Atten Percept Psychophys (2015) 77:1342–1357
DOI 10.3758/s13414-015-0855-z

Interdependent processing and encoding of speech and concurrent background noise

Angela Cooper · Susanne Brouwer · Ann R. Bradlow

Published online: 14 March 2015

Abstract Speech processing can often take place in adverse listening conditions that involve the mixing of speech and background noise. In this study, we investigated processing dependencies between background noise and indexical speech features, using a speeded classification paradigm (Garner, 1974; Exp. 1), and whether background noise is encoded and represented in memory for spoken words in a continuous recognition memory paradigm (Exp. 2). Whether or not the noise spectrally overlapped with the speech signal was also manipulated. The results of Experiment 1 indicated that background noise and indexical features of speech (gender, talker identity) cannot be completely segregated during processing, even when the two auditory streams are spectrally nonoverlapping. Perceptual interference was asymmetric, whereby irrelevant indexical feature variation in the speech signal slowed noise classification to a greater extent than irrelevant noise variation slowed speech classification. This asymmetry may stem from the fact that speech features have greater functional relevance to listeners, and are thus more difficult to selectively ignore than background noise. Experiment 2 revealed that a recognition cost for words embedded in different types of background noise on the first and second occurrences only emerged when the noise and the speech signal were spectrally overlapping. Together, these data suggest integral processing of speech and background noise, modulated by the level of processing and the spectral separation of the speech and noise.

A. Cooper (✉) · A. R. Bradlow
Department of Linguistics, Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA
e-mail: akcooper@u.northwestern.edu

S. Brouwer
Department of Special Education, Utrecht University, Utrecht, The Netherlands

Keywords Selective attention · Speech perception · Implicit/explicit memory

In everyday conversations, listeners must sift through multiple dimensions of the incoming auditory input in order to extract the relevant linguistic content. The speech signal contains not only linguistic material but also indexical information, which includes the particular voice and articulatory characteristics of the speaker that would enable a listener to identify the speaker's gender or individual identity. Moreover, listeners must also contend with the fact that, in many situations, environmental noise will co-occur with the speech signal. Although robust evidence suggests that the linguistic and indexical dimensions of speech are integrally processed during speech perception (e.g., Mullennix & Pisoni, 1990), relatively little research has been conducted on whether or not linguistically irrelevant environmental noise[1] is also processed integrally and/or encoded in memory with the linguistic and indexical attributes of a speech event during speech processing. Thus, in the present study we investigated (a) the extent to which indexical speech features and background noise are processed interdependently at a relatively early stage of processing, using the Garner speeded classification paradigm (following Garner, 1974; Exp. 1), and (b) whether the consistency of concurrently presented background noise from a first to a second occurrence can serve as a facilitatory cue for recognition of the word as having occurred earlier in a list of spoken words (following Palmeri, Goldinger, & Pisoni, 1993, and Bradlow, Nygaard, & Pisoni, 1999; Exp. 2).

[1] The term "noise" can be used to refer to any extraneous background sound (e.g., dogs barking, even other speech streams). In this article, "noise" and "background noise" will refer specifically to the filtered white-noise and pure-tone samples used in the stimuli. We will use the term "environmental noise" to refer to extraneous background sounds more generally.

Integration of indexical and linguistic information

Traditional models of spoken word recognition have assumed that linguistic processing operates over abstract symbolic representations, and that nonlinguistic features of the speech signal, such as indexical information, are stripped away from the linguistic content during speech processing and encoding (see Pisoni, 1997, for a review). However, a growing body of literature has demonstrated that linguistic and indexical information are perceptually integrated and encoded during speech processing (e.g., Bradlow et al., 1999; Church & Schacter, 1994; Cutler, Andics, & Fang, 2011; Goldinger, 1996; Kaganovich, Francis, & Melara, 2006; Mullennix & Pisoni, 1990; Nygaard, Sommers, & Pisoni, 1994; Palmeri et al., 1993; Schacter & Church, 1992). For example, several studies have investigated this issue by using the Garner speeded classification paradigm (Garner, 1974) to determine how interdependent the processing of linguistic and indexical information are with one another (e.g., Cutler et al., 2011; Green, Tomiak, & Kuhl, 1997; Kaganovich et al., 2006; Mullennix & Pisoni, 1990). In the Garner task, listeners are asked to attend to one dimension and ignore the other dimension, which could be held constant (control), covary (correlated), or vary randomly (orthogonal). If these dimensions are processed independently of one another, then irrelevant variation in the unattended dimension should not have a substantive effect on response latencies for classifying the stimuli along the attended dimension relative to the control condition. However, integral processing of these dimensions would manifest as slower response latencies as a result of random variation in the unattended dimension (referred to as orthogonal interference) or faster classifications from the covariation of the stimulus dimensions (referred to as redundancy gain). Mullennix and Pisoni (1990) found asymmetrical orthogonal interference between phonetic and indexical dimensions of the speech signal. Listeners were slower at classifying initial consonants (either /b/ or /p/) when the talker gender varied randomly, as compared to the baseline control, in which gender was held constant. When identifying talker gender, listeners were similarly slowed by irrelevant phonetic variation, though to a lesser degree. Similar findings have been reported for the integral processing of vowel and talker identity (Cutler et al., 2011; Kaganovich et al., 2006), as well as final alveolar consonants (/s/, /t/) and talker identity (Cutler et al., 2011), where talker identity refers to two voices of the same gender associated with arbitrarily assigned names (e.g., Peter, Thomas).

Furthermore, a same-voice advantage has been found for the recognition of spoken words, whereby listeners were faster and more accurate at indicating whether or not a word had previously been presented when the item was produced by the same talker versus a different talker (e.g., Bradlow et al., 1999; Palmeri et al., 1993). Similarly, trial-to-trial talker changes have yielded performance decrements in word recognition and naming (Mullennix, Pisoni, & Martin, 1989), providing further evidence of an integral relationship between the processing of information about talker identity with the processing of linguistic information. Accruing perceptual experience with the voice characteristics of a talker has also been shown to enhance listeners' ability to extract linguistic content from speech produced by the familiar talker in adverse listening conditions (e.g., Johnsrude et al., 2013; Newman & Evers, 2007; Nygaard et al., 1994). These findings provide robust support for the notion that linguistic and indexical classifications are, in part, dependent on one another. This evidence has been used, in conjunction with the Garner task results discussed above, in support of the view that instead of retaining solely abstract linguistic representations, listeners also encode episodic details of the perceptual context in which speech occurred into (or along with) the representation of lexical items in their mental lexicons (Goldinger, 1996, 1998; Johnson, 2006; Pierrehumbert, 2001).

Noise in speech processing

This perceptual integration and retention in memory of linguistic and indexical information may not be surprising if one considers that talker and linguistic information coexist in a single speech stream and necessarily stem from the same sound source. However, our conversational interactions can take place in a variety of adverse listening conditions that may involve the mixing of the speech signal with environmental noise (see Mattys, Davis, Bradlow, & Scott, 2012, for a review). This raises the question of whether or not environmental sounds and speech are also perceptually integrated and retained in memory. Speech and environmental noise are typically produced from two distinct sound sources and thus have dissimilar acoustic signatures; therefore, it is conceivable that environmental noise could be relatively easily and effectively filtered out, or at least perceptually segregated from the speech signal, at an early stage of speech processing. Tomiak, Mullennix, and Sawusch (1987) reported an asymmetry between how listeners process what they believe to be speech and what they believe to be noise. Noise–tone analogues of fricative–vowel syllables were presented to listeners in a Garner paradigm. One group was informed that the stimuli were computer-generated noise–tone sequences and had to classify them along the noise and tone dimensions, whereas the other group was informed that the stimuli were speech and classified them along the fricative and vowel dimensions. Integral processing of the noise fricative and tone vowel dimensions was only found for the group who believed the stimuli were speech. This may suggest that noise (or what is believed to be noise) is processed in a fundamentally different manner than speech, engaging a different mode of processing and utilizing different processes.

Alternatively, in the context of the episodic models of speech perception (e.g., Goldinger, 1996), if listeners do encode all perceptual details of a given speech event in an integrated cognitive representation, then linguistically irrelevant background sounds that are extrinsic to the speech signal (i.e., from a difference source) would be expected to influence speech processing in a similar way to the influence of intrinsic, indexical information in the speech signal. Indeed, recent research has suggested that listeners construct context-specific, integrated representations during novel word learning. Creel, Aslin, and Tanenhaus (2012) trained English listeners on nonsense word-meaning associations, manipulating whether they were initially exposed to the items in white noise or in the clear and whether they were tested in white noise or in the clear. Learners were found to be faster and more accurate when the listening conditions matched between the training and test phases, suggesting that their newly forming lexical representations included details related to the extraneous nonspeech context of the initial exposure. Similarly, Pufahl and Samuel (2014) found that listeners who were exposed to words paired with background sounds (e.g., a dog bark) were less accurate at later transcribing heavily filtered words in a word recognition task when the background sound mismatched between exposure and test than when it matched, providing support for the notion that listeners retain the perceptual details of a speech event including details that are related to the presence of a background sound that is extrinsic to the speech signal.

