

Effects of stimulus naturalness and contralateral interferers on lexical bias in consonant identification

Brian Roberts, Robert J. Summers and Peter J. Bailey

Citation: [The Journal of the Acoustical Society of America](#) **151**, 3369 (2022); doi: 10.1121/10.0011395

View online: <https://doi.org/10.1121/10.0011395>

View Table of Contents: <https://asa.scitation.org/toc/jas/151/5>

Published by the [Acoustical Society of America](#)




**Advance your science and career
as a member of the**

ACOUSTICAL SOCIETY OF AMERICA

LEARN MORE



Effects of stimulus naturalness and contralateral interferers on lexical bias in consonant identification

Brian Roberts,^{1,a)}  Robert J. Summers,¹  and Peter J. Bailey²

¹*School of Psychology, Aston University, Birmingham, B4 7ET, United Kingdom*

²*Department of Psychology, University of York, Heslington, York, YO10 5DD, United Kingdom*

ABSTRACT:

Lexical bias is the tendency to perceive an ambiguous speech sound as a phoneme completing a word; more ambiguity typically causes greater reliance on lexical knowledge. A speech sound ambiguous between /g/ and /k/ is more likely to be perceived as /g/ before /ɪft/ and as /k/ before /ɪs/. The magnitude of this difference—the Ganong shift—increases when high cognitive load limits available processing resources. The effects of stimulus naturalness and informational masking on Ganong shifts and reaction times were explored. Tokens between /gɪ/ and /kɪ/ were generated using morphing software, from which two continua were created (“giss”–“kiss” and “gift”–“kift”). In experiment 1, Ganong shifts were considerably larger for sine- than noise-vocoded versions of these continua, presumably because the spectral sparsity and unnatural timbre of the former increased cognitive load. In experiment 2, noise-vocoded stimuli were presented alone or accompanied by contralateral interferers with constant within-band amplitude envelope, or within-band envelope variation that was the same or different across bands. The latter, with its implied spectro-temporal variation, was predicted to cause the greatest cognitive load. Reaction-time measures matched this prediction; Ganong shifts showed some evidence of greater lexical bias for frequency-varying interferers, but were influenced by context effects and diminished over time.

© 2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/10.0011395>

(Received 16 November 2021; revised 27 April 2022; accepted 2 May 2022; published online 23 May 2022)

[Editor: Karen S. Helfer]

Pages: 3369–3386

I. INTRODUCTION

Speech perception involves interaction between incoming sensory data and top-down constraints, allowing lexical knowledge to influence the pre-lexical processing of acoustic information (e.g., McClelland *et al.*, 2006). Consequently, the identification of a spoken item is influenced by its context, and this effect manifests as lexical bias—a greater tendency to identify the item as a word than a non-word. This lexical bias often becomes more pronounced when the stimulus is ambiguous or degraded (e.g., Mattys *et al.*, 2012). A widely used method for exploring lexical bias, first introduced by Ganong (1980), involves creating a set of consonant-vowel (CV) syllables by progressively increasing the voice onset time (VOT) of the initial stop such that its perception changes from voiced to unvoiced. A pair of continua is then derived from this set of CVs by appending one or other of two unchanging consonantal segments, such that one continuum spans from a word to a non-word and the other from a non-word to a word. The effect of lexical status on judgments is typically greater for items around the phoneme boundary—where auditory cues are most ambiguous—than at the ends of the continuum (e.g., Ganong, 1980; Fox, 1984). For example, an initial stop ambiguous between /g/ and /k/ is more likely

to be perceived as /g/ when preceding /ɪft/, because “gift” is a word, and as /k/ when preceding /ɪs/, because “kiss” is a word, leading to a difference in the position of the phoneme boundary along the continuum (e.g., Pitt and Samuel, 1993; Mattys and Wiget, 2011). This “Ganong effect” can also be measured for other phonemic contrasts, such as in place of articulation (e.g., Gianakas and Winn, 2019). A convenient measure of the extent of lexical bias based on the full dataset, not just the boundary shift at 50% identification, is obtained by computing the area difference between the two identification functions. This measure, sometimes called the lexical identification shift (e.g., Pitt and Samuel, 1993), is referred to henceforth as the *Ganong shift*. Lexical bias has also been explored using reaction-time measures (e.g., Fox, 1984).

In a discussion of the factors that influence speech recognition in adverse conditions, Mattys *et al.* (2009) proposed a distinction between a perceptual load and a cognitive load; the former involves any alteration to the signal that results in decreased acoustic integrity (e.g., the presence of background noise in the same ear as the signal), whereas the latter arises from an increased use of limited central processing resources owing to concurrent demands on attention or memory, independent of signal degradation (e.g., the cost of the effort required to ignore a masker). The extent of lexical bias is known to be influenced by several factors. These include perceptual loads—the quality of the

^{a)}Electronic mail: b.roberts@aston.ac.uk

stimulus and its reproduction, and the quality of the sensorineural representation of the stimulus—and the cognitive load experienced by the listener when making the judgments. For example, lexical bias tends to be greater when natural speech is either low-pass filtered or noise vocoded with only eight channels (Gianakas and Winn, 2019), for cochlear-implant users than for normal-hearing listeners (Gianakas and Winn, 2019), and when listeners perform a concurrent visual search task (Mattys and Wiget, 2011) or are distracted by the induction of acute anxiety (Mattys *et al.*, 2013). The locus of the greater lexical bias associated with increased cognitive load is sub-lexical, arising from impoverished encoding of the sensory input (Mattys and Wiget, 2011; Mattys *et al.*, 2014). However, there are two factors potentially relevant to the processing load on listeners that have received less attention. The first is stimulus naturalness, of interest because of the possibility that a reduction in naturalness without much loss of intelligibility might nonetheless increase the cognitive load on the listener. The second is acoustical informational masking (IM)—e.g., the presence of a speech-like interferer in the contralateral ear may affect the listener in several ways, including an increased cognitive load. These two factors are the focus of the experiments reported here.

Burton and Blumstein (1995) investigated the effect of stimulus naturalness on lexical bias by comparing boundary shifts for stimulus continua in which the initial stop differed only in VOT with those for otherwise comparable continua in which there were covarying changes in two other acoustic dimensions characteristic of the voicing distinction in natural stop-vowel syllables—namely, the amplitude of the release burst and of the aspiration noise (Lisker and Abramson, 1964). They investigated the effect of stimulus quality by presenting these stimuli alone or embedded in white noise of the same duration (10.5-dB signal-to-noise ratio with respect to the vowel). Although the boundary shifts observed in their study were fairly small, there was some evidence that lexical bias was increased by reducing stimulus quality, but there was no indication that stimulus naturalness had any effect. However, we are not aware of any studies that have explored the effects on lexical bias of more substantial differences in stimulus naturalness—henceforth taken to mean the extent to which the stimulus sounds plausibly to have been produced by a human vocal tract—while preserving good stimulus quality and hence relatively high intelligibility. Experiment 1 reported here addressed this issue by comparing the Ganong shifts and reaction times obtained for three types of stimuli; reaction times indicate processing speed and are often measured alongside Ganong shifts (e.g., Fox, 1984; Mattys and Wiget, 2011). The continua for the reference (REF) condition were derived from natural tokens with minimal acoustic processing (see the following section—monotonization of the CV portion, morphing between the endpoints, appending the terminal consonantal portion); the other two conditions used vocoded versions of these continua with sufficient channels to limit any loss of intelligibility but which differed from

the REF condition, and markedly from one another, in naturalness owing to the properties of the carriers used.

The second issue addressed by the study reported here is the effect of acoustical IM on processing load, as revealed by lexical bias. Speech perception often takes place in listening conditions that make communication challenging (see, e.g., Bregman, 1990; Darwin, 2008; Mattys *et al.*, 2012). Interfering sounds can impair speech intelligibility through energetic masking (EM), in which the encoding of features of the target speech is degraded in the response of the auditory nerve, or through IM arising in the central auditory system. The framework proposed by Mattys *et al.* (2009) encompasses adverse listening conditions broader than those involving acoustic interferers, and so in this context EM refers to any degradation of the signal or its encoding in the auditory nerve and IM refers to any kind of cognitive load, including a concurrent task in another modality. The effects of a perceptual load arise from EM, or more usually a combination of EM and IM components, whereas those of a cognitive load arise solely from IM (Mattys *et al.*, 2009). In this framework, the less accurate encoding of sensory information that results from the EM component leads to more diffuse lexical activation and an increased reliance on acoustic detail, whereas the IM component has the opposite effect—increased reliance on lexical information—because depletion of general-purpose central resources has a greater effect on the processing of acoustic detail than of lexical information (Mattys *et al.*, 2009; see also Mattys *et al.*, 2005). The speech signal is sparse on a frequency \times time representation, and so the impact of interfering speech on intelligibility often arises mainly from IM—particularly when there is only one interfering voice, similar in level or lower than the target voice (e.g., Brungart *et al.*, 2006; Darwin, 2008).

The IM caused by contralateral speech-like interferers depends mainly on the extent and velocity of formant-frequency variation in the interferer (Roberts and Summers, 2015, 2018, 2020; Summers *et al.*, 2012). In particular, frequency variation in the extraneous formants appears to hinder the extraction or integration of information about speech articulation carried by the time-varying formant-frequency contours of the target speech, an effect known as acoustic-phonetic interference (Roberts and Summers, 2018, 2020; Summers and Roberts, 2020). In contrast, differences in fundamental frequency (F_0) between target and interfering formants have a much smaller effect on intelligibility (Summers *et al.*, 2010, 2017; see also Summers and Roberts, 2020) and the amplitude contours of the interfering formants have little or no effect (Roberts *et al.*, 2010; Summers *et al.*, 2012; Roberts and Summers, 2015). The contralateral interferers used in experiment 2 were chosen to reveal whether frequency variation in the interferer is similarly potent in its effect on lexical bias.

IM can result from failures of object formation and selection—for example, properties of the interferer may intrude into the target percept—or from capacity limitations on the resources available for information processing (see,

e.g., Shinn-Cunningham, 2008). These aspects of IM can be regarded as *corrupting* and *disrupting* effects, respectively (e.g., Roberts *et al.*, 2014; Roberts and Summers, 2018). An example of IM deriving from corruption of target processing was described recently by Roberts *et al.* (2021): predictable changes in judgments of place of articulation for the initial consonant of CV syllables occurred when contralateral interferers (sine-wave analogues of F_2 with non-matching initial transitions) were present, as a result of mandatory dichotic integration of acoustic-phonetic information (see also Porter and Whittaker, 1980). The contribution of non-specific disrupting effects to the IM of speech is less well characterized and was the main concern of experiment 2. To our knowledge, only one study has even attempted to measure lexical/acoustic bias when target speech and masker were presented dichotically (Mattys *et al.*, 2009). A possible contribution from the disrupting component of IM is suggested by the close parallel between the importance of formant-frequency variation in a speech-like interferer for the IM it produces and a type of cross-modal interference known as the irrelevant sound effect (ISE), in which serial recall of visually presented digits or words is impaired by an acoustic distractor (Jones and Macken, 1993; for a review, see Ellermeier and Zimmer, 2014). Notably, the distractor must involve frequency change to produce the ISE—amplitude change alone is not sufficient (Tremblay and Jones, 1999)—and the ISE is usually greatest for distractors that involve complex spectro-temporal change, particularly speech-like sounds (e.g., Viswanathan *et al.*, 2014; Dorsi *et al.*, 2018).

