Auditory neural tracking and lexical processing of speech in noise: Masker type, spatial location, and language experience

Jieun Song, Luke Martin, and Paul Iverson

Citation: The Journal of the Acoustical Society of America **148**, 253 (2020); doi: 10.1121/10.0001477
View online: https://doi.org/10.1121/10.0001477
View Table of Contents: https://asa.scitation.org/toc/jas/148/1
Published by the Acoustical Society of America

JASA ARTICLE

Auditory neural tracking and lexical processing of speech in noise: Masker type, spatial location, and language experience

Jieun Song,[a] Luke Martin, and Paul Iverson
Department of Speech, Hearing and Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London, WC1N 1PF, United Kingdom

ABSTRACT:
The present study investigated how single-talker and babble maskers affect auditory and lexical processing during native (L1) and non-native (L2) speech recognition. Electroencephalogram (EEG) recordings were made while L1 and L2 (Korean) English speakers listened to sentences in the presence of single-talker and babble maskers that were colocated or spatially separated from the target. The predictability of the sentences was manipulated to measure lexical-semantic processing (N400), and selective auditory processing of the target was assessed using neural tracking measures. The results demonstrate that intelligible single-talker maskers cause listeners to attend more to the semantic content of the targets (i.e., greater context-related N400 changes) than when targets are in babble, and that listeners track the acoustics of the target less accurately with single-talker maskers. L1 and L2 listeners both modulated their processing in this way, although L2 listeners had more difficulty with the materials overall (i.e., lower behavioral accuracy, less context-related N400 variation, more listening effort). The results demonstrate that auditory and lexical processing can be simultaneously assessed within a naturalistic speech listening task, and listeners can adjust lexical processing to more strongly track the meaning of a sentence in order to help ignore competing lexical content.
 https://doi.org/10.1121/10.0001477
(Received 11 November 2019; revised 9 June 2020; accepted 9 June 2020; published online 16 July 2020)
[Editor: Adrian K. C. Lee] Pages: 253–264

I. INTRODUCTION

Speech perception in everyday noisy situations (e.g., parties or restaurants) is complex because these contexts put simultaneous demands on multiple levels of processing. For example, noises mask the acoustic information of a speaker at the auditory periphery, the listener must perceptually track the variable acoustics of the speaker's voice through a background of similar speakers from multiple spatial locations, and the listener must follow the meaning of the conversation while ignoring what other people are saying (e.g., Brungart, 2001; Shinn-Cunningham, 2008; Cooke *et al.*, 2008). This situation becomes more difficult when understanding speech in a non-native (L2) language; noise may have a greater effect on L2 listeners because their perceptual and linguistic processes are not as well developed for their L2 (see Lecumberri *et al.*, 2010, for a review), and it is also possible that the perceptual and cognitive demands of L2 speech communication reduce the spare capacity to focus attention in difficult listening conditions (e.g., Kahneman, 1973; Pichora-Fuller *et al.*, 1995; McCoy *et al.*, 2005).

The present study assessed the effect of intelligible maskers (i.e., single talker vs babble) and the spatial separation of the target and masker for native (L1) and L2 speakers of English, using simultaneous neural measures of auditory and lexical processing that can be applied to naturalistic speech recognition tasks (e.g., listening to podcasts or sentences). The primary aim was to use manipulations that affect auditory attention for speech (i.e., intelligible vs unintelligible maskers, spatial separation of target and maskers) to help understand previous listening-effort differences found for L1 and L2 listeners (Song and Iverson, 2018). The more general aim was to use simultaneous electroencephalogram (EEG) measures of auditory and lexical processing to provide a more integrated view of speech recognition under challenging conditions.

The ability to auditorily track a target talker through noise was assessed using analyses of EEG recordings that measure the degree to which neural activity in the auditory cortex becomes phase-locked to slow amplitude fluctuations in the speech signal (2–8 Hz; e.g., Ahissar *et al.*, 2001; Luo and Poeppel, 2007; see Ding and Simon, 2014, for a review). In general, listeners have stronger neural tracking to talkers they were asked to attend to than to talkers or noises they were are asked to ignore (e.g., Ding and Simon, 2012; Kerlin *et al.*, 2010).

