010**). "The perceptual organization of sine-wave speech under competitive conditions," J. Acoust. Soc. Am. **128**, 804–817.

Roberts, B., Summers, R. J., and Bailey, P. J. (**2014**). "Formant-frequency variation and informational masking of speech by extraneous formants: Evidence against dynamic and speech-specific acoustical constraints," J. Exp. Psychol. Hum. Percept. Perform. **40**, 1507–1525.

Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. B **336**, 367–373.

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

Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cognit. Sci. **12**, 182–186.

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

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

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.

Thompson, E. R., Iyer, N., Simpson, B. D., Wakefield, G. H., Kieras, D. E., and Brungart, D. S. (**2015**). "Enhancing listener strategies using a payoff matrix in speech-on-speech masking experiments," J. Acoust. Soc. Am. **138**, 1297–1304.

Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.

Wang, D. L. (**2005**). "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. L. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.