These studies using an exposure-test experimental paradigm have indicated that, after being exposed to spoken words embedded in background noise in the exposure phase, a change in the concurrent noise of a given item will impact how well listeners recognize that spoken word at test. Creel et al. (2012) took their findings as evidence for "highly-specific, integral representations" (p. 1033). However, the results in Creel et al. (2012) may have stemmed from the fact that the words heard with environmental noise were indeed acoustically and perceptually distinct from the versions heard without noise. Although it could be the case that listeners store integrated representations of the environmental noise itself along with the word exemplar, it is also plausible that listeners store segregated representations of speech and noise, whereby encoded word exemplars would have spectro-temporal gaps due to masking from the noise. According to this view, if the word "ball," for instance, were presented concurrently with white noise at 3–4 kHz and a 0-dB signal-to-noise ratio, then a listener would segregate this noise from the speech signal and store the word with a spectral gap at 3–4 kHz. For example, in Creel et al. (2012), the superior performance of the participants whose listening conditions matched at both exposure and test could be explained by the fact that hearing a word at test that perfectly matched its presentation at exposure either activated the integral speech + noise representation formed during exposure or activated the speech + spectral-gap representation formed during exposure. Pufahl and Samuel (2014) posited a similar idea to explain their findings, when they noted that pairing a word with noise resulted in a degraded speech input and that later word recognition with the same word–noise pairing might have been improved by the fact that listeners had had previous exposure to that "pattern of residual information left in the word" (p. 26).

In order to gain a more detailed view of the processing and storage of spoken words with concurrent background noise, the present work contrasted two tasks and two stimulus conditions. In Experiment 1, we investigated processing dependencies between indexical speech features and background noise at an early stage of processing with the Garner speeded classification paradigm. Through this experiment, we examined the extent to which two different types of nonlinguistic information—namely indexical (talker identity and gender) information and noise—are integrally processed. In Experiment 2, we then probed the joint encoding of spoken words and background noise with a continuous recognition memory paradigm. Since prior work had only utilized stimuli in which the dimensions to be encoded were either inherent to the same signal (e.g., linguistic and indexical speech features) or spectrally overlapping (e.g., background sounds and speech that both covered the full spectral range presented to participants), we compared listener responses under conditions of either spectral overlap or spectral segregation between the speech and noise signals, to examine how energetic masking mediates the specificity of encoding speech signals in memory.

Experiment 1: Speeded classification

In Experiment 1, we examined the processing of speech and noise information using a speeded classification paradigm (Garner, 1974). The dimensions of the speech signal were divided into indexical features that are intrinsic to the speech input—namely gender (Exp. 1A) and within-gender talker identity (Exp. 1B). Examining listeners' classification of these different dimensions of speech in the presence of random noise variations, as well as classifying noise in the presence of random speech variations along these dimensions (gender or talker), would provide insight into which, if any, dimensions of the speech signal are processed independently of extraneous background noise in the context of a task that involves minimal access to the long-term mental lexicon. As such, this task can inform our understanding of speech-in-noise perception at a relatively early stage of processing. Furthermore, in order to investigate whether the extent of acoustic overlap in the spectral domain between two auditory signals has an impact on whether or not the speech and noise dimensions are processed independently, we also included conditions manipulating the spectral overlap of these dimensions

in each experiment. That is, we manipulated whether energetic masking of the speech by the noise was present or absent.

If all concurrently presented perceptual details of a speech event, including those that are intrinsic to the speech signal (gender and talker) and those that are extrinsic to the speech signal (environmental noise), are perceptually integrated, we should observe slower reaction times in the orthogonal condition (which incorporated variation along both dimensions) than in the control condition (in which only one dimension varied). Furthermore, we may hypothesize that the integrality of the dimensions can vary as a function of the ease with which listeners can strip away elements of the auditory context that are extraneous to the speech signal. Under this view, perceptual separation of the speech and noise should be particularly facilitated when the noise and speech are highly acoustically distinct. In the present study, this would predict an asymmetry in the magnitudes of the difference in reaction times between orthogonal and control conditions between the spectrally separated condition and the spectrally overlapped condition. Since it might be more difficult to perceptually segregate the speech and noise dimensions in the spectrally overlapped condition, we might predict greater orthogonal interference in this condition than in the spectrally separated condition.

Finally, previous research has suggested that asymmetric interference effects arise from the relative discriminability of the dimensions, such that more-discriminable dimensions should be more difficult to ignore (Cutler et al., 2011; Garner, 1974). Discriminability is typically indexed by comparing reaction times in the control conditions of the two dimensions, with shorter latencies indicating easier discriminability. However, the prior work on the integration of speech and indexical information discussed above had not made this discriminability comparison within the same study, using the same materials and experimental setup. The present work included two indexical speech features, gender and within-gender talker identity, to further investigate the impact of discriminability on the integration of speech and noise. These indexical dimensions vary in their relative eases of discriminability, with gender (Exp. 1A) being easier to classify than within-gender talker identity (Exp. 1B).

Experiment 1A: Perceptual integration of talker gender and background noise

Method

Participants Eighty-three American English listeners, who reported having no speech or hearing deficits at the time of testing, participated in this experiment. Participants who had experience with more than one language before the age of 11 were required to have learned English first and not to have been exposed to the other language for more than 5 h per week. Listeners were randomly assigned to either the nonenergetic-masking (NEM) or the energetic-masking (EM) condition. In order to be included in the analyses, participants were required to attain at least 90 % classification accuracy for both dimensions of the Garner (1974) task, resulting in the exclusion of 11 participants (eight from the NEM and three from the EM group). This yielded 36 participants in the NEM (22 female, 14 male; mean age = 20 years) and 36 participants in the EM (22 female, 14 male; mean age = 20 years) condition.

Stimuli The stimulus materials included 96 English disyllabic, initial-stress words produced by one male and one female American English talker. The words were produced in citation form and recorded at a 22,050-Hz sampling rate. Acoustic analyses performed on the 96 stimuli items produced by the two talkers revealed that the mean difference in fundamental frequencies between the male and female talkers was 86 Hz (M_{male} = 144 Hz, M_{female} = 230 Hz). The average F0 ranges (calculated as the difference between the mean F0 minimum and F0 maximum across words) were 165 Hz for the male talker and 146 Hz for the female talker.

The set of materials was digitally processed in Praat (Boersma & Weenink, 2013) to yield the four different noise and masking stimulus sets (Fig. 1). The stimuli were first normalized for duration, low-pass filtered at 5 kHz, and normalized for root-mean-squared (RMS) amplitude to 65 dB. For the nonenergetic masking (NEM) condition, two sets of stimuli were constructed, with each set including all 96 recorded words: For one set, the speech files were combined with narrow band-pass-filtered white noise from 7 to 10 kHz, and for the other set, they were combined with a 6-kHz pure tone (Fig. 1, right column). Similarly, two sets of stimuli were constructed to produce items for the energetic masking (EM) condition: The low-pass-filtered speech files were combined with either narrow band-pass-filtered white noise from 3 to 4 kHz (Set 1) or a 3-kHz pure tone (Set 2; Fig. 1, left column). In total, there were four sets of stimuli: NEM-noise, NEM-tone, EM-noise, and EM-tone.

Procedure Whereas masking condition (NEM or EM) was a between-subjects manipulation, stimulus dimension (gender, noise) and stimulus set condition (control, correlated, orthogonal) were within-subjects manipulations. When making classification judgments, all participants were required to attend to either the gender (male vs. female) or the noise (pure tone vs. white noise) dimension. They completed both of these judgments in each of the three stimulus set conditions, blocked by stimulus dimension. This resulted in a total of six sets of trials, and all 96 words were presented in every set with no repetitions of items (i.e., a total of 96 trials per stimulus set condition). The orders of stimulus dimension and stimulus set were counterbalanced across participants, as were response button order and which words were presented with which particular

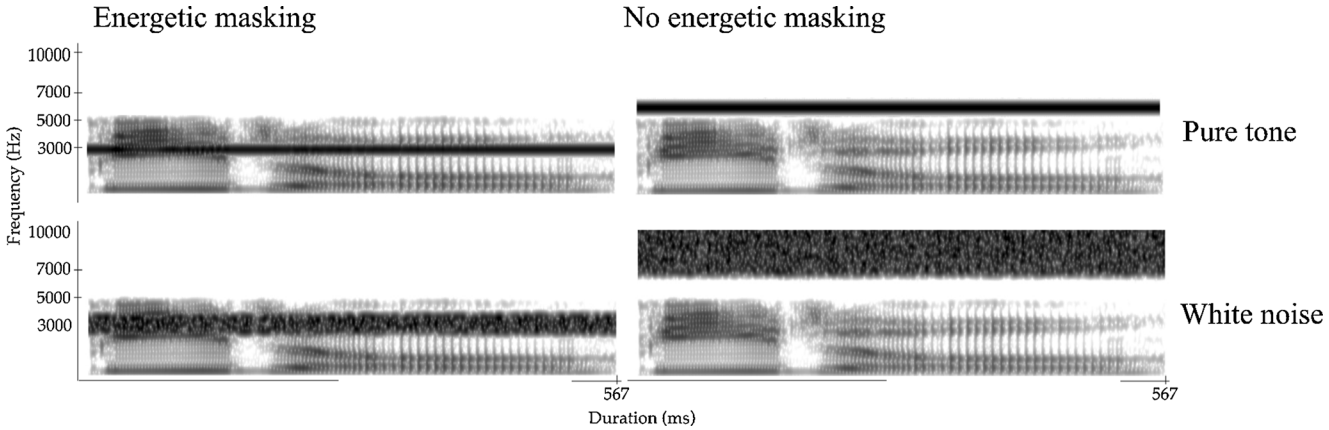

Fig. 1 Sample spectrograms of the four types of stimuli for the same word. The left column depicts energetic-masked items (3-kHz pure tone, top; and 3- to 4-kHz white noise, bottom). The right column depicts non-energetic-masked items (6-kHz pure tone, top; and 7- to 10-kHz white noise, bottom). All stimuli were normalized to 567 ms

gender–noise combinations (e.g., male talker with pure tone, female talker with white noise).