A cross-modal effect like the ISE clearly cannot be based on the intrusion of specific properties of the masker into the target percept. Experiment 2 was designed to investigate the possibility that IM of speech by acoustic interferers involves not only a specific corrupting component (Roberts *et al.*, 2021) but also a non-specific disrupting component. This was done by measuring changes in lexical bias and reaction time caused by listening to speech accompanied by different kinds of contralateral interferer. It was hypothesized that any increase in cognitive load would arise primarily from formant-frequency change in the interferer, such that only interferers involving spectro-temporal change would increase lexical bias (larger Ganong shifts) and slow reaction times. In summary, the experiments reported here were intended to explore the extent to which lexical bias and consonant identification reaction times are influenced by stimulus naturalness, using syllables subject either to minimal processing or high-resolution vocoding with different carriers (experiment 1), and by acoustical IM in circumstances where attention to the masker is not required, using noise-vocoded syllables accompanied by contralateral interferers (experiment 2).

II. EXPERIMENT 1

In this experiment, identification functions, reaction-time profiles, Ganong shifts, and overall reaction times were

measured using word to non-word and non-word to word VOT continua presented without accompanying sounds. Measuring Ganong shifts and reaction times together is useful because the former depend specifically on the extent of lexical bias, whereas any manipulation that slows overall processing will affect the latter, irrespective of the direction of change in the relative balance of acoustic and lexical influences. These measures were compared across conditions that differed mainly in stimulus naturalness but also, to a lesser extent, in stimulus quality and therefore intelligibility. One condition involved minimal processing of natural syllables and used tokens of high quality and naturalness (REF); the other two conditions involved high-resolution vocoding of those syllables, affecting naturalness while minimizing any loss of intelligibility.

Two types of carrier were used to create modulated-noise-band (MNB) and modulated-sine-band (MSB) versions of the paired VOT continua. MSB vocoding produced stimuli that were considerably sparser in spectrum than the natural tokens from which they were derived and which had a highly unnatural timbre. In contrast, MNB vocoding produced stimuli with a similar spectral density to natural speech and a relatively familiar timbre, akin to whispering. No acoustic interferers were used in this experiment, but in terms of the framework for adverse listening conditions proposed by Mattys *et al.* (2009), the perceptual load associated with high-resolution vocoding should have an EM component (here, signal degradation) associated with any loss of target intelligibility and an IM component (cognitive load) associated with the extent of change in stimulus naturalness. The EM component should increase the reliance on acoustic detail but should be small given that—at least for word identification in sentences—there is no intelligibility cost for MNB and MSB vocoding relative to natural speech when the number of channels is ≥ 16 (Villard and Kidd, 2021). The IM component should act to increase the reliance on lexical information and was predicted to be considerably larger for the MSB than for the MNB stimuli. The direction of change and overall size of the Ganong shifts should depend on the balance of the two influences, suggesting here little change in lexical bias relative to the REF condition for the MNB stimuli, but a substantial increase for the MSB stimuli.

A. Method

1. Listeners

All listeners were students or members of staff at Aston University and received either course credits or payment for taking part. They were first tested using a screening audiometer (Interacoustics AS208, Assens, Denmark) to ensure that their audiometric thresholds at 0.5, 1, 2, and 4 kHz did not exceed 20 dB hearing level in either ear. All listeners who passed the audiometric screening took part in a training session designed to familiarize them with the task and stimuli (see Sec. II A 3). Thirty-six listeners (five males) successfully completed the experiment (mean age = 23.6 years,

range = 18.3–49.1). All listeners 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 syllables “gift,” “kift,” “giss,” and “kiss” were spoken by a British female talker; several tokens of each were recorded. The best exemplars of /gI/ and /kI/ were excised from these recordings and monotonized ($F_0 = 180$ Hz) using the STRAIGHT software package (Kawahara *et al.*, 1999). The plosives /g/ and /k/ share a velar place of articulation but differ in VOT (shorter vs longer) and the extent of aspiration (less vs more) in syllable-initial position. An eight-item continuum spanning from /gI/ to /kI/ was created by morphing between these endpoints using STRAIGHT, giving a more naturalistic progression of voicing cues than would be possible using manual editing. Two continua were created from this continuum, one changing from word to non-word (“gift” to “kift”) and the other from non-word to word (“giss” to “kiss”); this was done by excising a single token of /ft/ and a single token of /s/ from the initial recordings and splicing the same token, as appropriate, on to each member of the continuum. These continua—which, apart from being monotonized, had a natural-sounding timbre and quality—constituted the REF condition.

The other conditions were created by passing these continua through a noise vocoder (Shannon *et al.*, 1995; Roberts *et al.*, 2011) and a sine vocoder (Hill *et al.*, 1968; Dorman *et al.*, 1997), written in PRAAT (Boersma and Weenink, 2016), to create modulated-noise-band and modulated-sine-band versions, respectively. Higher resolution was used for the MSB stimuli (32 channels) than for the MNB stimuli (16 channels) owing to their relative spectral sparsity, which can impair intelligibility relative to otherwise matched MNB stimuli (e.g., Souza and Rosen, 2009; Rosen *et al.*, 2015). The center frequencies of the constituent bands of the vocoded stimuli were equally spaced on a log frequency scale (i.e., equal spacing in semitones) over the range from 0.2 to 8.0 kHz. To generate them, each REF stimulus was first filtered into the required number of bands using a 16th-order Butterworth filter (96 dB/octave roll-off) and the envelope of each band was extracted using a Kaiser window acting as an ~50 Hz low-pass filter, implemented by the “To Intensity...” function in PRAAT. From informal listening, the best vocoding was achieved by padding the beginning of the original signal with 40 ms of silence, so that the initial plosive burst was better represented in the envelope. The envelope of each band was used to modulate an appropriate carrier; this was a Gaussian noise source with corresponding lower and upper cut-off frequencies for the MNB stimuli and a sine tone with the same center frequency (i.e., geometric mean of the band cut-off frequencies) for the MSB stimuli. Each band was scaled to have the same root mean square (rms) level as that of the corresponding band in the REF stimulus and the bands were summed together.

Finally, all items in the three conditions were matched to the same rms level. Figures 1 and 2 show the sound waves and spectrograms illustrating the endpoint stimuli for the “gift”–“kift” and “giss”–“kiss” continua, respectively, in each of the three conditions. See the supplementary material for all stimuli and the parameters used to create them.¹

3. Procedure

During testing, listeners were seated in front of a computer screen and a keyboard in a double-walled sound-attenuating chamber (Industrial Acoustics 1201 A, Winchester, UK). The experiment began with a training session, which used only the endpoint stimuli from the two continua and was intended to ensure that these stimuli were reliably classified as /g/ and /k/ for positions 1 and 8, respectively, in each of the three conditions. Stimuli were blocked by condition but mixed by continuum within blocks, and blocks were always run in the order REF, MNB, and MSB. Mixing by continuum typically produces more reliable Ganong shifts (Pitt and Samuel, 1993). To help familiarize listeners with the three stimulus types, each block began by presenting a sentence processed to have the same acoustic source properties as those of the syllables in the subsequent trials. Listeners then heard eleven repetitions of the four endpoint stimuli from the appropriate condition, presented in random order. Each trial was initiated by pressing the “space bar,” 500 ms after which the token was presented diotically. Listeners were asked to identify the initial consonant as quickly as possible while maintaining accuracy, responding by pressing “G” or “K” on the keyboard, as appropriate. Reaction times for these judgments do not seem to depend on whether listeners are asked to identify the initial consonant or the whole syllable (Miller and Dexter, 1988). No feedback on listeners’ responses was given. The complete training session comprised 132 trials (11 repetitions \times 3 conditions \times 2 continua \times 2 endpoints). The results for the first repetition were discarded; scores of $\geq 9/10$ for the REF condition and $\geq 7/10$ for the MNB and MSB conditions were required to proceed to the main session. Listeners who did not pass the first time were allowed to repeat the training session once.

In the main session, listeners completed three blocks of trials (one each for the REF, MNB, and MSB stimuli); unlike in the training session, block order was counterbalanced across listeners. Each block comprised eleven repetitions of all eight members of the two continua under diotic presentation, for which listeners were again asked to identify the initial consonant. A new randomization of trial order was used for each repetition and listener. Reaction times were recorded from the onset of the stimulus file for the REF stimuli (Fox, 1984), but from 40 ms later for the MNB and MSB stimuli to compensate for the 40-ms padding used to optimize the vocoding of the initial stop (see Sec. II A 2). Given that these trials included stimuli with intermediate and ambiguous cues to the voicing category, the first repetition (16 trials) was treated as practice and the results were

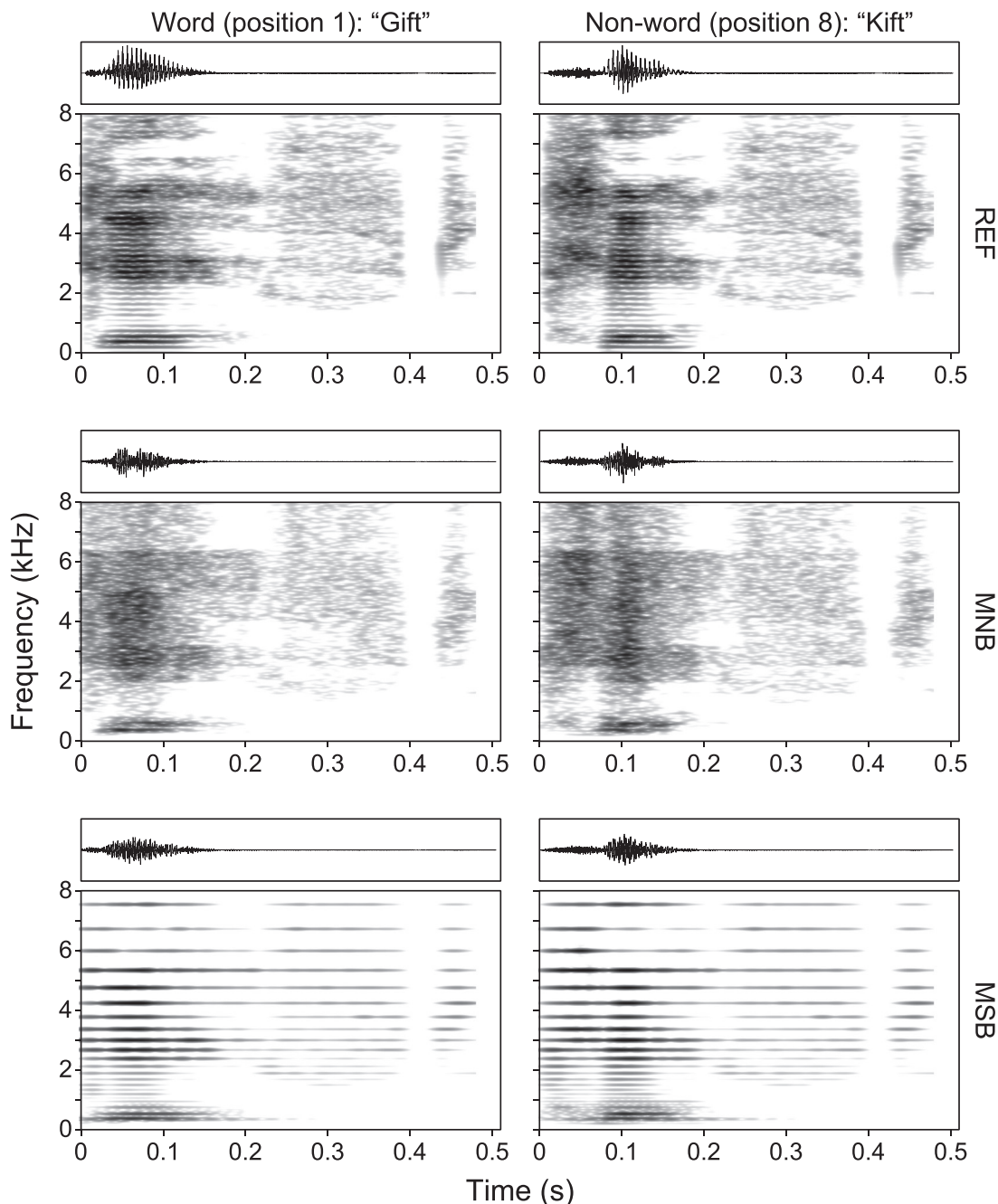


FIG. 1. Stimuli for experiment 1—exemplars from the “gift”–“kift” continuum (word to non-word) for the REF, noise-vocoded (MNB), and sine-vocoded (MSB) conditions. The upper section of each panel shows the sound wave and the lower section shows the corresponding narrowband spectrogram. For each panel, the stimulus is shown aligned to the appropriate start time for the reaction-time measure (see main text for details), as indicated by the 0-ms mark on the x axis. The left- and right-hand panels illustrate the targets for continuum positions 1 ([g]) and 8 ([k]), respectively. Descending rows of panels illustrate the stimuli for the REF, MNB, and MSB conditions, respectively.

discarded, leaving ten repetitions. Hence, the identification functions and reaction time measures obtained for each listener were based on 160 trials per condition. Two listeners did not meet the inclusion criterion of a score $\geq 7/10$ for each endpoint in every condition, and so they were excluded and replaced. The whole experiment typically took ~ 70 min to complete; listeners were free to take short breaks between the component parts if they wished.