Lexical processing was quantified using the N400 response, an event-related potential (ERP) occurring approximately 400 ms after word onset. The N400 has been used extensively in psycholinguistics research and its exact causes and interpretation are a matter of debate (e.g., Brown and Hagoort, 1993; Federmeier, 2007; Hagoort, 2008; Kutas and Federmeier, 2000; see Lau *et al.*, 2008, for a review). However, there is a broader sense in which the magnitude of the N400 can be used as a measure of lexical-semantic effort during word recognition (i.e., lexical access; Song and

[a]Electronic mail: jieun.song@ucl.ac.uk

Iverson, 2018). For example, the N400 is greater when listeners are recognizing low-frequency words with many lexical competitors than when recognizing high-frequency words with few competitors (e.g., Smith and Halgren, 1987; Winsler *et al.*, 2018); lexical selection is harder in the former case and requires greater neural resources. The N400 also varies with the predictability of words within a semantic context (i.e., greater for incongruent/low-predictability words in sentences than for high-predictability words), and this context-related variation can be used as a separate measure of semantic processing within sentences (Kutas and Hillyard, 1980; see Lau *et al.*, 2008, for a review). That is, greater variation in the N400 due to semantic predictability indicates that the listener is making more use of the sentence context, whereas smaller context-related variation can indicate that the listener is adopting more of a word-by-word recognition strategy. We examined both the magnitude of N400 and how it varied with semantic predictability. We also used newer measures of lexical processing that can be applied to continuous speech, similar to those used for auditory neural tracking, in order to assess lexical processing within our single-talker maskers (Broderick *et al.*, 2018; Broderick *et al.*, 2019).

Single-talker and babble maskers can place different demands on peripheral and central processing. Babble maskers are normally constructed to have relatively constant amplitude, whereas single talkers have natural amplitude fluctuation. Single talkers are thus less effective maskers at a peripheral level because they contain dips in amplitude where the target is relatively unmasked (e.g., Freyman *et al.*, 2004; Rosen *et al.*, 2013). However, linguistic content in single-talker maskers can cause additional interference. For example, intelligible maskers can produce lexical activation and processing that affect the recognition of the target speech (e.g., Brouwer and Bradlow, 2016; Cooke *et al.*, 2008). Single-talker maskers are likewise less disruptive when spoken in a language that the listener does not understand (Rhebergen *et al.*, 2005; Van Engen and Bradlow, 2007; Brouwer *et al.*, 2012) and require more effort to ignore when less degraded acoustically (Wöstmann *et al.*, 2017). Single-talker maskers can also be difficult in terms of auditory organization (i.e., stream segregation and selection) because they resemble the target more than do steady maskers like babble (e.g., Brungart, 2001). This effect of masker type on auditory organization appears to interact with the spatial separation of the target and masker. That is, masking effects are generally reduced when the target and distractor are at different spatial locations (e.g., see Blauert, 1983; Shinn-Cunningham, 2005, for a review), but this spatial release of masking is greater when the masker and target are otherwise hard to separate due to their similarity (i.e., two talkers) compared to when the masker is a stationary noise (Freyman *et al.*, 1999; Arbogast *et al.*, 2005). In short, this fairly straightforward contrast between single-talker maskers and multi-talker babble can differentially affect peripheral masking, perceptual organization, auditory spatial attention, and lexical processing.

One could predict, however, that these manipulations would have relatively little effect on neural tracking of attended speech because EEG measures of this process have been shown to be fairly robust to background noise or acoustic distortions except at very low signal-to-noise ratios (SNR; Fuglsang *et al.*, 2017; Ding *et al.*, 2014). That being said, linguistic interference from an intelligible masker has been claimed to reduce neural tracking of attended speakers (Dai *et al.*, 2018); this previous work involved vocoding and did not compare the effects of single-talker and babble maskers. Also, neural tracking is clearly enhanced by top-down attention (e.g., Ding and Simon, 2012; Kerlin *et al.*, 2010) and thus may be affected by increased listening effort in these difficult conditions. In terms of our lexical measure, N400 generally increases when word recognition becomes more difficult, but lexical-semantic processing can be disrupted (e.g., smaller N400 differences depending on context) when the speech signal is too strongly degraded by noise (e.g., Obleser and Kotz, 2011; Obleser *et al.*, 2007). It is not clear how the N400 will be affected by the additional lexical activation of an intelligible masker (e.g., Brouwer and Bradlow, 2016) or whether the overall greater cognitive demands of the single-talker condition will reduce the resources available for lexical-semantic processing, thereby reducing context-related N400 variation (Schmidt *et al.*, 2015; Otsuka and Kawaguchi, 2007).

L2 listeners have greater speech recognition difficulties in noise than do L1 listeners (e.g., Black and Hast, 1962; Cooke *et al.*, 2008; see Lecumberri *et al.*, 2010, for a review), and our expectation had been that L2 listeners would likewise have poorer neural tracking of a target talker presented with a single-talker masker. However, we have found that L2 listeners actually track target talkers more strongly than do L1 listeners (Song and Iverson, 2018). It seems likely that their increased neural tracking reflects active mechanisms that help compensate for their difficulties with L2 speech, although it is not clear exactly which perceptual or cognitive demands of the listening situation produces this increased neural tracking. Interestingly, older adults and hearing-impaired adults also appear to have greater neural tracking than normal-hearing younger adults (Presacco *et al.*, 2016; Brodbeck *et al.*, 2018; Decruy *et al.*, 2020). Our original motivation for this study was to examine how different listening conditions modulate this increased entrainment, in order to better understand why it arises. For example, Song and Iverson (2018) only had single-talker maskers, which were presented to a different ear from the target, and it is possible that more difficult listening conditions (e.g., colocated targets and distractor or greater acoustic masking produced by babble) might produce enhanced entrainment for L1 listeners that resemble what we have found for L2 listeners under easier conditions. Such results would suggest that the enhanced tracking for L2 listeners might be caused by perceptual factors (i.e., experience-related processing of acoustic phonetic variation).