In the control conditions, the attended dimension varied randomly and the unattended dimension was held constant. One gender control set, for example, included words spoken by both the male and female talkers embedded only in white noise. The control set for the noise condition, on the other hand, presented words with both white-noise and pure-tone backgrounds spoken by a single talker. Each participant completed one gender and one noise control set, which were counterbalanced across participants. In the correlated condition, one value of the gender dimension was consistently paired with one value of the noise dimension. For instance, one set included words with a pure-tone background produced by the female talker and words with a white-noise background produced by the male talker, whereas the other set consisted of words in white noise produced by the male talker and words with a pure tone produced by the female talker. Which correlated set a participant received was also counterbalanced across participants. In the orthogonal condition, the attended and unattended dimensions varied randomly, whereby all items produced by both male and female talkers in both white noise and a pure tone were presented to the participant for classification along the attended dimension. Before each stimulus dimension condition, a brief familiarization phase was presented in order to orient listeners to the task procedures for that particular condition. Each familiarization phase consisted of ten trials (five items for each response option) using stimulus items not contained in the test phase.

Stimuli were presented over Sony MDR-V700 headphones at a comfortable listening volume in sound-attenuated booths. Upon hearing each item, participants were instructed to classify it on the basis of the appropriate attended dimension as quickly and accurately as possible by pressing one of two buttons on a response box.

Results

Percent correct classifications were calculated for each dimension (Table 1). The response latencies for correct trials, measured from the onset of the stimulus to the onset of the buttonpress, were also obtained (Table 1).

Only the latencies of correct responses were submitted for analysis. The data were analyzed using linear mixed-effects regression models (LMER; Baayen, Davidson, & Bates, 2008), with log-transformed reaction times as the dependent variable. Outlier trials that deviated by more than three standard deviations from the mean log reaction time of the condition were excluded from the analysis. The stimulus set was contrast-coded to investigate the following comparisons: control versus correlated (ContCorr) and control versus orthogonal (ContOrtho). Although we included fixed effects in the model to investigate both orthogonal interference (ContOrtho) and redundancy gain (ContCorr), we will report only the orthogonal interference results (see Appendix 1 for the redundancy gain findings), since redundancy gain is not as robust an indicator of perceptual integration as is orthogonal interference.[2] Additional contrast-coded fixed effects included dimension (gender vs. noise) and masking condition (EM, NEM), as well as the interactions of the stimulus set contrasts (ContCorr, ContOrtho) with dimension and masking condition. Random intercepts for participants and items were included. The model also contained random slopes by participants for the stimulus set contrasts and dimensions. Random slopes by items for the stimulus set contrasts, dimensions, and masking conditions were also included. Model comparisons were performed to determine whether the inclusion of each of

[2] Redundancy gain has been found to be possible when processing separable dimensions (Biederman & Checkosky, 1970). It should thus not be taken as robust evidence of perceptual integration in and of itself (see Eimas, Tartter, Miller, & Keuthen, 1978).

Table 1 Mean reaction times (in milliseconds) for masking condition, dimension, and stimulus set with accuracy

Masking	Dim.	Control		Orthogonal		Interference (ms)
		RT (ms)	Accuracy	RT (ms)	Accuracy	Ortho-Ctrl (std. err.)
Energetic	Gender	979	97 %	982	97 %	3 (10)
	Noise	989	94 %	1016	95 %	27 (12)
No energetic	Gender	990	96 %	997	97 %	7 (8)
	Noise	961	96 %	986	95 %	25 (12)

Also shown is the mean orthogonal interference (orthogonal – control), in milliseconds, with standard errors (in parentheses)

these fixed factors and their interactions made a significant contribution to the model.

Figure 2 shows individual participants' mean difference scores, depicting orthogonal interference (orthogonal – control) for each dimension and masking condition. From this figure, it is evident that the majority of listeners showed some interference in each of the conditions (as indicated by positive values). In all, 67 % (EM) and 58 % (NEM) of the participants showed positive interference values for gender classifications, and 64 % (EM) and 72 % (NEM) of the participants for noise classifications. In line with these observations, the results of the LMER analyses revealed a significant main effect of ContOrtho ($\beta = -0.047$, $SE\ \beta = 0.006$, $\chi^2(1) = 40.53$, $p < .05$), whereby reaction times were slower overall in orthogonal than in control conditions.

On the basis of the mean differences between the control and orthogonal conditions, it appears that classification along the noise dimension was subject to greater orthogonal interference than was classification along the gender dimension. Indeed, this was confirmed by a significant ContOrtho × Dimension interaction ($\beta = 0.024$, $SE\ \beta = 0.004$), $\chi^2(1) = 39.35$, $p < .05$. Separate LMERs were performed on the gender and

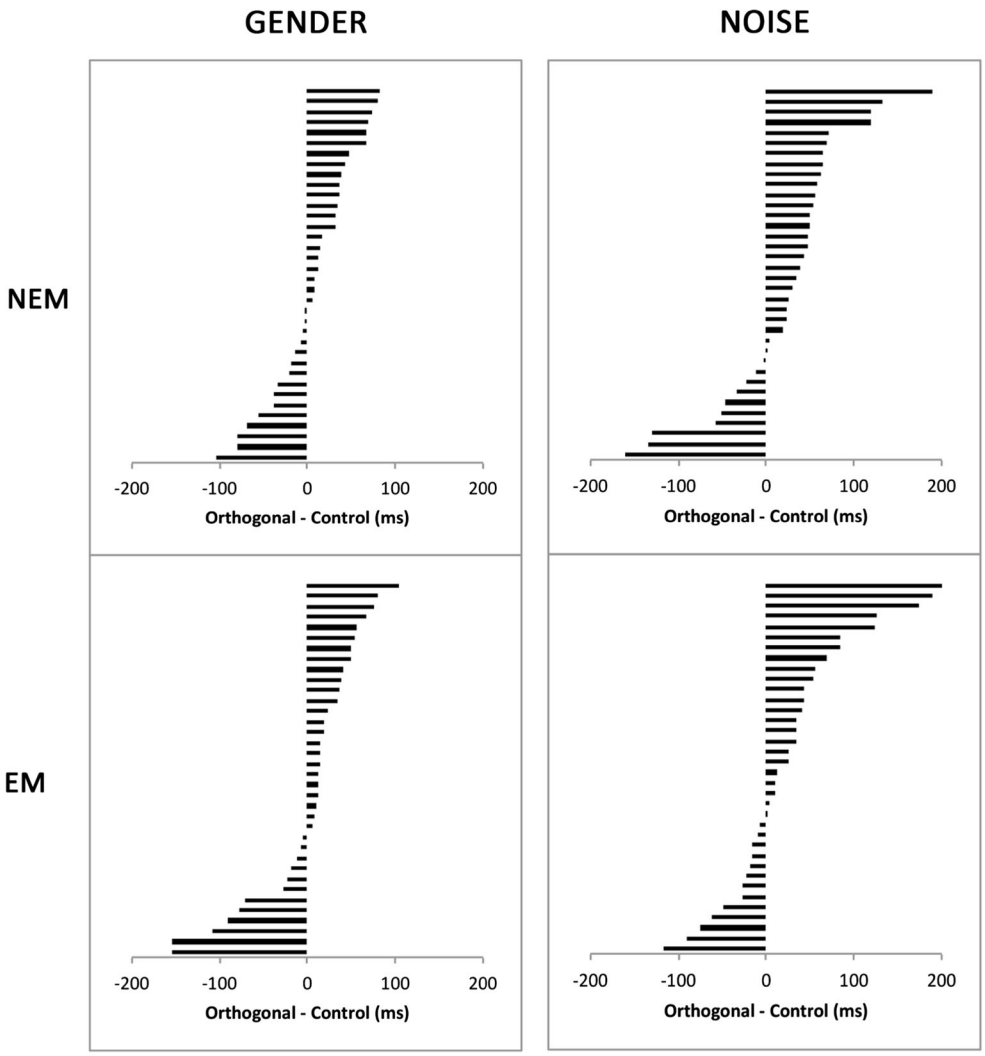

Fig. 2 Mean orthogonal interference (orthogonal – control) for each participant: left column, gender classification; right column, noise classification. Top row, non-energetic-masking (NEM) group; bottom row, energetic-masking (EM) group. Positive values indicate orthogonal interference