All stimuli were presented at a sampling rate of 44.1 kHz and with 10-ms raised-cosine onset and offset ramps. They

were played at 16-bit resolution over Sennheiser HD 480-13II earphones (Sennheiser, Hannover, Germany) *via* a Sound Blaster X-Fi HD sound card (Creative Technology Ltd., Singapore), a pair of programmable attenuators (Tucker-Davis Technologies TDT PA5, Alachua, FL), and a headphone buffer (TDT HB7, 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 (type 4153). All stimuli were presented at 72 dB sound pressure level (SPL).

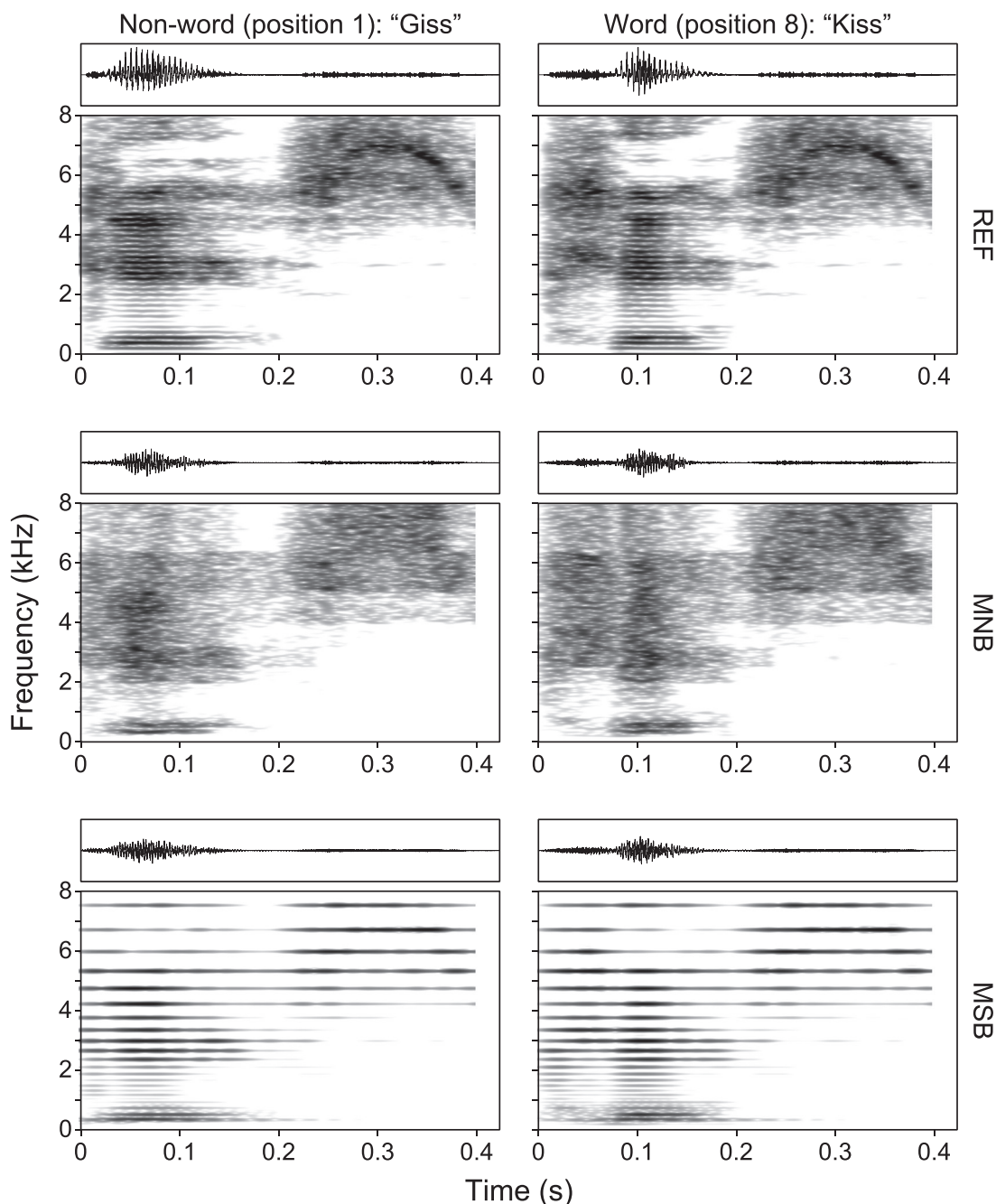


FIG. 2. Stimuli for experiment 1—exemplars from the “giss”–“kiss” continuum (non-word to word) for the REF, noise-vocoded (MNB), and sine-vocoded (MSB) conditions. Otherwise as for Fig. 1.

4. Data analysis and availability

Identification functions were computed for each listener in each condition and changes in these functions were used to assess changes in judgments of voicing for the initial plosive across conditions. Our measure of lexical bias, the Ganong shift, was calculated as the percentage of the total area falling between the identification functions for “giss”–“kiss” and “gift–kift” (see, e.g., Pitt and Samuel, 1993; Mattys and Wiget, 2011), where the area of each function corresponds to the mean proportion of /k/ responses when averaged across all positions in that continuum. Given

that corresponding positions in the two continua were acoustically identical for the initial consonant-vowel portion, differences in identification between continua must arise from top-down influences. Positive values of the Ganong shift indicated that the listener was biased towards hearing the lexical item (i.e., “gift” or “kiss”) and negative values indicated that the listener was biased towards hearing the non-lexical item (i.e., “kift” or “giss”). Differences in the size of the Ganong shift across conditions are expressed as changes in percentage points (% pts).

Mattys and Wiget (2011) noted that the relationship between response latency and the size of the Ganong shift is

both complex and unclear. For example, some studies have reported an association between faster reaction times and smaller Ganong shifts (e.g., Fox, 1984; Miller and Dexter, 1988), whereas others have found the opposite association (e.g., McQueen, 1991; Pitt and Samuel, 1993, 2006). Rather than dividing the observed reaction times into slow, medium, and fast partitions, our approach here has been to focus on average reaction times as a general measure of overall processing speed. For each listener in every condition, the mean of median reaction times was computed for each position in each continuum; this measure protects against the effects of outliers. When combining reaction time measures across listeners and across continuum positions, the mean of the mean of medians was used.

All statistical analyses were computed using R 4.1.0 (R Core Team, 2020) and the *ez* analysis package (Lawrence, 2016). The measures of effect size reported here for repeated-measures analysis of variance (ANOVA) are eta squared (η^2) and partial eta squared (η_p^2). All pairwise comparisons (two tailed) were computed using the restricted least-significant-difference test (Snedecor and Cochran, 1967; Keppel and Wickens, 2004). The research data underlying this publication are available online from a repository hosted by Aston University.²

B. Results

Figure 3 shows the results for identifying the initial consonant; the outcomes for each condition are shown in descending pairs of panels (REF, MNB, and MSB). The left- and right-hand panels show the mean proportions of /k/ responses and mean reaction times, respectively, for each position along the two continua (with inter-subject standard errors). The identification functions and reaction-time profiles for the non-word to word continuum and the word to non-word continuum are indicated by dotted lines (filled square symbols) and dashed lines (open circle symbols), respectively. Each pair of panels also includes the mean Ganong shift—i.e., the difference in area between the two identification functions—and the overall mean reaction time for all responses to the two continua. Figure 4 summarizes the differences in Ganong shift and overall reaction time across conditions.

Consider first the identification functions. Listeners produced clear and systematic patterns of judgments for both continua in all three conditions, in almost all cases progressing monotonically from few to mainly /k/ responses as position number increased. Since the initial CV portion of each syllable was acoustically identical for each corresponding position in a given condition, differences in judgments of the initial consonant between the two continua must be a consequence of the final portion of the syllable—presumably the lexical context it provided. As expected, listeners were almost always more likely to judge ambiguous tokens as starting with whichever consonant specified a word; the regions of difference between the two identification

functions are shown shaded in gray whenever mean judgments were biased towards the word items. One-way ANOVA showed that the effect of condition on the extent of this lexical bias was highly significant [$F(2,70) = 14.435$, $p < 0.001$, $\eta^2 = 0.292$].

The Ganong shifts observed were consistent with the range of means typically reported in studies using this metric (~5%–30%). Mean Ganong shifts for the MSB stimuli (15.1%) were larger than those for the REF (9.9%) or MNB stimuli (6.7%); these differences were highly significant [REF vs MSB: difference = 5.2% pts, $t(35) = 3.719$, $p = 0.001$; MNB vs MSB: difference = 8.4% pts, $t(35) = 4.585$, $p < 0.001$]. The stronger lexical bias observed for the MSB stimuli relative to the MNB stimuli is noteworthy given the relatively small differences in slope between the corresponding identification functions and the greater number of channels used to synthesize the former (32 vs 16), which (other factors being equal) might have been expected to increase the fidelity of the acoustic-phonetic information represented. Somewhat contrary to expectations, the mean Ganong shift was significantly smaller for the MNB stimuli than for the REF stimuli [difference = 3.3% pts, $t(35) = 2.199$, $p = 0.035$], suggesting that there was a greater reliance on acoustic detail when listeners judged the MNB stimuli. Relative to the REF and MNB stimuli, responses to MSB stimuli were biased towards /g/. This is apparent in the rightward shift of the identification functions for the MSB stimuli and is also reflected in the change in overall proportion of /k/ responses across conditions (REF = 46.3%, MNB = 44.4%, and MSB = 41.1%). Finally, there is some evidence of greater stimulus ambiguity for both types of vocoded stimuli, as indicated by the greater number of items for which an appreciable lexical bias was observed (five positions for the MSB and MNB stimuli, three for the REF stimuli). Note, however, that this was associated with larger Ganong shifts only for the MSB stimuli.