For the most part, L2 speech recognition is marked by delayed N400 responses or reduced variation in the N400

https://doi.org/10.1121/10.0001477

related to semantic context (e.g., larger N400 for highly predictable words), both of which indicate poorer lexical access and semantic processing (Hahne, 2001; Song and Iverson, 2018; Stringer and Iverson, 2019a; cf, Hahne and Friederici, 2001). Moreover, the N400 effect related to context in quiet conditions is smaller for individual L2 listeners who have more difficulty with speech recognition in noise (Stringer and Iverson, 2019a). In contrast, L1 speakers have sometimes been found to increase their N400 magnitude for L2 accents, likely employing additional lexical resources to overcome difficulties at pre-lexical levels (Song and Iverson, 2018; Romero-Rivas *et al.*, 2015). But, more difficult accents can also suppress the N400 or context-related differences in the response (e.g., Goslin *et al.*, 2012; Stringer and Iverson, 2019a) in much the same way that the N400 can become suppressed with noise (e.g., Obleser and Kotz, 2011). Using neural tracking and N400 measures together can allow us to compare different ways L1 and L2 listeners cope with challenging situations by picking apart these different levels of speech processing (i.e., auditory, lexical).

The aim of the present study was to investigate within a single study how the demands of the listening conditions and the language experience of the listeners affect auditory and lexical-semantic processing. The previous work suggests that noise can reduce both auditory tracking and lexical processing (e.g., Ding *et al.*, 2014; Dai *et al.*, 2018; Obleser and Kotz, 2011), increased effort can enhance auditory and lexical processing (e.g., Song and Iverson, 2018), but there may be trade-offs between effort at the auditory and lexical levels given that listeners have a limited pool of cognitive resources (e.g., Kahneman, 1973). It is reasonable to expect that lexical-semantic processing will be disrupted to a greater extent with intelligible single-talker maskers than with babble maskers. However, it is unclear whether this disruption will produce decreased lexical processing or increased lexical processing because of additional effort, whether these lexical effects are linked to changes in auditory processing, and whether these effects vary with language experience.

We made EEG recordings while native English and Korean adults listened to English sentences in a background of intelligible single-talker and unintelligible babble maskers and in a condition without any noise. The masker was either colocated with the target at the front of the head or 45° to the side, simulated in insert earphones using head-related transfer functions (HRTFs; Algazi *et al.*, 2001). The two spatial positions were presented at different SNRs (+3 and −7 dB) to make the intelligibility of the target similar regardless of the spatial manipulation. We chose relatively high SNRs in the masker conditions that would only begin to affect intelligibility rather than severely disrupting speech processing. Subjects were asked to press a button whenever they heard a catch trial that did not make sense (semantically anomalous sentence), and the accuracy of the button response was used to assess their speech comprehension performance. The sentences varied in final-word predictability in order to assess effects of semantic context on the N400.

II. METHODS

A. Subjects

Twenty monolingual native speakers of British English and 21 native speakers of Korean participated in this study. One English subject was excluded from the analyses due to excessive noise in the EEG data (e.g., artefacts caused by movements). All English and Korean subjects were adults under 35 years old with the average age of 24.8 and 28.8 years old, respectively. They reported that they did not have any hearing, language, or other neurological impairments and were right-handed. The Korean subjects were second-language speakers of English; they were living in London at the time of testing, and their length of residence in English-speaking countries did not exceed 3 yr (average: 1.4 yr). They also reported that they had started learning English in school in South Korea at an average age of 12 years old and had not lived in English-speaking countries before the age of 18.

B. Stimuli and apparatus

English sentences were recorded by a female native speaker of Standard Southern British English. The sentences had different levels of final-word cloze probability that have been validated with non-native speakers to allow for measurement of the N400 response (Stringer and Iverson, 2019b). High cloze probability sentences consisted of highly constraining sentence contexts followed by congruent final words (e.g., *There are three pictures hanging on the wall.*). Low cloze probability sentences were neutral sentences (e.g., *There are many dirty marks on the wall*). Semantically anomalous sentences consisted of highly constraining sentence contexts followed by incongruent final words (e.g., *There are three pictures hanging on the pain*). A total of 725 sentences were used in the experiment; there were 300 high cloze and 300 low cloze probability sentences, comprising approximately 82.8% of the stimuli (41.4% each), and 125 anomalous sentences, comprising approximately 17.2% of the stimuli. Each sentence was presented only once with the average duration of 2.26 s.