noise data with the same fixed- and random-effects structure as above, but with the fixed effect of dimension (and any interactions containing it) removed. For the gender dimension, a significant effect of ContOrtho was found ($\beta = -0.004$, $SE\ \beta = 0.007$), $\chi^2(1) = 23.97$, $p < .05$, indicating that listeners were slowed by irrelevant noise variation when classifying gender in the orthogonal condition, as compared to the baseline control. Similarly, for the noise dimension, a significant main effect of ContOrtho ($\beta = -0.059$, $SE\ \beta = 0.011$), $\chi^2(1) = 26.33$, $p < .05$, revealed that listeners were slower at classifying the noise dimension when the gender dimension varied randomly. No additional main effects (dimension, masking condition) or any other interactions reached significance ($\chi^2 < 2.18$, $p > .05$). In sum, these findings indicate that although significant orthogonal interference was found in both the gender and noise dimensions, the magnitude of this interference differed as a function of the dimension of classification.

In order to determine whether one dimension was inherently easier to discriminate than the other, the reaction times of the control conditions were compared. The LMER model contained fixed effects for dimension and masking condition, as well as random intercepts for participants and items. The model also included random slopes for dimension by participants and by items, as well as a random slope for masking condition by items. The model comparisons revealed no significant effects of dimension, masking condition, or the Masking Condition × Dimension interaction ($\chi^2 < 2.63$, $p > .05$). This indicates that neither spectral separation nor the dimension of classification had a substantive impact on the speed with which listeners made their classifications. This lack of a significant difference may seem surprising, given the relatively smaller, but significant, average differences between control and orthogonal reaction times. However, an examination of the individual participants' data revealed a wide range of individual variation in response speed to classifications of the different dimensions in the control conditions, making any seeming difference not statistically reliable.

The results of Experiment 1A suggest that the processing of gender and the processing of noise information are interdependent, since irrelevant variation in either dimension resulted in interference in classifying the other dimension. However, these interference effects were asymmetric, with irrelevant variation in the gender dimension causing greater interference when classifying noise than in the reverse direction. A comparison of the response latencies in the control conditions for these dimensions indicated that both dimensions were equally discriminable.

In order to further investigate the relationship between baseline classification speed and susceptibility to orthogonal interference, in Experiment 1B we examined whether a different indexical dimension of the speech signal—namely within-gender talker identity—is perceptually integrated with background noise, since talker identity is purportedly more challenging to classify than making a male–female judgment (Cutler et al., 2011). Prior research has suggested that an asymmetry in the discriminability of the dimensions may result in an asymmetry in the magnitudes of the interference effects (e.g., Cutler et al., 2011), with the slower dimension of classification being more susceptible to interference from the faster dimension than the faster dimension is susceptible to the slower one. Specifically, Cutler et al. (2011) found that within-gender talker classification was both slower in the control condition and subject to greater interference from irrelevant phonetic variation, whereas the reverse was the case for gender–phonetic classifications in Mullennix and Pisoni (1990). However, Experiment 1A showed asymmetric integration of extrinsic noise information and intrinsic gender, even though noise and gender classifications were accomplished with comparable speed. Experiment 1B allowed us to examine whether the asymmetry shown in Experiment 1A would reverse as a result of an asymmetry in discriminability, with greater interference effects in the talker than in the noise dimension, as would be predicted by an account that links lower discriminability in the attended dimension to greater interference effects.

Experiment 1B: Perceptual integration of talker identity and background noise

Method

Participants Eighty-seven American English listeners participated in this experiment. All satisfied the same participant criteria outlined in Experiment 1A. On the basis of their classification accuracy performance (<90 % correct) in the Garner task, 15 participants were excluded (nine from the NEM and six from the EM group), resulting in 36 listeners in the NEM (23 female, 13 male; mean age = 20 years) and 36 listeners in the EM (29 female, seven male; mean age = 20 years) condition.

Stimuli The same 96 words from Experiment 1A were used in this experiment, produced by two female American English talkers, one of which was the same female talker as in Experiment 1A. The talker from Experiment 1A had a mean F0 of 230 Hz, whereas the other female talker had a mean F0 of 197 Hz. The first female talker had a mean F0 range of 146 Hz, and the second talker a range of 99 Hz. The mean difference between the male and female talkers from Experiment 1A was 86 Hz, relative to a mean difference of 33 Hz in the present experiment. Identical processing procedures were performed on these speech files, yielding pure-tone- and white-noise-combined stimuli for each of the two masking conditions: NEM and EM.

Procedure Listeners were required to attend to either talker identity by using arbitrarily assigned names (Sue vs. Carol) or noise (white noise vs. pure tone) in one of the two masking conditions (NEM or EM). All other task procedures were identical to those of Experiment 1A. As in the previous experiment, a brief familiarization phase preceded each stimulus dimension condition. In this case, it not only oriented listeners to the task procedures for that particular condition, but also allowed them to learn the names associated with the two talkers' voices. The participants completed three stimulus set conditions (control, correlated, orthogonal) for each stimulus dimension (talker, voice), for a total of six sets of trials.

Results

Percent correct classifications were tabulated for each dimension (Table 2). The mean interference effects as well as the reaction times for correct responses for each dimension and masking condition are also presented in Table 2.

Response latencies (Fig. 3) were log-transformed and analyzed using LMER models. Outliers that satisfied the same criteria as in Experiment 1A were excluded from the analysis. These models contained the same fixed- and random-effects structure as in Experiment 1A, whereby stimulus set was contrast-coded to examine control versus correlated (ContCorr) and control versus orthogonal (ContOrtho), with additional fixed effects of dimension (talker vs. noise) and masking condition (EM, NEM), as well as the interactions of the stimulus set contrasts with dimension and masking condition. As in Experiment 1A, only the orthogonal interference results will be reported here (see Appendix 1 for the redundancy gain findings). Figure 3 reveals that the majority of listeners showed orthogonal interference, with 53 % (EM) and 69 % (NEM) of participants for talker classification and 67 % (EM) and 72 % (NEM) of participants for noise classification showing positive interference values.

Consistent with these observations, a significant main effect of ContOrtho was obtained ($\beta = -0.046$, $SE\ \beta = 0.007$), $\chi^2(1) = 33.20$, $p < .05$, such that participants produced slower reaction times overall in orthogonal than in control conditions. We also found a significant main effect of dimension ($\beta = 0.051$, $SE\ \beta = 0.010$), $\chi^2(1) = 20.80$, $p < .05$, with slower reaction times across conditions when identifying talkers than when identifying noise. The main effect of masking condition did not reach significance ($\chi^2 = 0.44$, $p = .51$).

Furthermore, the mean response latencies (Table 2) suggest that the magnitudes of orthogonal interference in the EM condition were asymmetrical, with greater interference from irrelevant talker variation on noise classification (M = a 37-ms interference effect) than the reverse (M = a 13-ms interference effect). This was reflected in a significant Masking Condition × ContOrtho × Dimension interaction ($\beta = 0.046$, $SE\ \beta = 0.008$), $\chi^2(1) = 31.25$, $p < .05$. Similar LMER analyses, as described above, were conducted to further investigate this three-way interaction. Indeed, a significant effect of ContOrtho was found in both masking conditions of the talker dimension [NEM: $\beta = -0.063$, $SE\ \beta = 0.010$, $\chi^2(1) = 27.94$, $p < .05$; EM: $\beta = -0.038$, $SE\ \beta = 0.012$, $\chi^2(1) = 8.92$, $p = .0028$]. For the noise dimension, the interference effect was significant for the EM condition ($\beta = -0.053$, $SE\ \beta = 0.013$), $\chi^2(1) = 13.31$, $p < .05$, but marginal for the NEM condition ($\chi^2 = 2.78$, $p = .095$).

To determine the relative classification ease for a given dimension, the reaction times of the control conditions were compared. The LMER model contained fixed effects for dimension and masking condition, random intercepts for participants and items, random slopes for dimension by participants and by items, as well as a random slope for masking condition by items. A significant main effect of dimension ($\beta = 0.068$, $SE\ \beta = 0.013$), $\chi^2(1) = 22.72$, $p < .05$, was found, whereby listeners were slower at classifying talker identity than classifying noise in the control condition. No significant effect of masking condition or Dimension × Masking Condition interaction was found ($\chi^2 < 0.61$, $p > .05$).