Stimulus type (condition) also had a significant effect on overall reaction times [$F(2,70) = 6.273$, $p = 0.003$, $\eta^2 = 0.152$], but showed a different pattern from its effect on the size of the Ganong shifts. Mean reaction times were longer for MNB (991.5 ms) and MSB (1004.2 ms) stimuli relative to the REF condition (947.4 ms); these differences were significant [REF vs MNB: 44.0 ms, $t(35) = 2.512$, $p = 0.034$; REF vs MSB: 56.8 ms, $t(35) = 2.914$, $p = 0.019$]. The small difference between the MNB and MSB conditions was not significant [12.8 ms, $t(35) = 1.004$, $p = 0.322$]. Inspection of the reaction-time profiles across continuum positions shows two patterns that merit consideration. First, the mean of median values per listener was typically longer for the most ambiguous stimuli (i.e., tokens obtaining in the region of 50% /k/ responses on the identification function). This pattern was clearest in the REF condition, for which the positions corresponding to the longest reaction times differed between the two continua in a way concordant with the lexical bias shown in the identification functions. Second, the other striking aspect of the reaction-time profiles is that,

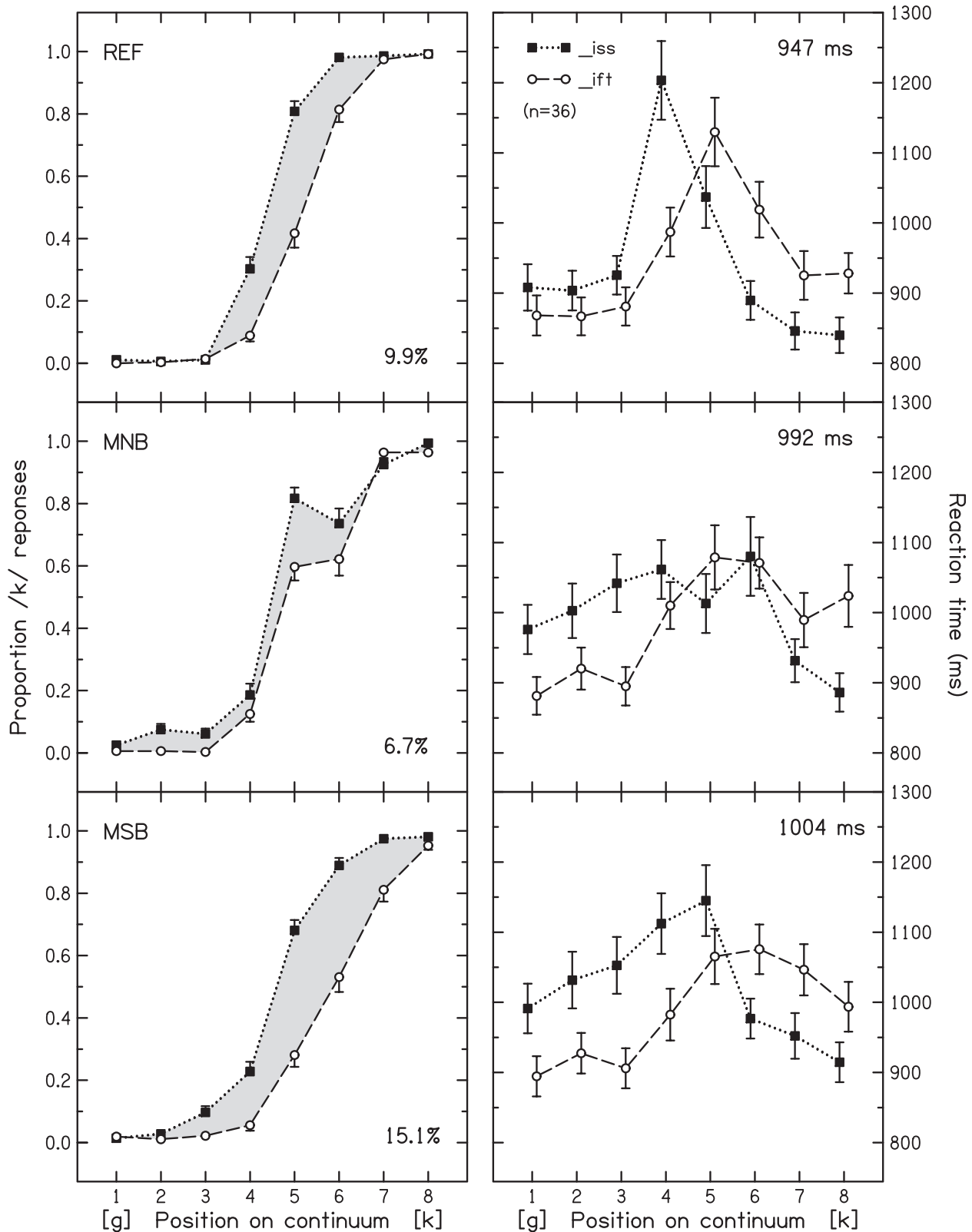


FIG. 3. Results for experiment 1—identification functions and reaction-time profiles for the two continua in each condition. Descending rows show the results for the REF, noise-vocoded (MNB), and sine-vocoded (MSB) conditions, respectively. Each left-hand panel shows the mean proportion of /k/ responses and the inter-subject standard error ($n = 36$) for each position along the non-word to word continuum (“giss”–“kiss,” filled square symbols and dotted lines) and the word to non-word continuum (“gift”–“kift,” open circle symbols and dashed lines). The Ganong shift corresponds to the difference between the two functions as a percentage of the total area and is summarized in the bottom-right of each panel; this area is shaded in gray wherever mean judgments were biased towards the word items, which was the case in almost all instances. Each right-hand panel shows the corresponding mean of median reaction times and inter-subject standard errors for each position on the two continua. For clarity, the data for the two continua are shown slightly offset along the abscissa. The overall mean reaction time is summarized in the top-right of each panel.

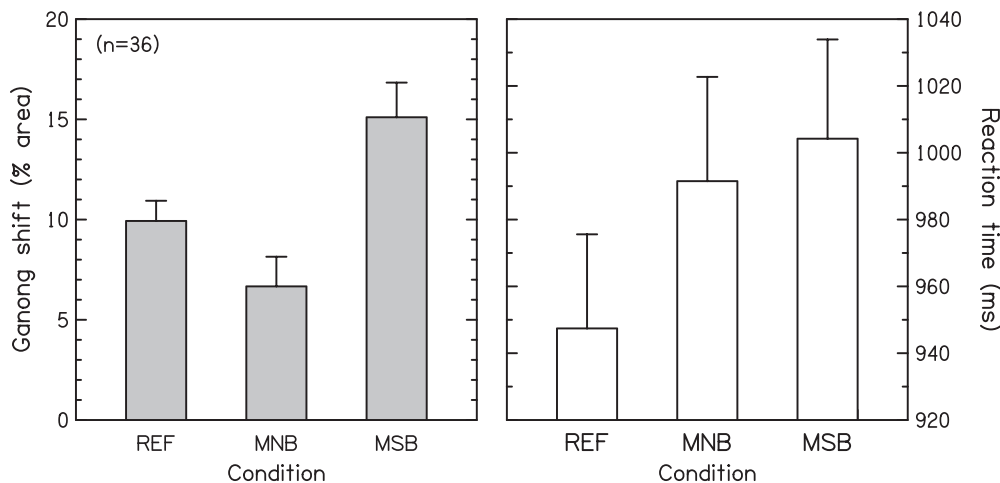


FIG. 4. Results for experiment 1—summary of Ganong shifts and overall reaction times for the three conditions. Mean Ganong shifts and inter-subject standard errors ($n = 36$) are shown for each condition (bottom axis) in the left-hand panel (gray bars); the corresponding overall mean reaction times and inter-subject standard errors are shown in the right-hand panel (open bars).

towards the endpoints of the continua, reaction times were longer for the non-lexical than for the lexical response, producing a characteristic crossover pattern that can be seen in all three conditions.

To explore whether the effects of stimulus ambiguity and lexical vs non-lexical decisions were significant, we began by conducting a three-factor repeated-measures ANOVA—condition (stimulus type, T) \times continuum (C) \times position (P)—on the median reaction-time data; the statistical outcomes are presented in Table I. Except for the main effect of continuum ($p = 0.104$), all main effects and interactions were significant (range: $p = 0.008$ to $p < 0.001$), and so for each condition pairwise comparisons were used to examine these contrasts. The effects of ambiguity were explored by comparing reaction times when averaged across the inner four positions on the continua (positions 3–6, more ambiguous stimuli) with those when averaged across the outer four positions (positions 1–2 and 7–8, less ambiguous stimuli). Reaction times were significantly longer for the more ambiguous inner positions in all three conditions (REF: mean difference = 107.1 ms, $t(35) = 8.294$, $p < 0.001$; MNB: mean difference = 78.6 ms, $t(35) = 6.099$, $p < 0.001$; MSB: mean difference = 72.1 ms, $t(35) = 7.893$, $p < 0.001$). The effects of lexical vs non-lexical decisions were explored

by collapsing the data into word (“_ift” and “g” or “_iss” and “k”) and non-word responses (“_ift” and “k” or “_iss” and “g”) and comparing them. Reaction times were significantly longer for non-word than for word responses in all three conditions (REF: mean difference = 44.6 ms, $t(35) = 5.224$, $p < 0.001$; MNB: mean difference = 77.5 ms, $t(35) = 5.457$, $p < 0.001$; MSB: mean difference = 74.2 ms, $t(35) = 4.205$, $p < 0.001$).

C. Discussion

Most notably, the effect of stimulus type on the size of the Ganong shifts showed a different pattern from its effect on the overall reaction times. The latter were longer for the vocoded stimuli than for the reference case, but were similar for both types of vocoding, indicating a similar slowing of processing speed and implying a similar reduction in stimulus quality caused by the vocoding (despite the use of high resolution processing with ≥ 16 channels). The different pattern found for the Ganong shifts suggests that they were heavily influenced by the extent of spectral sparsity of the vocoded stimuli—MSB carriers are relatively sparse compared with MNB carriers—and the associated changes in stimulus naturalness. Note also that the sine bands were equally spaced on a log frequency scale and hence were inharmonic, a pattern not characteristic of natural speech. Presumably, listeners needed actively to ignore the unusual carrier for the MSB stimuli, requiring a different allocation of attention and increasing the cognitive load they experienced, leading to greater lexical bias. If lower stimulus naturalness had instead made it easier to ignore the carrier, a reduced rather than an increased Ganong shift should have been observed. The origin of the small but significant fall in the Ganong shift for the MNB condition (relative to the REF case) is less clear, but it may be related to the finding that listeners allocate more attention to acoustic detail when listening to speech in noise (Mattys *et al.*, 2009), and so may be the result of the MNB stimuli being perceived as if they

TABLE I. Results for experiment 1—reaction times. Effects of condition (stimulus type), continuum, and position on the median reaction times of the responses to the initial consonant. Summary of the three-way repeated-measures ANOVA; all significant terms are shown in bold.

Factor	<i>df</i>	<i>F</i>	<i>p</i>	η_p^2
Condition: Stimulus type (T)	(2, 70)	6.273	0.003	0.152
Continuum (C)	(1, 35)	2.790	0.104	0.074
Position (P)	(7, 245)	34.727	<0.001	0.498
T \times C	(2, 70)	5.208	0.008	0.130
T \times P	(14, 490)	5.989	<0.001	0.146
C \times P	(7, 245)	29.726	<0.001	0.459
T \times C \times P	(14, 490)	4.161	<0.001	0.106

were speech embedded in noise. The lower likelihood of /k/ responses for the MSB stimuli probably results from the absence of noise in the representation of aspiration in the target syllable.

At this point, it should be acknowledged that an alternative explanation for the observed pattern of Ganong shifts, of less theoretical interest than the relationship between stimulus naturalness and cognitive load, cannot be ruled out without further research. This is the possibility that the MNB vocoding led to relatively poorer intelligibility of the terminal parts of the syllable (/ft/ and /s/), and hence less scope for lexical compensation, whereas the MSB vocoding led to relatively poorer intelligibility of the initial stop (/g/-/k/ continuum), and hence more scope for lexical compensation. This limitation could be addressed in future work by exploring the extent to which the outcomes for this experiment generalize across a wide range of pairs of word to non-word and non-word to word continua, vocoded in similar ways.