The maskers were created from English stories that had been read by the same female speaker who recorded the sentences (*The Secret Garden*, Burnett, 1909; and *Lazy Jack*, Ross, 1985). For the single-talker masker, the stories were processed so that they had some of the continuous acoustic properties found in babble; pauses were edited so that they did not exceed 25 ms, and low-frequency amplitude modulations (<1 Hz) were attenuated by filtering the broadband Hilbert envelope of the signal. Because the acoustic similarity between competing voices (e.g., *f*0, spectral qualities) is an important factor affecting the amount of masking in a two-talker situation (e.g., Brungart, 2001), we used the same speaker for the single-

talker masker and target to eliminate that factor from our comparison of single-talker and babble maskers. The babble masker was also created based on the same recordings, which removed any speaker-related acoustic differences between the two maskers. In order to ensure that the babble was acoustically uniform and unintelligible, the stories were segmented into short sections of 1.5–2.5 s, excluding any that were silent for longer than 15% of the interval. These were then multiplied by a Hann window and spliced back together in a random order. Twelve of these random sequences were combined to create the babble.

For reference, Fig. 1 displays the modulation spectra for our target speech and two maskers. These modulation spectra are the summed within-channel modulations using a gammatone filter bank, calculated from the front-end of the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen *et al.*, 2013). The target speech had a typical speech modulation pattern with greatest modulation amplitudes in the 2–8 Hz range. The single-talker distractor had a similar modulation pattern but with reduced amplitude as a result of pause durations being reduced in the distractor; target sentences were separated by silent gaps. Adding multiple speech streams for the babble reduced the magnitude of modulations further; the amplitude modulations of speech streams have incoherent phase with each other when randomly summed and, thus the amplitude modulations tend to cancel out. Previous work comparing auditory tracking of target and distractor speech typically has used counterbalanced designs to match the target and distractor materials (e.g., Ding and Simon, 2012; Rimmele *et al.*, 2015; Song and Iverson, 2018). The present study did not match the target and maskers in this way. It therefore does not make sense to compare, for example, the auditory tracking of the target and the babble masker given

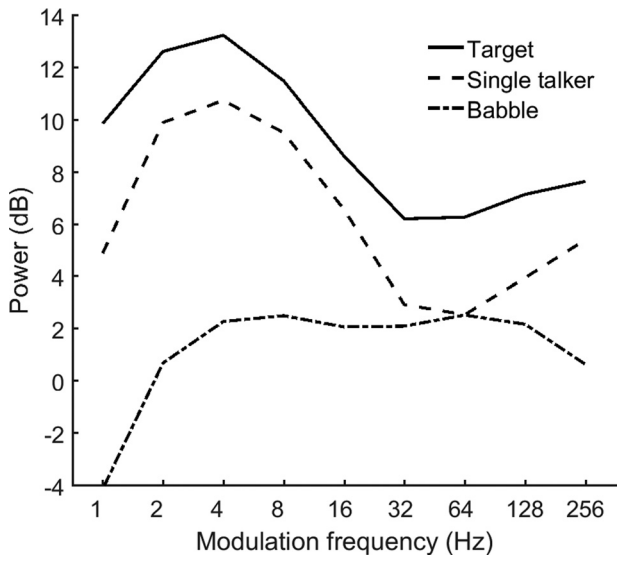

FIG. 1. Average modulation spectra for target sentences and single-talker and babble maskers. The spectra were calculated with the multi-resolution speech-based envelope power spectrum model (mr-sEPSM; Jørgensen *et al.*, 2013).

that babble has greatly reduced amplitude modulations, and these modulation differences would themselves affect the neural tracking results.

The targets and maskers were processed with HRTFs that reproduced the acoustic effects of presenting sound at different spatial locations (Algazi *et al.*, 2001). The target signal was always presented at 0° (front of the head), and the masker was either at the same location as the target or placed 45° to the left. In order to make intelligibility similar between the two conditions, the SNR was 3 dB for the colocated condition and −7 dB for the 45° condition; these SNR levels were determined based on results of our pilot work. All audio stimuli were recorded with a sampling rate of 44 100 Hz and a 16-bit quantisation. Stimuli were presented via Etymotic ER-1 insert earphones (Illinois) at 67 dB sound pressure level.

C. Procedure

Subjects were instructed to pay attention to the target sentences and ignore the single-talker or babble noise in the background. They were also asked to press a button whenever they heard a semantically anomalous sentence. The experiment consisted of a total of ten blocks (2 blocks * 5 conditions − 2 masker types * 2 spatial configurations, and no masker) with each lasting approximately 4 min. Sentences with all three levels of probability (i.e., high, low, and anomalous) were randomly mixed within each block, and the order of the blocks was randomised for each subject. Subjects were given a short break between blocks.