Table 2 Mean reaction times (in milliseconds) for masking condition, dimension, and stimulus set with accuracy

Masking	Dim.	Control		Orthogonal		Interference (ms)
		RT (ms)	Accuracy	RT (ms)	Accuracy	Ortho-Ctrl (std. err.)
Energetic	Talker	1068	97 %	1081	96 %	13 (12)
	Noise	992	97 %	1029	95 %	37 (12)
No energetic	Talker	1047	97 %	1067	96 %	20 (8)
	Noise	993	96 %	1016	94 %	23 (17)

Also shown is the mean orthogonal interference (orthogonal – control), in milliseconds, with standard errors (in parentheses)

Discussion

The results of Experiment 1 suggest that certain indexical features of speech, such as gender and talker identity, are perceptually integrated with background noise during speech processing, even when the speech and noise signals are spectrally nonoverlapping. Experiment 1A demonstrated mutually dependent processing of gender and noise information, because significant orthogonal interference effects were found for classification along both dimensions. Our findings also revealed a processing asymmetry, whereby listeners were more affected by irrelevant gender variation in the noise classification task than by irrelevant noise variation in the gender classification task. The results from Experiment 1B demonstrated a similar asymmetry with respect to the magnitudes of the interference effect found for classification along each of the two dimensions. However, this interference appeared to be

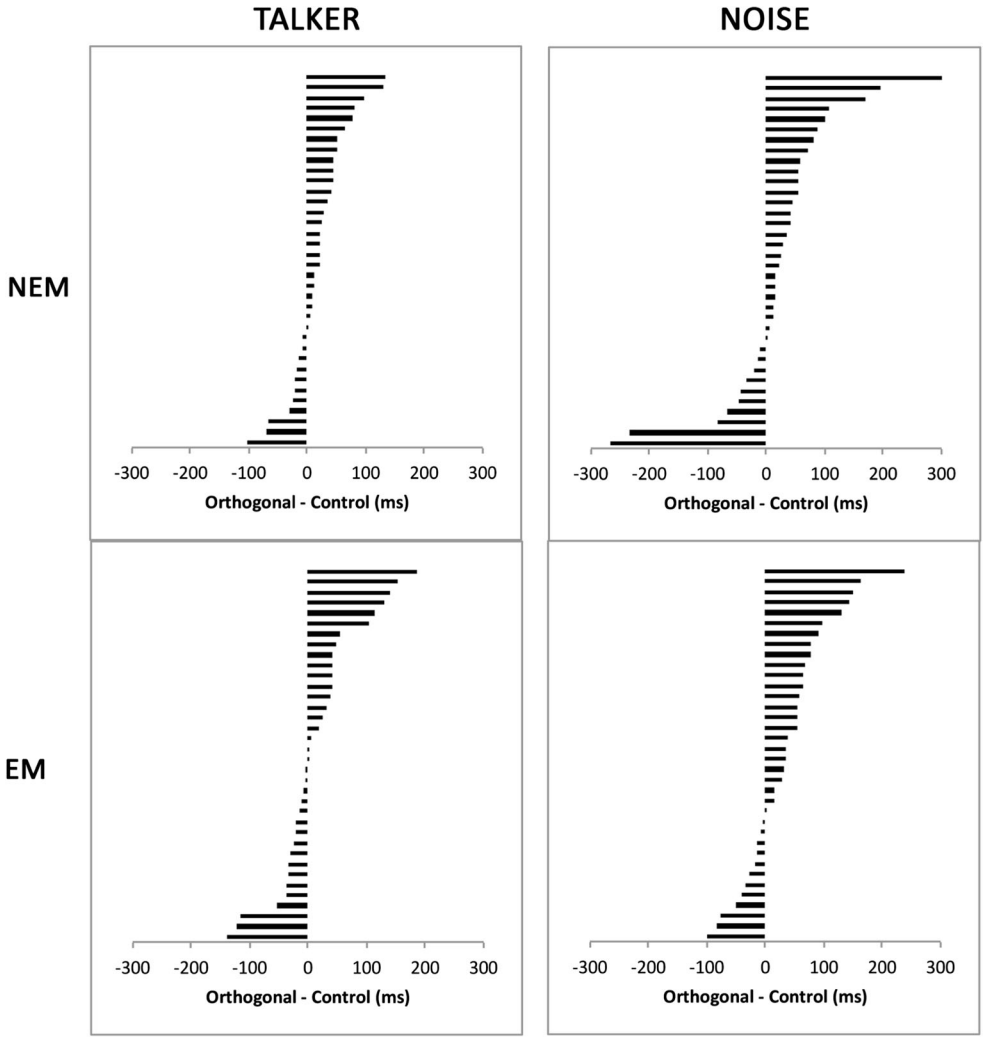

Fig. 3 Mean orthogonal interference (orthogonal – control) for each participant: left column, talker classification; right column, noise classification. Top row, non-energetic-masking (NEM) group; bottom row, energetic-masking (EM) group. Positive values indicate orthogonal interference

modulated by masking condition, since it was found in both the EM and NEM conditions for the talker dimension, but only in the EM condition for the noise dimension (although there was a trend toward interference in the NEM condition).

Additionally, Experiments 1A and 1B allowed us to examine the role of discriminability in the magnitudes of these interference effects, since talker identity is relatively more difficult to discriminate than gender. Cutler et al. (2011) posited a relationship between the sizes of orthogonal interference effects and how difficult it is to classify a given dimension (as indexed by reaction times), such that more difficult decisions will yield longer reaction times and, subsequently, greater interference effects. However, on the basis of the present experiments, the interference asymmetries found in Experiments 1A and 1B do not appear to be related to a discrepancy in classification difficulty between the noise and indexical dimensions. Indeed, asymmetric orthogonal interference was found in Experiment 1A, in which there was no significant difference in discriminability between the gender and noise dimensions in the control conditions. Furthermore, in Experiment 1B, a greater degree of orthogonal interference was found for the noise dimension, despite the fact that listeners were slower to make talker identity classifications than to make noise classifications in the control conditions. Thus, given these findings, inherent processing difficulty does not appear to be the primary factor influencing the directionality and magnitude of orthogonal interference effects.

The present results extend previous work examining the processing dependencies in speech perception (e.g., Mullennix & Pisoni, 1990). Prior work had reported that indexical and linguistic properties of the speech signal are perceptually integrated during speech processing. The findings of the present study suggest that listeners integrate indexical features of the speech signal with temporally concurrent auditory

information—in this case, background noise. However, the processing asymmetries found for noise and both indexical speech properties (gender and talker identity) indicate that although context-specific information and indexical speech information are coupled, they are unevenly weighted during processing.

One possible explanation for the asymmetry between the speech and noise dimensions pertains to the relative salience of these dimensions. The Garner (1974) task involves selective attention, whereby listeners must attend to one dimension while ignoring the other. Tong, Francis, and Gandour (2008), examining the processing dependencies between consonants, vowels, and lexical tones in Mandarin Chinese, found that irrelevant segmental variation led to greater interference for lexical-tone classification than did irrelevant tone variation for segmental classification. Tong et al. posited that the information value of a given dimension could play a substantive role in selective attention, such that listeners may opt to attend to features that are more informative in resource-demanding situations, resulting in an asymmetry between dimensions in their susceptibilities to orthogonal interference. In their study, information value was determined by calculating the probability of the dimension occurring in a communicative system, a criterion by which segmental information is substantially more informative than tone information. In the context of the present findings, noise could be considered to provide less information generally for listeners, and thus to be less salient than linguistic information, making it more susceptible to interference from variation in a more-salient dimension. Indeed, from infancy, humans are purportedly biased toward listening to speech over nonspeech (Vouloumanos & Werker, 2007). Although we cannot quantify the relative saliences of speech versus noise by the same metric used by Tong et al., the relatively greater functional relevance of speech over noise should not be controversial. Thus, it could be that the observed processing asymmetries between dimensions that are intrinsic to the speech signal and extraneous background noise result from asymmetries in the information value of the dimensions being processed, with gender and talker features (speech-intrinsic dimensions) having greater information value than noise and pure-tone information (a speech-extrinsic, background dimension).

We note that the size of these interference effects are smaller than those in prior work with the Garner task using speech stimuli (e.g., Mullennix & Pisoni, 1990). This occurred despite the fact that the overall response times are relatively long (averages of 979–1081 ms in this study, as compared to 456–657 ms in Mullennix & Pisoni, 1990). One possible explanation is that low variability along the classification dimensions may have led to relatively smaller interference effects. For instance, Mullennix and Pisoni found that increasing the number of talkers for gender classification (up to 16) led to more robust orthogonal interference. In the present study we employed just two talkers and two noise types, which could have contributed to smaller interference effects. Moreover, it is also conceivable that dimensions intrinsic to the speech signal are more robustly perceptually integrated, by virtue of the fact that cues to classification of speech-intrinsic dimensions may be co-present within the same signal. For example, upon hearing the word "pill," cues that identify the initial consonant can in part be used to identify the talker. However, with speech and noise dimensions, the speech signal does not hold any cues to help identify the noise type, which may result in relatively smaller orthogonal interference effects. With regard to the relatively long average reaction times in the present study, Mullennix and Pisoni noted that as the number of individual items increased, so too did the reaction times. If one considers the amount of item variability in the present experiment, the longer reaction times are perhaps not surprising, given that there were 96 different disyllabic words (relative to between two and 16 different monosyllabic words in Mullennix & Pisoni, 1990). Moreover, none of the 96 items were repeated within a given condition, unlike in Mullennix and Pisoni (1990), in which items were repeated between four and 32 times within a condition. These factors likely contributed to the overall longer reaction times observed in the present work.