The finding that reaction times were longer for the most ambiguous tokens is in accord with the results of previous studies (e.g., Fox, 1984). The high peaks in the reaction-time profile for the REF condition, relative to the MNB and MSB conditions, may be a consequence of the smaller proportion of ambiguous stimuli in the REF condition. Similarly, the finding that reaction times were longer for the non-lexical than for the lexical response is consistent with the results of auditory lexical decision tasks, which indicate that non-words usually yield longer reaction times than otherwise comparable words (e.g., Goh *et al.*, 2009; see also Connine and Clifton, 1987).

III. EXPERIMENT 2

In this experiment, the effects of contralateral interferers on initial consonant identification were explored in a context where differences in acoustic source characteristics—and hence in stimulus naturalness—were controlled by using the same form of vocoding for the target syllables and interferers. MNB stimuli were used, rather than MSB, because the former sounded more like natural speech and because the latter in experiment 1 were associated with Ganong shifts more than twice as large, and so potentially offered little headroom for any additional lexical bias in the presence of interferers to be manifest. Dichotic presentation was used such that any masking caused by the different interferers, when present, would be informational rather than energetic. Listeners were not required to attend the interferers.

There are limits on the ability to listen with independent ears (e.g., Gallun *et al.*, 2007) and so Ganong shifts and overall reaction times were compared for monaural target syllables when presented either alone or in the presence of contralateral interferers intended to cause different amounts of IM. It was predicted that contralateral interferers involving spectro-temporal variation would cause more IM (Roberts *et al.*, 2010, 2014, 2021; Roberts and Summers 2015, 2018, 2020; Summers *et al.*, 2012), and that the

consequent depletion of general-purpose central resources would have a greater effect on the processing of acoustic detail than of lexical information, leading to an increased reliance on the latter. This outcome would be manifest as increased Ganong shifts and slower reaction times when the target syllables were accompanied by frequency-varying interferers, relative to interferers without such variation. The effect of spatial uncertainty on these measures was also explored because greater spatial uncertainty typically increases IM (see, e.g., Kidd *et al.*, 2008).

A. Method

Except where described, the same method was used for experiment 1. Thirty listeners (five males) passed the training and successfully completed the experiment (mean age = 21.5 years, range = 18.3–36.5); two of these listeners also took part in experiment 1. In experiment 2, only MNB versions of the syllables were used and they were presented monaurally, either always in the left ear (fixed-ear targets) or with spatial uncertainty (random-ear targets). Targets were presented alone in condition 1 (C1), but in the other conditions, they were accompanied by one of three types of contralateral interferer—those with a constant amplitude envelope (C2), or those whose within-band envelope of amplitude variation was either the same across bands (C3) or different (C4). Stimuli were blocked by spatial cue (fixed- vs random-ear targets) but mixed within blocks by interference condition as well as by continuum.

Interferers were derived from the set of envelopes defining the corresponding noise-vocoded targets; independent samples of Gaussian noise were used for corresponding channels. In C4, each interferer was created by taking the set of amplitude envelopes defining the 16 bands of the target and applying them differently to the 16 bands of the interferer by swapping the top and bottom channel envelopes, then the next-to-top and next-to-bottom envelopes, and so forth. Finally, the rms level of each band was matched to that of the corresponding band of the target. Hence, like the target syllables, the interferers used in C4 involved implied spectro-temporal variation. In C2 and C3, each interferer was created by applying the same amplitude envelope to each channel in the interferer and then rms-matching that channel to its counterpart in the target. The amplitude envelope for each channel was set to be constant in C2; 120-ms linear onset and offset ramps were used, a duration chosen roughly to match the average time to reach peak amplitude from the onset of the target syllable. Using a significantly shorter onset ramp would have had the unwanted consequence of there being more energy in this interferer during the critical initial stop than for the other types of interferer. The amplitude envelope for each channel in C3 was derived from the wideband amplitude envelope for the C4 interferer. Hence, the interferers in C2 had a trapezoidal amplitude envelope and those in C3 had amplitude variation but without spectro-temporal variation. In C2–C4,

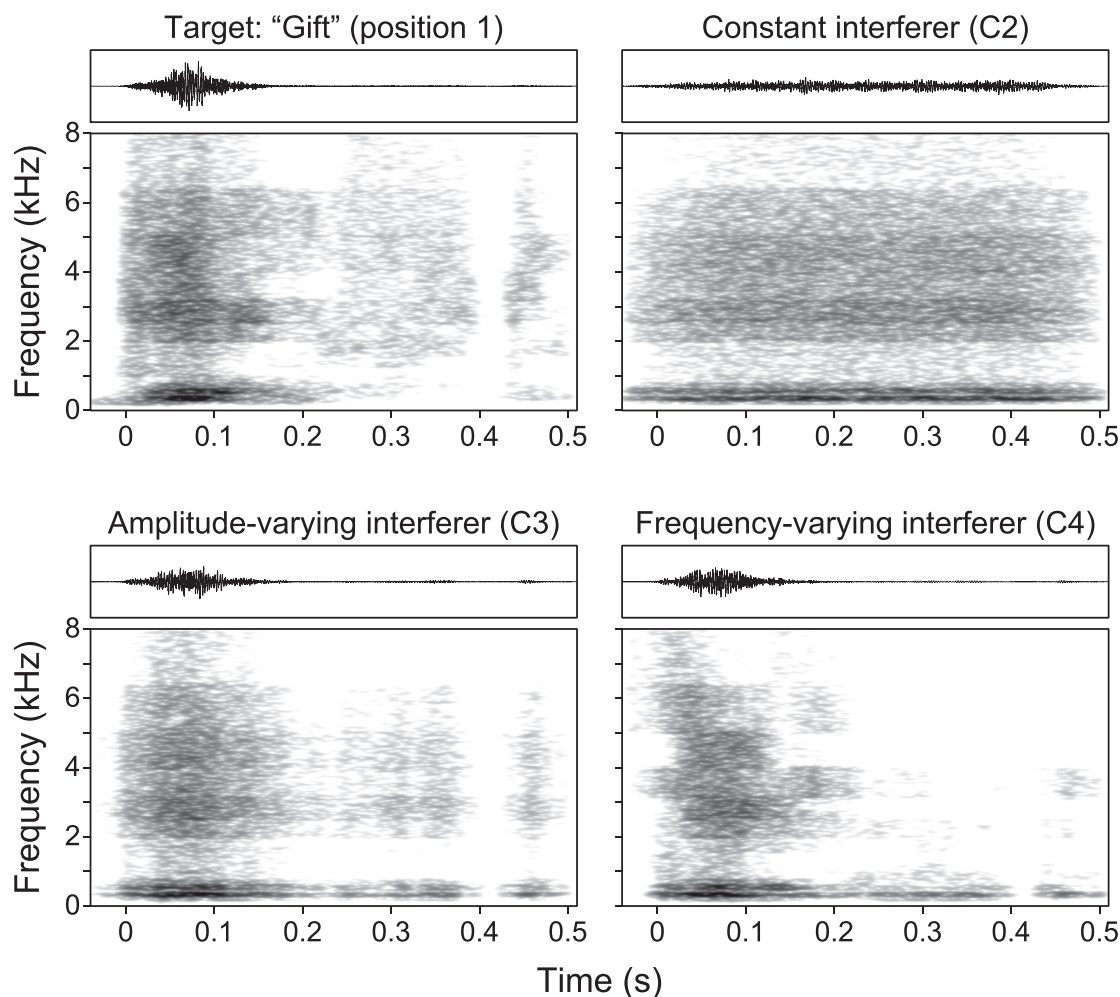


FIG. 5. Stimuli for experiment 2—illustrative examples of interferers, derived from the noise-vocoded stimulus “gift” (position 1 on word to non-word continuum). The upper part of each panel shows the sound wave and the lower part shows the corresponding narrowband spectrogram. For each panel, the 0-ms mark on the x axis indicates the start time used for the reaction-time measure (see main text for details). The top-left panel shows the target stimulus and the other panels show the three types of interferer derived from it in conditions C2 (top right), C3 (bottom left), and C4 (bottom right). The amplitude envelope for each band of the interferer was constant (C2), time-varying but the same across bands (C3), or time-varying and different across bands (C4); see main text for details of how these interferers were created. Note that only one of the three types of interferer (C4, bottom right) involved spectro-temporal variation.

the overall rms level of the interferer was always matched to that of the corresponding target. Figure 5 shows the sound waves and spectrograms illustrating an example syllable and the three types of interferer derived from it. A presentation level of 72 dB SPL in the target ear was used for all conditions involving contralateral interferers (i.e., C2–C4), but this was raised by 3 dB in C1 roughly to account for the reduced loudness arising from the purely monaural stimulation.

The experiment comprised two blocks, run on different days and each lasting ~60 min. Whether the fixed- or random-ear targets were tested first was counterbalanced across listeners. Each block began with a short training session similar in form to that used for the MNB stimuli in experiment 1 (44 trials); no interferers were present and the spatial configuration matched that used in the main session. Listeners were advised that, on any given trial in the main session, the syllable would be presented in one ear and that it may or may not be accompanied in the other ear by an

interfering sound; they were asked to try to focus their attention on the syllable, ignoring the interferer as best they could when it was present. To control for the possibility in this experiment that the task-irrelevant ear of presentation for the target stimulus might induce the Simon effect (Simon and Rudell, 1967), in which responses are quicker and more accurate when stimulus and response are spatially congruent, the layout of the keyboard was modified such that the response key for /g/ was directly above that for /k/. Each main session comprised eleven repetitions of all members of the two continua across all four interference conditions. As before, the first repetition (64 trials) was treated as practice and the results were discarded, leaving ten repetitions. Hence, in each block, the identification functions and reaction time measures obtained for each listener were based on 160 trials per interference condition (i.e., 640 trials in total). One listener did not meet the inclusion criterion (score $\geq 7/10$ for all endpoints) and so was excluded and replaced.