D. EEG recording and analysis

The EEG was recorded with a Biosemi Active Two System with 64 electrodes (Ag/AgCl; Amsterdam, Netherlands) mounted on an elastic cap and 7 external electrodes (nose, left and right mastoids, two vertical and horizontal EOG electrodes) with a sampling rate of 2048 Hz. Electrode impedances were kept between ±25 kΩ.

Preprocessing of the EEG recordings was performed offline in MATLAB using the Fieldtrip toolbox (Oostenveld *et al.*, 2011). They were re-referenced to the average of left and right mastoids and high-pass filtered at 0.1 Hz using a zero-phase Butterworth filter. A zero-phase 40-Hz low-pass Butterworth filter was also applied for the N400 analysis. Noisy channels were interpolated. Independent component analysis (ICA) was used to remove eye artefacts. The recordings were downsampled to 256 Hz for N400 analyses and 64 Hz for neural tracking analyses to improve computational efficiency.

1. Neural tracking analysis

Multivariate Temporal Response Functions (mTRFs; Crosse *et al.*, 2016) were generated in backward models that mapped the EEG data from each subject back onto the Hilbert envelopes of the sentences that they had heard. The mTRFs were trained over 0–400 ms time lags between the EEG and speech signals for each individual sentence. A

https://doi.org/10.1121/10.0001477

tenfold cross-validation procedure (David *et al.*, 2007) was then used to predict the speech signals (i.e., amplitude envelopes) using the models. That is, the sentences were randomly divided into ten groups, regardless of condition, and the amplitude envelopes of the target sentences of each group were predicted using an average model that was trained on the sentences that had been left out of the target group. Coherence was calculated to quantify the phase locking between the actual and predicted amplitude envelopes (i.e., accuracy of reconstruction based on EEG); the data were segmented into 1-s Hann windows with 50% overlap, and coherence was calculated from the cross-spectral density of the fast Fourier transform (FFT) of the two signals, divided by the power spectrum of each signal.

2. N400 analysis

The EEG data were segmented into 1000 ms epochs (200 ms pre-stimulus and 800 ms post-stimulus) that were time-locked to the onset of each final word of the target sentences. Trials were baseline-corrected by subtracting the pre-stimulus average. Trials were then rejected if the amplitude was not within the range of $\pm 150\,\mu$V. We performed a nonparametric permutation analysis (Maris and Oostenveld, 2007) to examine the scalp distribution and time scale of the response. This analysis avoids the problem of multiple comparisons by creating clusters in neighbouring time points and electrodes. These clusters were calculated based on averages for each subject for each relevant condition, averaging over conditions irrelevant to that test. The significance of a difference between two conditions (e.g., high cloze vs low cloze probability conditions) was obtained by the Monte Carlo method. That is, averages for each subject were randomly assigned to each condition regardless of which condition they originally belonged to, and 1000 of these random partitions were generated. The statistical significance was then determined by calculating the proportion of random partitions that had greater cluster-level statistics than real data (summed t-values in a cluster). Because it is difficult to use a cluster-based permutation analysis to investigate interaction effects involving more than two variables, N400 amplitude was quantified for mixed-effects analyses by averaging the amplitude in the 300–500-ms window across five midline electrodes (Fz, FCz, Cz, CPz, and Pz), following previous N400 studies (e.g., Strauß *et al.*, 2013; Song and Iverson, 2018).

3. Additional analyses of lexical processing

The story materials used in the masker were not designed for N400 analyses, but methodologies have been recently developed to extract a neural component related to lexical processing for continuous speech (Broderick *et al.*, 2018; Broderick *et al.*, 2019). In the present analysis, the semantic similarity of each stimulus word to its preceding context was calculated based on a computational analysis of English words. Specifically, a word2vec model (Mikolov *et al.*, 2013) was calculated based on three British English corpora to create a semantic similarity space of words (the British National Corpus, the ukWaC, and the English Wikipedia). Following Broderick *et al.* (2019), a semantic similarity index was calculated for each stimulus word by dividing one by the Euclidean distance in the word2vec space between the vector of each word and the average vector of all preceding words in that sentence. For the first word of each sentence in the masker, the Euclidean distance was calculated between the vector of that word and the average vector of all words in the previous sentence. A higher semantic similarity value indicated that the word was more semantically related to the context. Only content words were used in this calculation.