Although the present study suggests integral processing of speech-intrinsic and -extrinsic features, one could also consider an alternative explanation for these findings that appeals to low-level processing mechanisms.[3] The information necessary to distinguish between two levels of a particular dimension (e.g., two female talkers) is in part carried by the frequency composition of the signal. In order to make the appropriate classifications, listeners must extract information about the relative frequency characteristics of the two talkers. It is conceivable that the interference of noise with gender and talker classification demonstrated in Experiment 1 may have arisen as a result of masking in the EM condition or of some spread of masking in the NEM condition. For example, in the EM condition, some of the indexical characteristics of the talkers may have been masked, since the noise overlapped with some parts of the spectra that carried the talker information. Moreover, it is possible that even in the NEM condition, despite the noise and the speech signal being spectrally separated, a spread of masking could have in part obscured the frequency composition necessary to make gender or talker classifications.

This explanation would likely predict a differential in gender or talker classification difficulty as a function of the type of noise presented concurrently with the speech signal, such that the presence of band-pass-filtered white noise should yield greater masking, and consequently slower response speeds, than the presence of a pure tone. However, a comparison of

[3] We thank James Sawusch for pointing out this alternative explanation.

the reaction times within both the gender and talker control conditions (in which the noise background was consistent) found no significant differences between classifications made in the band-pass-filtered white-noise versus pure-tone conditions, or any significant interaction with energetic masking (EM or NEM), $\chi^2 < 1.32$, $p > .25$, suggesting that listeners were not slower on the gender or talker classification tasks in the more heavily masked band-pass-filtered noise condition than in the single-frequency masking of the pure-tone condition, as would be predicted by a purely low-level explanation for the interference effects found in Experiment 1. It remains possible that the observed interference from irrelevant noise variation for gender and talker classification was due to masker uncertainty in the orthogonal condition, rather than to perceptual integration of the speech and noise signals, but this too would implicate a central (informational masking) rather than peripheral (energetic masking) locus for the observed interference effect. It remains for future work to determine exactly how speech and background noise interfere with each other during classification along noise and speech dimensions, respectively, but the presently available evidence seems to implicate some degree of higher-level processing involvement. In Experiment 2, we sought to provide further evidence of the integrality, or at least persistent association, of concurrently presented speech and noise from a task that taps into a later stage of processing—namely, the continuous recognition memory paradigm.

Experiment 2: Continuous recognition memory

In Experiment 2, we examined whether background noise is encoded and represented in memory for spoken words in a continuous recognition memory paradigm, with two critical departures from previous work that had addressed the encoding of speech and concurrently presented background noise (e.g., Creel et al., 2012; Pufahl & Samuel, 2014). First, considering that the background noise variation implemented in both experiments of the present study was neither intrinsic to the speech signal nor phonetically relevant (Sommers, Nygaard, & Pisoni, 1994), we anticipated that its influence on the recognition of a word as having occurred previously within a test session would more closely resemble the influence of amplitude variation than that of either talker or speaking rate variation. Bradlow et al. (1999) directly compared the influences of talker, speaking rate, and amplitude variation on recognition memory for spoken words with both implicit and explicit versions of the task. Whereas both talker and rate consistency facilitated old-item recognition in both the implicit and explicit versions of the task, an influence of amplitude consistency only emerged in the explicit version (described in detail below).[4] We therefore adopted this task as a means of investigating whether the integration of speech and noise extends beyond perceptual classification to recognition memory for spoken words. Secondly, whereas the stimuli in previous studies had involved speech and noise signals that overlapped in both the temporal and spectral domains, in Experiment 2 we used exactly the same stimuli as in Experiment 1, in which the effects of spectral overlap versus spectral separation were directly compared.

Method

Participants Forty-four monolingual American English listeners from the undergraduate student body at Northwestern University participated in this experiment, none of whom had also participated in Experiment 1. They reported having normal speech and hearing and were paid for their participation. Twenty of these participants were included in the EM condition (15 female, five male; M_{age} = 19 years) and 24 in the NEM condition (21 female, three male; M_{age} = 19 years).

Stimuli The stimulus materials consisted of 139 disyllabic spoken words, of which 96 were experimental items, 15 were memory load items, 20 were filler items, and eight were practice items. The four different noise and masking versions (NEM and EM with both a pure tone and narrow-band noise) of the 96 experimental items were taken from Experiment 1B, from just one of the female speakers. A pretest determined that participants (n = 16, none of whom also participated in the recognition memory test or in Exp. 1) recognized the words in both noise types with a very high degree of accuracy (≥93 %).

Procedure Listeners were tested in a sound-attenuated booth, and the stimuli were presented binaurally over headphones at a comfortable listening volume. A list started with 15 practice trials and 30 memory load trials. These memory load items were included to equate performance between the stimuli occurring early in the list and the stimuli occurring later in the list. Twenty filler items were randomly presented interspersed with the 192 test trials. Each item (word + noise) was presented and repeated once (except for the fillers) after a lag of 4, 8, or 16 intervening items, with the repetition itself counting as the last intervening item. All lags occurred with equal frequencies. The word was repeated with the same noise (pure tone followed by pure tone, or white noise followed by white noise) or with a

[4] The results from explicit memory studies, particularly those utilizing traditional recognition memory tasks, have had mixed success in demonstrating specificity effects (see Goh, 2005, and Pufahl & Samuel, 2014, for reviews). However, as Goh noted, the continuous recognition memory task has yielded more consistent findings.

different type of noise (pure tone followed by white noise, or vice versa). The probabilities of a same-noise versus a different-noise repetition were equal. On each trial, listeners were required to choose from one of three response options: "old–same" (heard before in the list with the same noise), "old–different" (heard before in the list but with a different noise), or "new" (the word was new to the list). The session lasted about 15 min.

2. Results

The practice, memory load, and filler items were not included in the final data analysis. The mean accuracy scores by response types (old–same, old–different, and new) for each masking condition and lag are provided in Appendix 2. Following Bradlow et al. (1999), *d*-prime scores were calculated. A hit was defined as an "old–same" response to a stimulus that was repeated with the same noise. A false alarm was defined as an "old–same" response to a stimulus that was repeated with different noise. This measure allowed us to determine whether listeners could explicitly recognize variation in noise while also accounting for any response bias. At each lag, a one-sample *t* test determined whether the *d*-prime score differed significantly from zero.

Figure 4 shows the mean *d*-prime scores at each lag for the EM (A) and NEM (B) conditions, and Table 3 provides the descriptive and test statistics by lag and masking condition. For the EM condition, *t* tests revealed that the *d*-prime scores were significantly greater than zero, except at lag 16. This is consistent with previous research showing that *d*-prime decreases as a function of lag (e.g., Bradlow et al., 1999). For the NEM condition, *t* tests indicated that at no lag was the *d*-prime score significantly greater than zero, indicating that spectral separation of the speech and noise effectively eliminated any cost of inconsistent background noise for spoken word recognition memory.

Next, we used LMER models with *d*-prime scores as the dependent variable and with lag (4 vs. 8 vs. 16) and masking (NEM vs. EM) as fixed effects to directly compare the *d*-prime scores across conditions. Participant was included as a random intercept and lag was included as a random slope by participant. Lag was centered (i.e., 0 at lag 8) and masking was contrast-coded (−0.5 vs. +0.5). Model comparisons were performed to determine whether the inclusion of each of these fixed effects made a significant contribution to the model. This analysis demonstrated significant main effects of lag ($\beta = -0.024$, $SE\ \beta = 0.098$), $\chi^2(1) = 6.29$, $p < .05$, and of masking ($\beta = -0.523$, $SE\ \beta = 0.195$), $\chi^2(1) = 6.96$, $p < .01$. The interaction between masking and lag did not reach significance ($\beta = -0.030$, $SE\ \beta = 0.019$), $\chi^2(1) = 2.68$, $p = .10$.