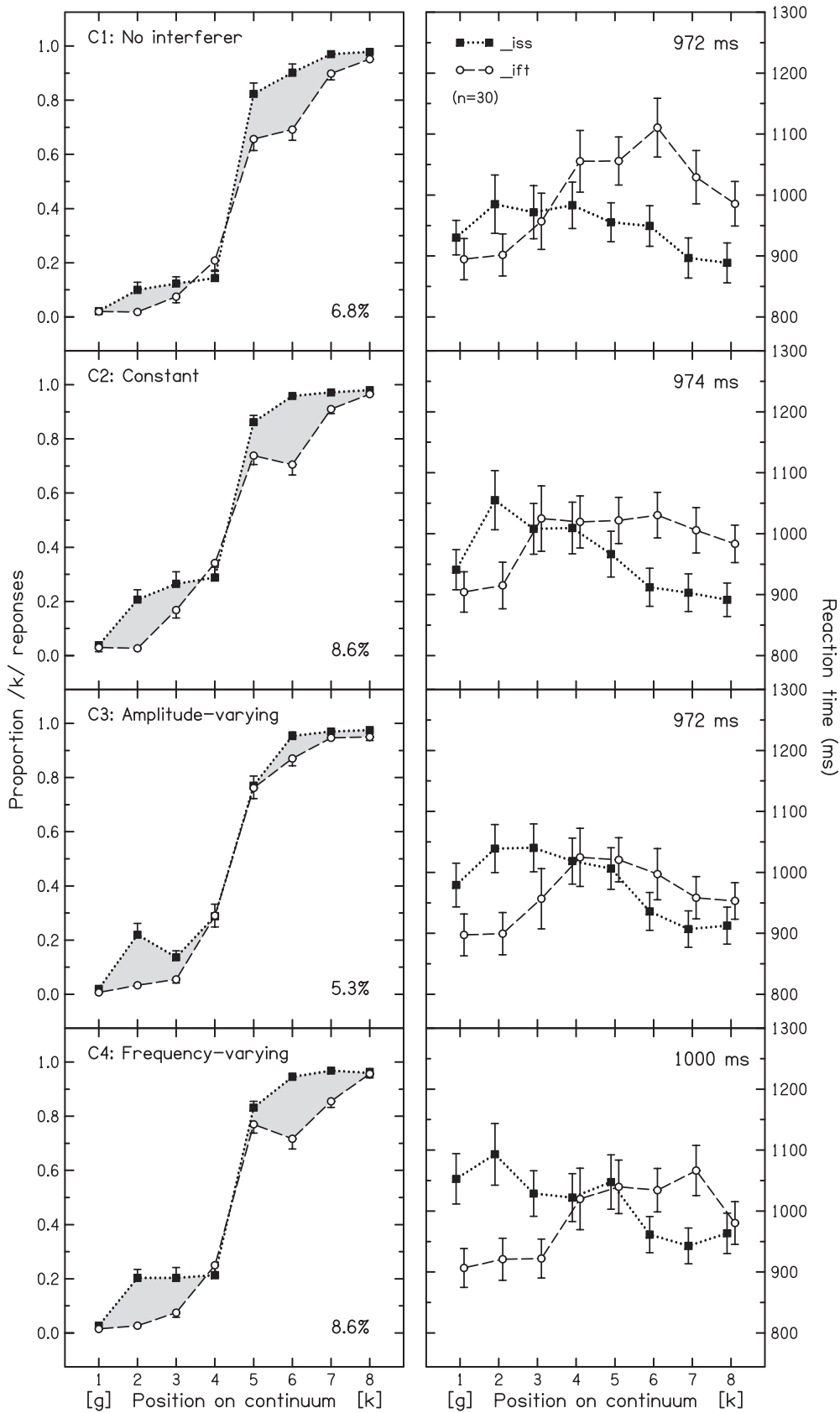


FIG. 6. Results for experiment 2—identification functions and reaction-time profiles for the two continua in each condition (C1–C4, in descending rows). The results ($n = 30$) have been collapsed across spatial cue (fixed- vs random-ear presentation of targets) because this factor had neither a significant main effect nor an interaction with condition for either measure. The four conditions correspond to the target-only case (C1) and the target plus contralateral interferer cases (C2–C4), for which the amplitude envelope for each band of the interferer was constant (C2), time-varying but the same (C3), or time-varying and different (C4). Otherwise as for Fig. 3.

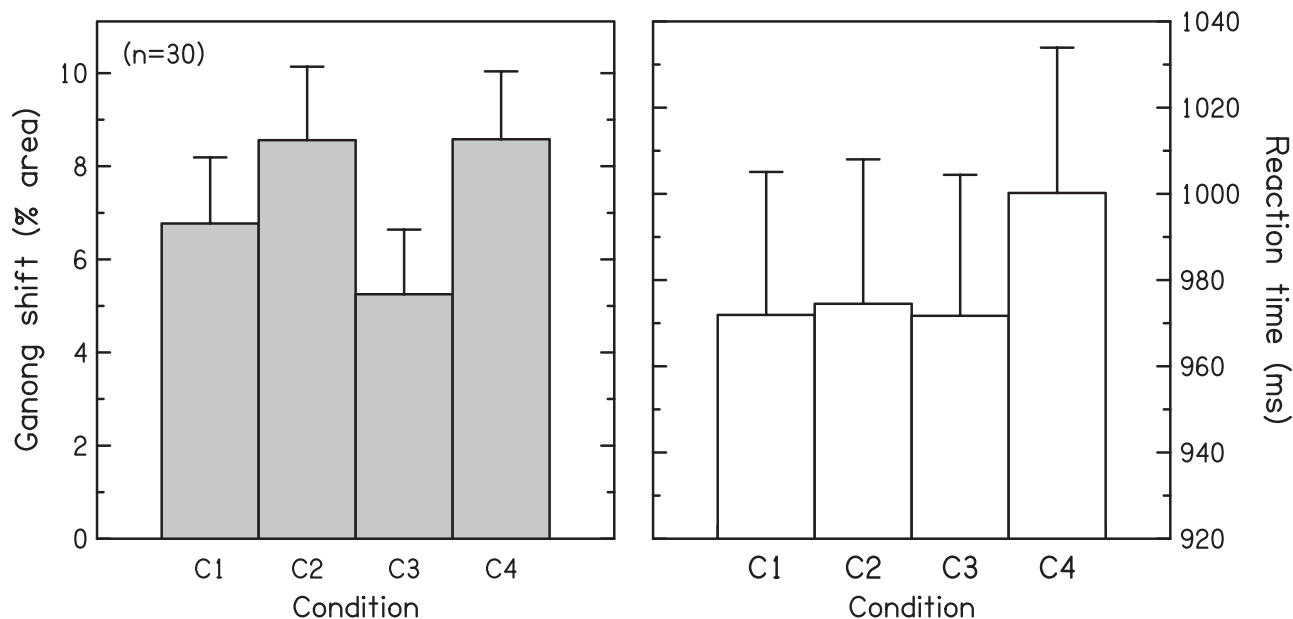


FIG. 7. Results for experiment 2—summary of Ganong shifts and overall reaction times for the four conditions (collapsed across spatial cue). Mean Ganong shifts and inter-subject standard errors ($n = 30$) are shown for each condition (bottom axis) in the left-hand panel (gray bars); the corresponding overall mean reaction times and inter-subject standard errors are shown in the right-hand panel (open bars).

B. Results

Figures 6 and 7 show the results for experiment 2 after averaging across fixed- and random-ear targets. This was done because both cases gave rise to very similar identification functions and reaction-time profiles (see the supplementary materials for versions of Fig. 6 showing the results separately for fixed- and random-ear targets¹), and there was no appreciable effect of this factor on the Ganong shifts or overall reaction times obtained. Figure 6 shows the identification functions and reaction-time profiles—including mean Ganong shifts and overall mean reaction times—in descending pairs of panels for each condition (C1–C4), and Fig. 7 summarizes the overall differences in Ganong shift and reaction time across conditions. Once again, the identification functions show that listeners produced clear and systematic patterns of judgments for both continua in all conditions, in most cases progressing monotonically from few to mainly /k/ responses with increasing position number. However, the pattern of contributions across continuum positions to the Ganong shift was not as expected. Typically, lexical bias is greatest for the most ambiguous tokens and decreases towards the endpoints. This standard pattern is interpreted as arising from a greater reliance on memory to disambiguate more perceptually ambiguous initial stops, whereas the endpoints require little or no disambiguation. In contrast, the contribution of positions 4 and 5—where the transition from mainly /g/ to mainly /k/ responses usually occurred—to the area difference between the two identification functions in C1 (no interferer) was smaller than for the corresponding MNB condition in experiment 1. This outcome was also manifest in the interference conditions (C2–C4). To our knowledge, this pattern has not

previously been reported for Ganong shifts and so its possible origin is considered in Sec. III C. Notwithstanding the cause of this anomaly, lexical bias was still evident at several continuum positions and so, in principle, there was scope for it to be increased as a consequence of any IM generated by the presence of an acoustic interferer.

Two-way ANOVA showed that there was neither a main effect of target-ear consistency (fixed vs random) on lexical bias [$F(1,29) = 0.004, p = 0.953, \eta_p^2 < 0.001$] nor an interaction of target-ear consistency with condition [$F(3,87) = 0.376, p = 0.770, \eta_p^2 = 0.013$]. The main effect of the condition on lexical bias was significant [$F(3,87) = 3.399, p = 0.021, \eta_p^2 = 0.105$], but the pattern obtained was not the one predicted. Rather than finding similar Ganong shifts for C1–C3 and a significantly greater shift for C4—the only case involving a frequency-varying interferer—C2 (constant-amplitude interferer) had a high mean almost identical to that for C4 and the effect of condition was driven mainly by C3 (amplitude-varying interferer). Specifically, there was less lexical bias for C3 than for C2 [difference = 3.31% pts, $t(29) = 2.459, p = 0.020$] or C4 [difference = 3.33% pts, $t(29) = 2.724, p = 0.011$]; none of the other pairwise comparisons were significant.

Further inspection of the data suggested that the size of the Ganong shifts declined over time in all conditions, but most markedly for C1 and C2, leading to a pattern towards the end of the listening session that was more consistent with that predicted. These observations were investigated by computing the Ganong shifts separately for the first and last five repetitions (irrespective of target-ear consistency). The mean values were as follows: first quarter: C1–C4 = 13.1%, 13.4%, 10.4%, and 11.3%, respectively; last quarter: C1–C4 = 2.8%, 1.8%, 4.7%, and 5.1%, respectively. These

changes correspond to declines in the mean Ganong shifts for the later trials, relative to the earlier ones, ranging from 55% to 79%; for comparison, relative declines from the first to the second halves of the trials in each block for experiment 1 were considerably smaller, ranging from 8% to 22%. It also merits note that the Ganong shift was numerically larger in C1 for the first-quarter data in experiment 2 than for the corresponding MNB condition in experiment 1 (6.7%, see Fig. 3). Two-way ANOVA showed that there was a highly significant main effect of time on lexical bias [$F(1,29) = 32.511$, $p < 0.001$, $\eta_p^2 = 0.529$]; there was no main effect of condition [$F(3,87) = 0.094$, $p = 0.963$, $\eta_p^2 = 0.003$], but there was a trend towards a significant interaction between condition and time [$F(3,87) = 2.377$, $p = 0.075$, $\eta_p^2 = 0.076$]. These aspects of the results are also discussed in Sec. III C (see the supplementary materials¹ for a figure showing the identification functions and Ganong shifts separately for the first and last quarter of trials in experiment 2; also included is a table showing mean Ganong shifts and inter-subject standard errors separately for earlier and later trials in experiment 1).

One might have expected slower reaction times when the target was subject to spatial uncertainty. Although the overall mean reaction time was nominally longer (by 13.5 ms) when there was spatial uncertainty, two-way ANOVA showed that there was neither a main effect of target-ear consistency (fixed vs random) [$F(1,29) = 0.274$, $p = 0.605$, $\eta_p^2 = 0.009$] nor an interaction of target-ear consistency with condition [$F(3,87) = 1.313$, $p = 0.275$, $\eta_p^2 = 0.043$]. However, the main effect of the condition on reaction time was highly significant [$F(3,87) = 9.373$, $p < 0.001$, $\eta_p^2 = 0.244$]. Moreover, the pattern of results obtained was as predicted. Specifically, reaction times were slower for C4 (frequency-varying interferer) than for any other condition [C4 vs C1, difference = 28.3 ms, $t(29) = 4.186$, $p < 0.001$; C4 vs C2, difference = 25.7 ms, $t(29) = 3.715$, $p < 0.001$; C4 vs C3, difference = 28.5 ms, $t(29) = 5.487$, $p < 0.001$]. Mean reaction times for C1–C3 were within 3 ms of one another and none of these differences were significant.

As in experiment 1, inspection of the reaction-time profiles across continuum positions showed a similar tendency towards longer reaction times for non-lexical than lexical responses. Once again, responses also tended to be slower away from the continuum endpoints, although this pattern was not as clearly defined as that seen in experiment 1, despite the broadly similar slopes of the identification functions. To explore whether these effects were significant, we again began by conducting a three-factor repeated-measures ANOVA—interference condition (I) \times continuum (C) \times position (P)—on the median reaction-time data; the statistical outcomes are presented in Table II. Except for the main effect of continuum ($p = 0.384$) and the three-way interaction ($p = 0.166$), all main effects and interactions were significant ($p < 0.001$), and so the effects of stimulus ambiguity and lexical vs non-lexical decisions were examined using the same approach as that taken in experiment 1. Reaction times for the more ambiguous inner positions (3–6) were significantly longer than for the outer positions (1–2 and 7–8) in all four conditions [C1: mean

TABLE II. Results for experiment 2—reaction times. Effects of interference condition, continuum, and position on the median reaction times of the responses to the initial consonant. Summary of the three-way repeated-measures ANOVA; all significant terms are shown in bold.