Similar to the neural tracking analysis, mTRFs (Crosse *et al.*, 2016) were calculated that mapped the neural signal back to the variation in semantic similarity of each word to its context. A semantic similarity function was generated for the stimulus materials that marked the onset of each content word with a 300-ms-width Gaussian window and with the amplitude of the Gaussian modulated by the degree of semantic similarity. The EEG and semantic similarity functions were downsampled to 64 Hz. The mTRFs were trained over 0–800 ms time lags on all other signals. An optimal ridge regression (lambda) value (100) was selected for training among a range of λ values (10^{-5}–10^{4}) with a "leave-one-out" cross-validation approach (i.e., leaving each stimulus block out of its training set). To quantify how strongly semantic similarity was represented in the EEG signals, Pearson correlations were calculated between the original and predicted signals (i.e., accuracy of reconstruction).

III. RESULTS

Behavioral responses were coded as the proportion of correct recognition of anomalous catch trials (Fig. 2); false alarms for non-anomalous sentences were low (mean proportion: English listeners, 0.03; Korean listeners, 0.09). The error rates were similar to our previous study (Song and Iverson, 2018); the task is more difficult than a typical sentence recognition task because it requires continuous vigilance for infrequent catch trials. A mixed-effects logistic regression analysis was performed using the button response for anomalous sentences (correct vs wrong) as the dependent variable with each sentence stimulus and subject as random intercepts and language (English vs Korean listeners) and condition as independent variables. Effect coding was used for language. The condition was effect coded using four contrasts that tested the overall effect of masking (no masker vs the other conditions), masker type (single-talker vs babble), spatial location of masker (0° vs 45° from target), and the interaction between masker type and spatial location. All mixed-effects models in this paper were fitted with the lme4 package (Bates *et al.*, 2014) in *R*, following the same procedure described above. The lmerTest package (Kuznetsova *et al.*, 2017) was used to obtain p-values.

As expected, accuracy was significantly higher in the no-masker condition than in the other masker conditions,

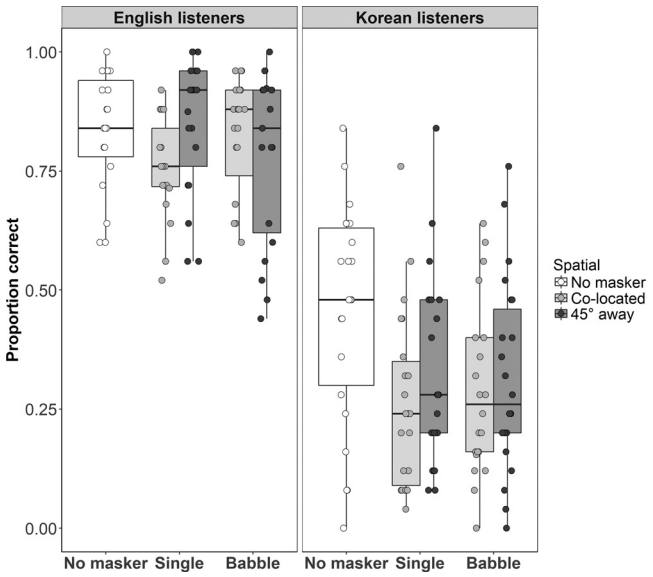

FIG. 2. Boxplots showing the proportions of correct detection of anomalous sentences by listener group (English listeners, Korean listeners), masker (no masker, single-talker masker, babble masker), and spatial position of the masker (colocated, 45° away). The mean scores of individual subjects for each condition are marked with dots.

$b = -0.573$, $z = -2.589$, $p = 0.010$, L1 English listeners were significantly more accurate than L2 Korean listeners were, $b = 1.383$, $z = 9.847$, $p < 0.001$, and this overall masking effect was greater for Korean than for English listeners, $b = 0.311$, $z = 3.268$, $p = 0.001$.

In this study, the SNR had been varied to produce similar behavioral accuracy for the two spatial manipulations, and the single-talker masker had been designed to produce similar performance to the babble (i.e., shortened pauses, reduced low-frequency amplitude modulation, and added difficulty because the same talker recorded both target and masker). There were likewise no main effects of masker type or spatial location, $p > 0.05$. However, there was a three-way interaction between language, masker type, and spatial location, $b = -0.343$, $z = -1.983$, $p = 0.047$. As shown in Fig. 2, there was a larger spatial release from masking in the single-talker condition (i.e., greater performance with a 45° separation) than for the babble in accord with previous results (Freyman *et al.*, 1999; Arbogast *et al.*, 2005) but this had less effect on the scores of L2 speakers. *Post hoc* mixed-effects logistic regression analyses were performed for each listener group, which found that the interaction between masker type and spatial location was only significant for English listeners, $b = -0.263$, $z = -2.109$, $p = 0.035$. The interaction of masker type and spatial location in the main model did not reach significance, $p = 0.093$.