Table 3 Mean *d*-prime scores (with standard errors in parentheses) and *t* and *p* values, as well as the mean percent hit and false alarm rates (with standard errors again in parentheses) for each lag and masking condition

	Lag 4	Lag 8	Lag 16
EM			
d′	0.42 (0.20)	0.62 (0.20)	0.38 (0.22)
t (*p*)	2.12 ($p < .05$)	2.94 ($p < .01$)	1.72 ($p < .10$)
Hit rate	68 (4.05)	68 (4.10)	66 (3.40)
False alarm rate	54 (4.54)	45 (5.17)	53 (5.29)
NEM			
d′	0.12 (0.11)	−0.02 (0.13)	−0.35 (0.14)
t (*p*)	1.08 ($p = .29$)	−0.12 ($p = .90$)	−2.43 ($p = .02$)
Hit rate	68 (3.15)	60 (3.70)	51 (3.16)
False alarm rate	66 (2.90)	61 (2.88)	63 (3.96)

EM energetic masking, *NEM* no energetic masking

Overall, these results revealed a performance decrement when recognizing that the word of a current trial had occurred earlier in the list with a different background noise only when the noise was spectrally overlapping with the speech signal and when the repeated item was presented less than 16 trials later than the first occurrence. At later lags and when the noise and speech were spectrally separated, there was no different-noise cost in performance.

Discussion

The results of Experiment 2 indicate that the influence of co-occurring background noise on recognition memory for spoken words is modulated by the spectral separation of the speech and noise signals: An inconsistent noise cost in this explicit memory task was only found when the noise and speech were spectrally overlapping. The significant positive *d*-prime scores of 0.4 and 0.6 for lags of 4 and 8 words, respectively, are consistent with the *d*-primes of the

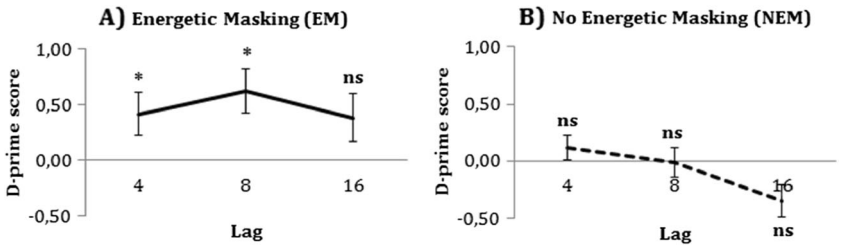

Fig. 4 Mean *d*-prime scores as a function of lag for the EM (A) and NEM (B) conditions. Asterisks indicate scores significantly greater than zero, and "n.s." indicates scores not significantly greater than zero

comparable task in Bradlow et al. (1999; i.e., an explicit task with amplitude as the source of variation from the first to the second occurrence yielded a *d*-prime range of 0.3–1.2 across lags 2, 8, 16, and 32). Thus, the present findings indicate that background noise may be included amongst the nonlinguistic attributes of a spoken word that may be encoded or associated with the cognitive representation of the word in the mental lexicon. The present findings suggest that the effect of background noise variation is comparable to that of amplitude variation. Unlike talker or rate variation—both of which are linguistically relevant sources of variation in the sense that the acoustic cues to phonemic categories show talker and rate dependencies—both amplitude and background noise are less likely to exert comparable influences on linguistic category perception (see Sommers & Barcroft, 2006, and Sommers et al., 1994, for more on the phonetic relevance hypothesis). However, a crucial difference between amplitude and background noise is that, whereas amplitude variation is an acoustic attribute of the speech signal itself, background noise emanates from a separate source and is extraneous to the speech signal. Thus, the present findings extend prior work by providing evidence that, under certain circumstances, even variation that is clearly from a different source than the speech signal can influence spoken word encoding and retrieval.

General discussion

Taken together, the results of Experiments 1 and 2 provide insight into the extent and limits of the influence of co-occurring nonlinguistic acoustic information on the perception and encoding of spoken language. In Experiment 1, we investigated perceptual integration versus segregation at an early stage of processing with the Garner (1974) speeded classification paradigm. In Experiment 2, we then used the same stimuli to probe the joint encoding of spoken words and background noise with a continuous recognition memory paradigm. Critically, both experiments compared listener responses under conditions of either spectral overlap or spectral segregation between the speech and noise stimuli. The results of Experiment 1 suggest that speech and background noise are perceptually integrated at the level of processing tapped into by the Garner speeded classification task. The results point to an interdependence of the perceptual processes used to encode information about background noise with indexical information in the speech signal—specifically, gender (Exp. 1A) and talker identity (Exp. 1B)—and this interdependence at the level of perceptual classification appears to be largely independent of whether the noise and speech are spectrally overlapping or spectrally separated. Furthermore, a general pattern of asymmetry emerged, whereby noise classifications tended to be more adversely affected by variation in the speech signal than speech classifications were affected by noise variation. Although previous work has put forth the relative discriminability of the dimensions as a possible motivation for asymmetric dependencies, so that less discriminable or "harder" dimensions are more susceptible to interference than more discriminable or "easier" dimensions (e.g., Cutler et al., 2011), we can reject that possibility here, on the basis of the findings of the present work. The noise dimension, which was either easier to discriminate than (Exp. 1B) or equally discriminable as (Exp. 1A) the speech dimension, suffered greater interference from the speech dimension than the speech dimension suffered from it. We posit that the observed asymmetry may have arisen from a discrepancy in the information values of the speech and noise dimensions, with speech being relatively more informative (and thus exerting greater interference) than noise.

However, as we discussed in Experiment 1, an alternative explanation for these interference effects could be attributed to low-level processing mechanisms, whereby frequency masking from the noise and speech signals could have obscured key features that would have enabled classification. Although this alternative cannot be ruled out definitively, there are several reasons to suspect that the observed interference effects were not entirely due to a low-level analysis of the signal. First, this explanation would predict a differential in speech processing as a function of noise type, such that narrow band-pass-filtered white noise should cause greater masking than a pure tone, due to the greater energetic masking of the noise band than the pure tone (i.e., 3- to 4-kHz or 7- to 10-kHz bands vs. 3- or 6-kHz tones). However, as we noted above, the reaction times for classifications of speech as spoken by a male or a female (Exp. 1A) or by one of two females (Exp. 1B) did not significantly differ by noise type. A low-level explanation could perhaps be maintained for the noise classification conditions according to which it is harder to compare the amounts of masking by the various talkers. However, for the classification dimensions (noise and gender for Exp. 1A, noise and talker for Exp. 1B), the same masking was present in both the control and orthogonal conditions. Thus, the fact that the orthogonal conditions showed slower reaction times than the control conditions suggests that something was slowing processing above and beyond what could be accounted for just by low-level masking.

In Experiment 2, we used the same stimuli as in Experiment 1 in a continuous recognition memory paradigm, examining whether listeners' ability to discriminate new (first occurrence) from old (second occurrence) words in a list of spoken words was affected by having consistent versus varying background noise from the first to second occurrence. An explicit version of this task was used, in which participants' attention was explicitly drawn to the background noise by requiring them to indicate whether a word recognized as old (i.e., repeated in the list) had consistent (old–same response)

or varying (old–different response) background noise, relative to the first occurrence. Thus, unlike the speeded classification (Garner) task of Experiment 1, in which each trial could be responded to without reference to a previous trial, this task required participants to assess the match between two instances of a spoken word. We observed that the perception of spoken words could be affected by a change in the background noise. However, our results point to a constraint on the integration of speech and noise, such that the influence of noise inconsistency was only observed under the condition of spectral overlap between the speech and the background noise.

Together, the pair of experiments probed different aspects of the influence of background noise on the perception and encoding of speech. Experiment 1 focused on classification of speech + noise signals without necessarily requiring the involvement of the mental lexicon. That is, our implementation of the speeded classification (Garner) task with noise (tone or broadband) as one dimension and either gender (Exp. 1A) or talker (Exp. 1B) as the other dimension could be performed without contact with the mental lexicon. Thus, the overall finding of speech + noise perceptual integration in Experiment 1 suggests a rejection of models of speech processing in which speech signals and environmental signals extraneous to the attended speech signal are segregated very early in processing. However, the results of this experiment do not speak directly to the nature of lexical representations themselves.

Experiment 2 focused on memory for spoken words in a way that likely involved contact with the mental lexical, but at the same time required quite limited lexical or linguistic processing. In the continuous recognition memory paradigm, participants were not required to recognize the word itself, to perform any linguistic judgment, or to conduct any syntactic, semantic, or pragmatic processing. Nevertheless, it is likely that for word (rather than nonword or nonspeech) stimuli, the task was performed with some access to the mental lexicon. Thus, the finding of a recognition performance cost for words presented with inconsistent background noise across the first and second occurrences in the list in the EM condition of Experiment 2 raises the possibility that the integration of speech + noise suggested by the findings of Experiment 1 impinges on lexical processing as well as on extralexical speech classification. These findings thus may have implications for word recognition. Although prior work (e.g., Creel et al., 2012; Pufahl & Samuel, 2014) had found evidence for the integrality of speech and background noise, demonstrated by a decrement in word recognition performance when the noise changed from exposure to test, the negative influence on word recognition of background noise inconsistency might be attenuated, or even eliminated, when the speech and background noise are easily segregated in the spectral domain, given the NEM results of Experiment 2. Future research on the types of nonlinguistic information that listeners utilize during word recognition may consider comparing conditions of spectral segregation and spectral integration between linguistic and nonlinguistic dimensions, since the degree to which listeners include this extraneous information in the word recognition process may be modulated by how easily the information can be segregated from the speech signal.