Factor	<i>df</i>	<i>F</i>	<i>p</i>	η_p^2
Interference condition (I)	(3, 87)	9.373	<0.001	0.244
Continuum (C)	(1, 29)	0.781	0.384	0.026
Position (P)	(7, 203)	9.100	<0.001	0.239
I \times C	(3, 87)	18.239	<0.001	0.386
I \times P	(21, 609)	3.517	<0.001	0.108
C \times P	(7, 203)	20.956	<0.001	0.419
I \times C \times P	(21, 609)	1.301	0.166	0.043

difference = 62.8 ms, $t(29) = 7.233$, $p < 0.001$; C2: mean difference = 53.7 ms, $t(29) = 6.043$, $p < 0.001$; C3: mean difference = 49.8 ms, $t(29) = 6.607$, $p < 0.001$; C4: mean difference = 14.9 ms, $t(29) = 2.205$, $p = 0.036$]. Finally, reaction times were significantly longer for non-word than word responses in all four conditions [C1: mean difference = 37.2 ms, $t(29) = 2.857$, $p = 0.008$; C2: mean difference = 43.6 ms, $t(29) = 3.529$, $p = 0.001$; C3: mean difference = 42.2 ms, $t(29) = 3.825$, $p < 0.001$; C4: mean difference = 53.5 ms, $t(29) = 3.589$, $p = 0.001$].

C. Discussion

Once again, the effect of condition on the size of the Ganong shifts showed a different pattern from its effect on the overall reaction times. The results for the latter measure were as predicted—accompanying the monaural target with a contralateral interferer without spectro-temporal change had no discernible effect (C2 and C3), but reaction times were slowed when the interferer involved frequency variation (C4). This outcome is in accord with the results of studies using three-formant synthetic analogues of target sentences or CV syllables, which have indicated that the impact on intelligibility of extraneous formants acting as informational maskers is influenced primarily by the extent and velocity of formant-frequency variation in the interferer (Roberts *et al.*, 2010, 2014, 2021; Roberts and Summers 2015, 2018, 2020; Summers *et al.*, 2012). Interpreting the overall results for the Ganong shifts is complicated by the absence of any significant pairwise differences involving the reference case (C1), but the most puzzling aspect is that the results for C2—which did not involve frequency variation—were like those for C4 rather than C3. One possible explanation concerns the more uniform distribution of energy across the duration of the interferers in C2, compared with the interferers used in C3 and C4, leading to relatively greater energy accompanying the terminal consonantal segment in C2 (see Fig. 5). Perhaps this led to increased IM of the terminal segment, impairing to some extent the distinction between /ft/ and /s/ on which the lexical bias being measured was based? If so, however, one might also have expected to observe a slower processing speed in C2, but this was not the case.

A distinctive difference between the identification functions obtained for the MNB syllables here, compared with experiment 1, is the tendency for the difference in area between the two functions to be small for the most ambiguous stimuli (positions 4 and 5). This outcome was robust, occurring whether or not an acoustic interferer was present. Probably related to this distinctive pattern is the observation that the categorization function steepens between positions 4 and 5 compared with the MNB condition in experiment 1. This indicates more categorical judgments for these positions, which fits with the notion that memory had less of an influence on responding. The differences between the two experiments are unlikely to be a consequence of sampling differences, given that a large number of listeners took part in each experiment ($n \geq 30$) and they were drawn from the same population. There are, however, three differences in the way the two experiments were run that potentially might account for this difference in the results and for the difference in the overall tendency for Ganong shifts to decline over time. First, the target syllables were presented diotically in experiment 1 but monaurally here. This factor is unlikely to be important, however, given that [Mattys *et al.* \(2009\)](#) directly compared the effects of diotic and monaural presentation on the reliance of speech perception on lexical knowledge and found no evidence of any differences between them. Second, the stimuli were blocked by stimulus type in experiment 1 but intermixed by interference condition in experiment 2; how trials are blocked has previously been shown to influence the Ganong shifts obtained (see, e.g., [Pitt and Samuel, 1993](#)). Third, there were eight times as many trials involving MNB syllables in experiment 2 (1280 trials) than in experiment 1 (160 trials), giving greater opportunities for perceptual learning and for any changes in listening strategy that might arise from intermixing the interference conditions. In this context, the temporal evolution of the responses over the course of experiment 2 was investigated by comparing the identification functions and Ganong shifts for the first quarter and last quarter of the trials.

The first observation arising from the comparison of the first and last quarters is the strong general tendency in all conditions for the Ganong shifts to fall over time, an effect much larger than that observed between earlier and later trials in experiment 1. This outcome indicates that there was a tendency over time to pay less attention to the lexical status of the stimuli and instead to focus increasingly on the initial consonantal segment, which is the only part of the target syllable that varied between trials. In turn, this suggests that there was a strategic rebalancing of attention away from information derived from imprecise lexical access towards information from the acoustic signal. In this context, the tendency for the difference in area between the two identification functions to be small for the most central positions (4 and 5) on the continua—a pattern that appears to be more prominent in the final quarter of the trials—suggests that the tendency to focus on acoustic detail is particularly marked for the most ambiguous consonants. The second observation arising from the comparison is that the fall in the Ganong

shifts over time was rather greater for C1 and C2 than for C3 and C4. This implies a greater strategic rebalancing towards acoustic detail when little or no IM was expected (C2 and C1, constant-amplitude interferers or none) than when appreciable IM was expected (C4, frequency-varying interferers)—presumably because of impaired signal encoding—albeit that a greater fall in the Ganong shift might also have been anticipated for C3 (amplitude-varying interferers without frequency change).

Notwithstanding this further exploration of the current results, the absence of a clear principled reason for prioritizing the data from the last quarter of trials and the wide range of factors potentially influencing listeners' responses (e.g., blocking vs intermixing, nature of the acoustic cues signaling the distinction being judged, perceptual learning, and listening strategy) indicates the need for further research before any firm conclusions can be drawn on the size and pattern of Ganong shifts obtained in the different interference conditions. For example, it would be informative to discover whether intermixing contralateral interferers with monaural targets causes a closing up of the two identification functions for the central positions in the REF and MSB conditions in a way similar to that observed here for the MNB stimuli. Such an outcome would suggest that the influence of memory on the perception of the initial stop, which is generally assumed to be relatively automatic, is preempted by IM.

It has recently been shown that contralateral interferers can affect consonant place judgments for CV syllables that were always presented in the same ear and therefore heard without spatial uncertainty ([Roberts *et al.*, 2021](#)). Nonetheless, the absence here of any appreciable effect of spatial uncertainty on either measure is perhaps surprising because, for such short materials, one might have expected presenting the target syllable in the same ear throughout a block of trials to help listeners focus their attention and ignore the interferer.

IV. SUMMARY AND CONCLUDING DISCUSSION

The results for the two experiments reported here showed different patterns across conditions for our measures of lexical bias (Ganong shifts) and overall processing speed (response latency from the start of the stimulus). The findings for overall reaction times were as expected—vocoding the target syllables using either noise or sine carriers slowed processing (experiment 1) and accompanying noise-vocoded syllables with similarly vocoded interferers in the contralateral ear slowed processing only if the interferers involved frequency variation ([Roberts and Summers, 2015, 2018, 2020](#); [Summers *et al.*, 2012](#); [Roberts *et al.*, 2021](#)). The underlying reaction-time profiles were also consistent with previous research in showing evidence of longer response times for judgments of more ambiguous tokens (e.g., [Fox, 1984](#)) and for non-lexical than for lexical decisions (e.g., [Goh *et al.*, 2009](#)).

The Ganong shift findings have been interpreted in the context of the framework for listening to speech in adverse

conditions proposed by [Mattys et al. \(2009\)](#), in which the EM and IM components of such conditions can be identified. The former and the latter favor an increased reliance in speech perception on acoustic detail and lexical knowledge, respectively. Unlike overall reaction times, the size of the Ganong shifts observed in experiment 1 was heavily influenced by the spectral sparsity of the stimulus tokens, being more than twice as large for the MSB than for the MNB stimuli. Presumably, this outcome was a consequence of the relationship between the carrier used and the naturalness of the vocoded stimuli; MSB stimuli sound far less natural than their MNB counterparts, acting like a cognitive load on the perceiver and thereby increasing reliance on lexical knowledge. The contrasting absence of effects of stimulus naturalness reported by [Burton and Blumstein \(1995\)](#) may be attributable to the subtlety of the cue manipulations that they used to influence naturalness, compared with the greater salience of differences in naturalness achieved using vocoding with different carriers. The reason for the significantly *smaller* Ganong shifts found for the MNB stimuli relative to the REF stimuli (not vocoded) is not clear, but this result implies less lexical bias not more in judgments of the former. This outcome suggests that the effect of the modest signal degradation caused by vocoding with 16 channels outweighed the effect of the cognitive load associated with using a noise-only carrier, which in turn suggests that this load was small.

In experiment 2, the identification functions for the MNB syllables were consistent with previous research in that they were characterized by a broadly monotonic increase in the proportion of /k/ responses across the continua in all conditions, and were associated with evidence of lexical bias in the judgments made. However, the tendency for the most central continuum positions (4 and 5) to contribute less towards the overall Ganong shifts and the associated tendency towards more categorical judgments for those positions compared with the MNB syllables in experiment 1 suggests that contextual differences between the experiments were important, most particularly the intermixing of trials from the four conditions (no interferer and three types of interferer) within the same block. Furthermore, unlike overall reaction times, the pattern of Ganong shifts observed in the presence of contralateral acoustic interferers acting as informational maskers did not conform with that expected based on the established relationship between the extent and velocity of formant-frequency variation in a speech-like interferer and its impact on target intelligibility ([Roberts and Summers, 2015, 2018, 2020](#); [Summers et al., 2012](#); [Roberts et al., 2021](#)). Further exploration of these data indicated a strong tendency for the Ganong shifts to diminish over time and some evidence to suggest the emergence towards the end of the listening session of a pattern more similar to that predicted—namely, that frequency-varying interferers caused more IM. To draw any firmer conclusions, further research is needed to explore systematically the effects of a range of factors likely to influence lexical bias in the presence of acoustic interferers acting as informational maskers. These factors include the intermixing of stimulus conditions

and the roles of perceptual learning and listening strategy in shaping the temporal evolution of listeners' judgments.

Predicting the relative extent in speech perception of reliance on acoustic detail and lexical knowledge can be particularly challenging in the context of perceptual loads. For example, manipulations involving signal degradation are associated not only with an EM component, but also an IM component corresponding to the cognitive load which often arises as a secondary consequence of signal degradation ([Mattys et al., 2009](#); [Mattys et al., 2012](#)). Presumably, for example, the contribution of the latter accounts for the *increased* lexical bias (Ganong shifts) observed by [Gianakas and Winn \(2019\)](#) for place-of-articulation judgments made by cochlear implant listeners and by normal-hearing listeners responding to eight-channel MNB syllables. Note, however, that increased lexical bias was not apparent in some of the other contrasts they tested—e.g., there was no evidence of any lexical bias for the fricative contrast and little for the vowel contrast.