Figure 3 displays model details of our neural tracking analysis (i.e., channel weights over time for the mTRF models along with mean coherence plots). Decoder weights should be interpreted with caution as they do not necessarily reflect the spatial or temporal origin of the neural response of interest (Haufe *et al.*, 2014). However, the mTRF weights resembled those of previous neural tracking work that used backward models (Fuglsang *et al.*, 2017; Song and Iverson, 2018) with negative weights (blue) at 100 ms and positive weights (yellow) at 150–200 ms in fronto-central electrodes possibly related to the N1 and P2 auditory evoked responses. Likewise, the greatest neural tracking of the speech envelope (i.e., peaks in coherence plots) occurred in a delta-theta range (2–8 Hz). Korean and English listeners had similar model weights and frequencies of peak coherence.

A linear mixed-effects model analysis was performed on the neural tracking results; coherence values averaged in the delta-theta range (2–8 Hz) were used as the dependent variable and each subject was used as a random intercept. There was a language * masker interaction, $b = 0.002$, $t(152) = 3.091$, $p = 0.002$, demonstrating that the difference in coherence between the no masker and masker conditions was significantly greater for Korean than for English listeners, and there was a significant main effect of masker, $b = -0.003$, $t(152) = -4.187$, $p < 0.001$. A Tukey *post hoc* test was performed using the Multcomp package (Hothorn *et al.*, 2019). The results demonstrated that there was no significant difference between masker and no-masker conditions for English listeners, $b = 0.001$, $z = 0.720$, $p = 0.875$, but this was significant for Korean listeners, $b = 0.005$, $z = 5.024$, $p < 0.001$. That is, the maskers had little overall effect on coherence for English listeners at the SNR levels used here, but they were strong enough to suppress the coherence of Korean listeners. Moreover, the Tukey *post hoc* test demonstrated that there was a significant difference in coherence between Korean and English listeners in the no-masker condition, $b = 0.005$, $z = 2.714$, $p = 0.029$, but not in the masker conditions, $b = 0.001$, $z = 0.870$, $p = 0.800$. We thus replicated the enhanced entrainment for L2 speakers that we had found previously (Song and Iverson, 2018) but only in the condition that had no masking. This likely occurred because the masking conditions used in the current study (e.g., masker and target produced by the same talker) produced more suppression of the neural tracking by Korean listeners than in our previous study (two talkers with different accents).

Neural tracking was also significantly greater with babble than with single-talker maskers, $b = -0.004$, $t(152) = -3.702$, $p < 0.001$. On its own, it is difficult to know whether the greater peripheral masking in the babble condition increased listening effort and thereby neural entrainment or the increased effort for lexical processing in the single-talker condition suppressed neural tracking. There is some evidence in the literature that the latter is plausible (Dai *et al.*, 2018); vocoded single-talker maskers have been found to suppress neural tracking to a greater extent when they are comprehensible, more than when they are heard by listeners who do not perceive these vocoded signals as speech.

Neural tracking was not different overall depending on whether the target was colocated with the noise or spatially separated, $p = 0.346$, but the interaction between spatial separation and masker type (single vs babble) was significant,

https://doi.org/10.1121/10.0001477

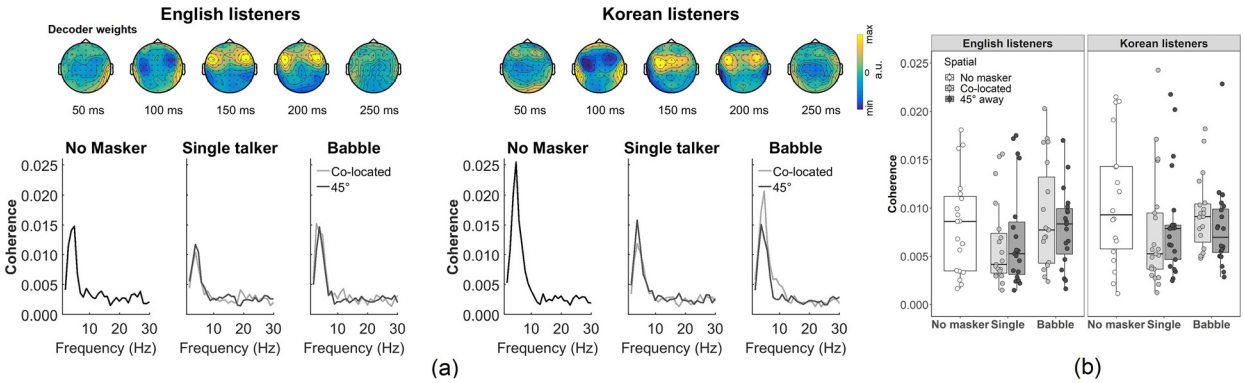

FIG. 3. (Color online) (a) Plots showing coherence values as a function of frequency by listener group (English listeners, Korean listeners), masker (no masker, single-talker masker, babble masker), and spatial position of the masker (colocated, 45° away). Topographic maps display averaged decoder weights for each listener group across different time lags with the scale plotted in arbitrary units (a.u.). (b) Boxplots showing average coherence values in the delta-theta range by listener group, masker, and spatial position. Dots on the boxplots represent average coherence values of each individual subject for each condition.