It is important to note that neither the speeded classification (Garner) nor the continuous recognition memory task requires lexical processing per se, and therefore the present work does not directly address the nature of lexical representations. Instead, the present work builds a case against models of speech + noise processing that involve early segregation of speech and concurrently presented environmental noise, and indicates potentially far-reaching effects of the specific conditions under which speech is experienced by listeners. Combined with converging evidence from other paradigms (e.g., Creel et al., 2012; Pufahl & Samuel, 2014) that involve lexical and/or other levels of linguistic processing, the present study contributes to a more comprehensive understanding of the cognitive and linguistic consequences of speech perception under adverse listening conditions.

Author note This research was supported by NIH-NIDCD Grant No. R01-DC005794, awarded to A.R.B. We thank Chun Liang Chan, Charlotte Vaughn, Vanessa Dopker, and Emily Kahn for their research and technical support. Thanks also to Matt Goldrick for advice on the statistical analyses and earlier drafts of the manuscript.

Appendix 1: Control versus correlated data

Gender versus noise classification The results of model comparisons in Experiment 1A revealed a significant main effect of ContCorr ($\beta = 0.064$, $SE\ \beta = 0.006$), $\chi^2(1) = 65.04$, $p < .05$, such that listeners were faster in correlated than in control conditions overall. A significant ContCorr × Dimension interaction ($\beta = –0.010$, $SE\ \beta = 0.004$), $\chi^2(1) = 7.08$, $p = .008$, was also found. Separate LMERs were performed on the gender and noise data with the same fixed- and random-effects structure as we described above, but with the fixed effect of dimension (and any interactions containing it) removed. For the gender dimension, a significant main effect of ContCorr ($\beta = 0.059$, $SE\ \beta = 0.007$), $\chi^2(1) = 45.08$, $p < .05$, was found, such that listeners were faster at identifying the gender of the talker in the correlated than in the control condition across masking conditions. Similarly, for the noise dimension, a significant main effect of ContCorr ($\beta = 0.070$, $SE\ \beta = 0.011$), $\chi^2(1) = 33.46$, $p < .05$, revealed that listeners were faster at classifying noise when the gender dimension covaried.

An examination of the mean differences between the control and correlated conditions for each dimension and masking condition revealed a stronger redundancy gain for noise classification than for gender classification in the EM condition

Table 4 Mean reaction times (in milliseconds) for masking condition, dimension, and stimulus set with accuracy

Masking	Dim.	Control		Correlated		Redundancy Gain (ms)
		RT (ms)	Accuracy	RT (ms)	Accuracy	Ctrl-Corr (std. err.)
Energetic	Gender	979	97 %	936	98 %	43 (9)
	Noise	989	94 %	935	96 %	54 (17)
No energetic	Gender	990	96 %	952	98 %	38 (11)
	Noise	961	96 %	927	97 %	34 (9)

Also shown is the mean redundancy gain (control – correlated), in milliseconds, with standard errors in parentheses

(Table 4). However, in the NEM condition, there appeared to be little difference between the dimensions in the magnitudes of the redundancy gain. Indeed, this was reflected in a significant three-way Masking Condition × ContCorr × Dimension interaction ($\beta = -0.016$, $SE\ \beta = 0.008$), $\chi^2(1) = 7.81$, $p = .02$. Separate LMERs were conducted for each masking condition and dimension. All of the comparisons generated significant results ($\chi^2 > 17.45$, $p < .05$), indicating that although covarying dimensions yielded significantly faster reaction times in all conditions, there was an asymmetry in the magnitudes of the redundancy gains.

Talker and noise classification Model comparisons in Experiment 1B yielded a significant main effect of ContCorr ($\beta = 0.049$, $SE\ \beta = 0.008$), $\chi^2(1) = 30.92$, $p < .05$, whereby listeners were faster overall in correlated than in control conditions. On the basis of the mean response latencies in Table 5, we see redundancy gains in both masking conditions of the talker dimension; however, a redundancy gain only appears to be present in the EM condition of the noise dimension. This was reflected by the significant three-way Masking Condition × ContCorr × Dimension interaction ($\beta = -0.060$, $SE\ \beta = 0.008$), $\chi^2(1) = 52.91$, $p < .05$. Subsequent LMERs were performed on the talker and noise data from each masking condition separately, with the same fixed structure as the full model but with the fixed effects of dimension and masking condition (and any interactions containing them) removed. Random intercepts for participants and items were included, as well as random slopes for stimulus set contrasts by participants and items. ContCorr was significant in both masking conditions of the talker dimension [NEM: $\beta = 0.090$, $SE\ \beta = 0.014$, $\chi^2(1) = 26.387$, $p < .05$; EM: $\beta = 0.052$, $SE\ \beta = 0.014$, $\chi^2(1) = 11.919$, $p < .05$]. However, for the noise dimension, it was significant in the EM condition ($\beta = 0.039$, $SE\ \beta = 0.015$), $\chi^2(1) = 6.2132$, $p = .013$, but not in the NEM condition ($\chi^2 = 1.498$, $p = .22$).

Table 5 Mean reaction times (in milliseconds) for masking condition, dimension, and stimulus set with accuracy

Masking	Dim.	Control		Correlated		Redundancy Gain (ms)
		RT (ms)	Accuracy	RT (ms)	Accuracy	Ctrl-Corr (std. err.)
Energetic	Talker	1068	97 %	1033	98 %	35 (14)
	Noise	992	97 %	980	97 %	12 (13)
No energetic	Talker	1047	97 %	988	97 %	59 (13)
	Noise	993	96 %	992	97 %	1 (11)

Also shown is the mean redundancy gain (control – correlated), in milliseconds, with standard errors in parentheses

Appendix 2

Table 6 Mean percent correct for the "old–same," "old–different," and "new" response types (with standard errors in parentheses) for lag and masking conditions

	Lag 4	Lag 8	Lag 16
EM			
Old–same	68 (3.15)	60 (3.70)	51 (3.16)
Old–different	34 (2.90)	39 (2.88)	37 (3.96)
New	92 (1.35)	92 (1.21)	94 (1.32)
NEM			
Old–same	68 (4.05)	68 (4.10)	66 (3.40)
Old–different	46 (4.54)	56 (5.17)	47 (5.29)
New	90 (1.86)	93 (0.99)	93 (1.19)

EM energetic masking, *NEM* no energetic masking

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Biederman, I., & Checkosky, S. F. (1970). Processing redundant information. *Journal of Experimental Psychology, 83,* 486–490. doi:10.1037/h0028841

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program]. Retrieved from www.praat.org

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61,* 206–219. doi:10.3758/BF03206883

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and

fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 521–533. doi:10.1037/0278-7393.20.3.521

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes, 27,* 1021–1038. doi:10.1080/01690965.2011.610597

Cutler, A., Andics, A., & Fang, Z. (2011). Inter-dependent categorization of voices and segments. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS XVII)* (pp. 552–555). Hong Kong: City University of Hong Kong, Department of Chinese, Translation and Linguistics.

Eimas, P. D., Tartter, V. C., Miller, J. L., & Keuthen, N. J. (1978). Asymmetric dependencies in processing phonetic features. *Perception & Psychophysics, 23,* 12–20. doi:10.3758/BF03214289

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Goh, W. D. (2005). Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 40–53. doi:10.1037/0278-7393.31.1.40

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1166–1183. doi:10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251–279. doi:10.1037/0033-295X.105.2.251

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics, 59,* 675–692. doi:10.3758/BF03206015

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science, 24,* 1995–2004. doi:10.1177/0956797613482467

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34,* 485–499. doi:10.1016/j.wocn.2005.08.004

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114,* 161–172. doi:10.1016/j.brainres.2006.07.049

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27,* 953–978. doi:10.1080/01690965.2012.705006

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47,* 379–390. doi:10.3758/BF03210878

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–378.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics, 35,* 85–103. doi:10.1016/j.wocn.2005.10.004

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5,* 42–46. doi:10.1111/j.1467-9280.1994.tb00612.x

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328. doi:10.1037/0278-7393.19.2.309

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam, The Netherlands: Benjamins.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology, 70,* 1–30. doi:10.1016/j.cogpsych.2014.01.001

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915–930. doi:10.1037/0278-7393.18.5.915

Sommers, M. S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *Journal of the Acoustical Society of America, 119,* 2406–2416. doi:10.1121/1.2171836

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96,* 1314–1324.

Tomiak, G. R., Mullennix, J. W., & Sawusch, J. R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *Journal of the Acoustical Society of America, 81,* 755–764.

Tong, Y., Francis, A. L., & Gandour, J. T. (2008). Processing dependencies between segmental and suprasegmental features in Mandarin Chinese. *Language and Cognitive Processes, 23,* 689–708. doi:10.1080/01690960701728261

Vouloumanos, A., & Werker, J. F. (2007). Listening to language at birth: Evidence for a bias for speech in neonates. *Developmental Science, 10,* 159–164. doi:10.1111/j.1467-7687.2007.00549.x