One factor that makes the balance of these effects hard to predict is that much of what is known about this balance is limited to the specific testing context used in the experiments in which the framework of listening in adverse conditions was first proposed ([Mattys et al., 2009](#)). In that framework, a segmentation task was used in which listeners rated their confidence that the stimulus phrase contained a particular target word. The stimuli were derived from natural productions of two-word phrases ranging from, e.g., “mild option” (lexically acceptable) to “mile doption” (lexically inconsistent). In this arrangement, the primary cues for the contrast were relatively long in duration—allophonic and prosodic differences around the word juncture—which are very different from the relatively brief consonantal cues that support the voicing contrast used in the experiments reported here. These differences may well influence both perceptual learning and listening strategy, such as the rate at which attention can be focused on the critical acoustic detail, and hence the size of the Ganong shifts obtained. Therefore, further research in this area might also consider measuring the extent of acoustic or lexical bias using a wider range of contrast cues.

There is an established body of evidence that changes in the Ganong shift can reliably indicate the extent to which dependence on lexical knowledge changes when the stimulus or its sensori-neural representation is degraded (e.g., [Gianakas and Winn, 2019](#)) or when listeners are under a cognitive load (e.g., [Mattys and Wiget, 2011](#)). The experiments reported here were intended to elucidate how Ganong shifts, and also reaction times, are influenced by stimulus naturalness and by acoustical IM. The results provide some evidence to suggest that Ganong shifts can be increased by reducing stimulus naturalness and by accompanying the target syllable with a frequency-varying interferer in the contralateral ear. Clearly, however, the notion that changes in the size of the Ganong shift might provide a straightforward measure of the non-specific cognitive load associated with an acoustic informational masker—akin to that associated

with engaging in a concurrent task (Mattys and Wiget, 2011; Mattys and Palmer, 2015) or being distracted by induced acute anxiety (Mattys *et al.*, 2013)—requires caution until further research clarifies in this context the effects of stimulus intermixing, perceptual learning, and listening strategy. Furthermore, any comparisons made for this purpose should be restricted to those between conditions using target stimuli with similar acoustic source properties and therefore stimulus naturalness (e.g., the same carrier type used for vocoding), as well as similar intelligibility when heard unaccompanied. For now, changes in response latency would appear to offer a more straightforward measure of changes in the extent of the disrupting effects of IM arising from the presence of interfering sounds, but reaction times do not directly indicate the relative extent of reliance on acoustic detail and lexical knowledge.

ACKNOWLEDGMENTS

This research was supported by Research Grant ES/N014383/1 from the Economic and Social Research Council (UK), awarded to Brian Roberts. Our thanks go to Paul Iverson for his suggestions regarding the likely impacts of sine and noise carriers on how listeners attended the vocoded syllables, and to Mark Pitt for his comments on earlier drafts of this paper. Poster presentations on the first experiment were given at the 175th Meeting of the Acoustical Society of America (Minneapolis, MN, May 2018), and at the Basic Auditory Science Meeting of the British Society of Audiology (Newcastle-upon-Tyne, UK, September 2018).

¹See supplementary material at <https://www.scitation.org/doi/suppl/10.1121/10.0011395> for all stimuli and parameters, and also for additional analysis of the results.

²See <https://doi.org/10.17036/researchdata.aston.ac.uk.00000539> (Last viewed May 15, 2022).

- Boersma, P., and Weenink, D. (2016). "PRAAT, a system for doing phonetics by computer (version 6.0.20) [software package]," <http://www.praat.org/> (Last viewed May 15, 2022).
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
- 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.
- Burton, M. W., and Blumstein, S. E. (1995). "Lexical effects on phonetic categorization: The role of stimulus naturalness and stimulus quality," *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 1230–1235.
- Connine, C. M., and Clifton, C., Jr. (1987). "Interactive use of lexical information in speech perception," *J. Exp. Psychol. Hum. Percept. Perform.* **13**, 291–299.
- Darwin, C. J. (2008). "Listening to speech in the presence of other sounds," *Philos. Trans. R Soc. B* **363**, 1011–1021.
- Dorman, M. F., Loizou, P. C., and Rainey, D. (1997). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *J. Acoust. Soc. Am.* **102**, 2403–2411.
- 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.
- Ellermeier, W., and Zimmer, K. (2014). "The psychoacoustics of the irrelevant sound effect," *Acoust. Sci. Technol.* **35**, 10–16.
- Fox, R. A. (1984). "Effect of lexical status on phonetic categorization," *J. Exp. Psychol. Hum. Percept. Perform.* **10**, 526–540.
- 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.
- Ganong, W. F. (1980). "Phonetic categorization in auditory word perception," *J. Exp. Psychol. Hum. Percept. Perform.* **6**, 110–125.
- Gianakas, S. P., and Winn, M. B. (2019). "Lexical bias in word recognition by cochlear implant listeners," *J. Acoust. Soc. Am.* **146**, 3373–3383.
- Goh, W. D., Suárez, L., Yap, M. J., and Tan, S. H. (2009). "Distributional analyses in auditory lexical decision: Neighborhood density and word-frequency effects," *Psychon. Bull. Rev.* **16**, 882–887.
- Hill, F. J., McRae, L. P., and McClellan, R. P. (1968). "Speech recognition as a function of channel capacity in a discrete set of channels," *J. Acoust. Soc. Am.* **44**, 13–18.
- 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.
- Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.* **27**, 187–207.
- Keppel, G., and Wickens, T. D. (2004). *Design and Analysis: A Researcher's Handbook*, 4th ed. (Pearson Prentice Hall, Englewood Cliffs, NJ).
- Kidd, G., Jr., Mason, C. R., Richards, V. M., Gallun, F. J., and Durlach, N. I. (2008). "Informational masking," in *Auditory Perception of Sound Sources*, *Springer Handbook of Auditory Research*, Vol. 29, edited by W. A. Yost and R. R. Fay (Springer, Boston, MA), pp. 143–189.
- Lawrence, M. A. (2016). "ez: Easy analysis and visualization of factorial experiments (R package version 4.4-0) [software]," <https://cran.r-project.org/package=ez> (Last viewed July 30, 2018).
- Lisker, L., and Abramson, A. S. (1964). "A cross-language study of voicing in stops: Acoustical measurements," *Word* **20**, 384–422.
- Mattys, S. L., Barden, K., and Samuel, A. G. (2014). "Extrinsic cognitive load impairs low-level speech perception," *Psychon. Bull. Rev.* **21**, 748–754.
- Mattys, S. L., Brooks, J., and Cooke, M. (2009). "Recognizing speech under a processing load: Dissociating energetic from informational factors," *Cogn. Psychol.* **59**, 203–243.
- Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). "Speech recognition in adverse conditions: A review," *Lang. Cogn. Process.* **27**, 953–978.
- Mattys, S. L., and Palmer, S. D. (2015). "Divided attention disrupts perceptual encoding during speech recognition," *J. Acoust. Soc. Am.* **137**, 1464–1472.
- Mattys, S. L., Seymour, F., Attwood, A. S., and Munafò, M. R. (2013). "Effects of acute anxiety induction on speech perception: Are anxious listeners distracted listeners?," *Psychol. Sci.* **24**, 1606–1608.
- Mattys, S. L., White, L., and Melhorn, J. F. (2005). "Integration of multiple speech segmentation cues: A hierarchical framework," *J. Exp. Psychol. Gen.* **134**, 477–500.
- Mattys, S. L., and Wiget, L. (2011). "Effects of cognitive load on speech recognition," *J. Mem. Lang.* **65**, 145–160.
- McClelland, J. L., Mirman, D., and Holt, L. L. (2006). "Are there interactive processes in speech perception?," *Trends Cogn. Sci.* **10**, 363–369.
- McQueen, J. M. (1991). "The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity," *J. Exp. Psychol. Hum. Percept. Perform.* **17**, 433–443.
- Miller, J. L., and Dexter, E. R. (1988). "Effects of speaking rate and lexical status on phonetic perception," *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 369–378.
- Pitt, M. A., and Samuel, A. G. (1993). "An empirical and meta-analytic evaluation of the phoneme identification task," *J. Exp. Psychol. Hum. Percept. Perform.* **19**, 699–725.
- Pitt, M. A., and Samuel, A. G. (2006). "Word length and lexical activation: Longer is better," *J. Exp. Psychol. Hum. Percept. Perform.* **32**, 1120–1135.
- Porter, R. J., Jr., and Whittaker, R. G. (1980). "Dichotic and monotic masking of CV's by CV second formants with different transition starting values," *J. Acoust. Soc. Am.* **67**, 1772–1780.
- R Core Team (2020). "R: A language and environment for statistical computing [software package]," <http://www.R-project.org/> (Last viewed July 31, 2020).

- 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. (2020). "Informational masking of speech depends on masker spectro-temporal variation but not on its coherence," *J. Acoust. Soc. Am.* **148**, 2416–2428.
- Roberts, B., Summers, R. J., and Bailey, P. J. (2010). "The perceptual organization of sine-wave speech under competitive conditions," *J. Acoust. Soc. Am.* **128**, 804–817.
- Roberts, B., Summers, R. J., and Bailey, P. J. (2011). "The intelligibility of noise-vocoded speech: Spectral information available from across-channel comparison of amplitude envelopes," *Proc. R Soc. B* **278**, 1595–1600.
- Roberts, B., Summers, R. J., and Bailey, P. J. (2014). "Formant-frequency variation and informational masking of speech by extraneous formants: Evidence against dynamic and speech-specific acoustical constraints," *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1507–1525.
- Roberts, B., Summers, R. J., and Bailey, P. J. (2021). "Mandatory dichotic integration of second-formant information: Contralateral sine bleats have predictable effects on consonant place judgments," *J. Acoust. Soc. Am.* **150**, 3693–3710.
- Rosen, S., Zhang, Y., and Speers, K. (2015). "Spectral density affects the intelligibility of tone-vocoded speech: Implications for cochlear implant simulations," *J. Acoust. Soc. Am.* **138**, EL318–EL323.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Shinn-Cunningham, B. G. (2008). "Object-based auditory and visual attention," *Trends Cogn. Sci.* **12**, 182–186.
- Simon, J. R., and Rudell, A. P. (1967). "Auditory S-R compatibility: The effect of an irrelevant cue on information processing," *J. Appl. Psychol.* **51**, 300–304.
- Snedecor, G. W., and Cochran, W. G. (1967). *Statistical Methods*, 6th ed. (Iowa University Press, Ames, IA).
- Souza, P., and Rosen, S. (2009). "Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech," *J. Acoust. Soc. Am.* **126**, 792–805.
- Summers, R. J., Bailey, P. J., and Roberts, B. (2010). "Effects of differences in fundamental frequency on across-formant grouping in speech perception," *J. Acoust. Soc. Am.* **128**, 3667–3677.
- Summers, R. J., Bailey, P. J., and Roberts, B. (2012). "Effects of the rate of formant-frequency variation on the grouping of formants in speech perception," *J. Assoc. Res. Otolaryngol.* **13**, 269–280.
- Summers, R. J., Bailey, P. J., and Roberts, B. (2017). "Informational masking and the effects of differences in fundamental frequency and fundamental-frequency contour on phonetic integration in a formant ensemble," *Hear. Res.* **344**, 295–303.
- Summers, R. J., and Roberts, B. (2020). "Informational masking of speech by acoustically similar intelligible and unintelligible interferers," *J. Acoust. Soc. Am.* **147**, 1113–1125.
- Tremblay, S., and Jones, D. M. (1999). "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound," *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 1005–1015.
- Villard, S., and Kidd, G., Jr. (2021). "Speech intelligibility and talker gender classification with noise-vocoded and tone-vocoded speech," *JASA Express Lett.* **1**(9), 094401.
- Viswanathan, N., Dorsi, J., and George, S. (2014). "The role of speech-specific properties of the background in the irrelevant sound effect," *Q. J. Exp. Psychol.* **67**, 581–589.