$b = -0.003$, $t(152) = -2.516$, $p = 0.013$, if small in magnitude (see Fig. 3). The three-way interaction with language (spatial separation * masker type * language) was not significant, $p = 0.459$. A *post hoc* Tukey test demonstrated that there was no significant effect of spatial separation for single-talker maskers, $b = 0.001$, $z = 1.374$, $p = 0.516$, but coherence was significantly reduced for babble maskers when there was 45° spatial separation, $b = -0.002$, $z = -3.083$, $p = 0.011$. Our spatially separated conditions were presented with a lower SNR (-7 dB) than when colocated (+3 dB), and it is possible that this entrainment difference was a direct result of increased energetic masking.

Figure 4 displays lexical processing results (N400; i.e., ERP waveforms time-locked to the final word of each target sentence). The N400 was greater (i.e., more negative) for words in low cloze probability sentences than for words in high cloze probability sentences, showing a typical context-related N400 effect. A linear mixed-effects analysis was performed with the average N400 amplitudes in the 300–500-ms time window as the dependent variable, sentence type (high cloze and low cloze probability), language, and condition as independent variables, and by-subject random intercepts. Effect coding was used for sentence type and language. There was a main effect of sentence type, $b = 0.800$, $t(342) = 10.635$, $p < 0.001$, confirming that N400 was modulated by semantic context. In addition, the interaction of sentence type and language was significant, $b = 0.408$, $t(342) = 5.423$, $p < 0.001$. That is, context-related N400 differences were greater for L1 than for L2 listeners, similar to previous findings (Hahne, 2001; Stringer and

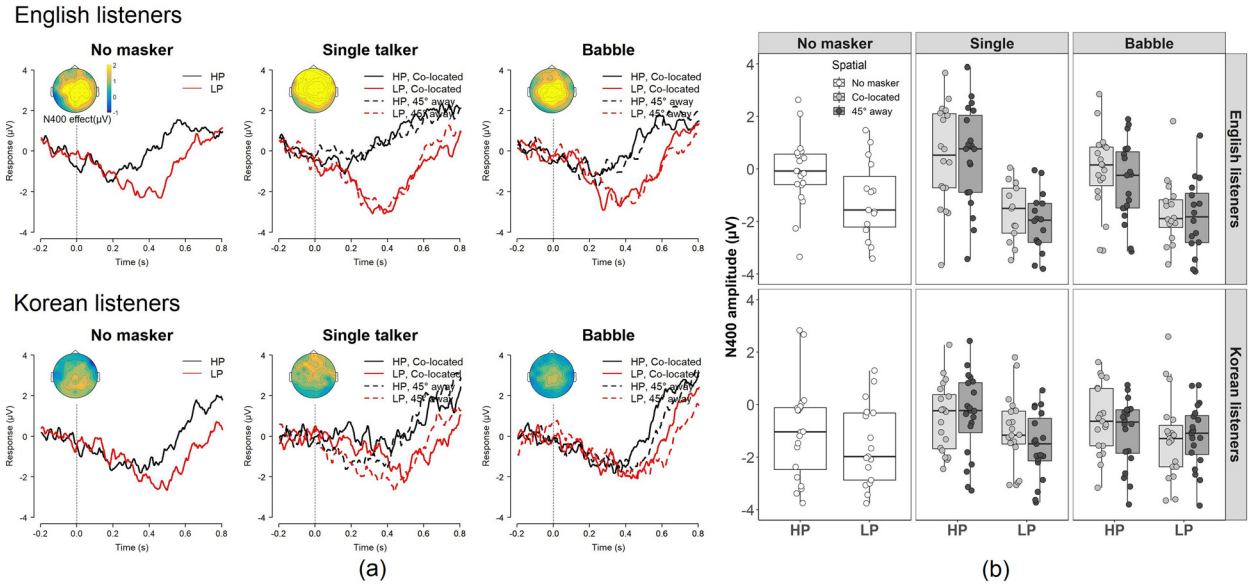

FIG. 4. (Color online) Results of the N400 analysis. (a) Grand average ERP waveforms for sentence-final words, divided by sentence type (HP, high cloze probability sentences; LP, low cloze probability sentences), masker (no masker, single-talker masker, babble masker), and spatial position of the masker (colocated, 45° away). Topographic plots show the mean N400 differences between HP and LP on the scalp for each masker type. (b) Boxplots showing averaged N400 amplitudes in the 300–500-ms time window for each listener group, sentence type, masker, and spatial location.

Iverson, 2019a; Song and Iverson, 2018), although the main effect of language was not significant, $p = 0.541$. L2 listeners thus had more difficulty in using the semantic structure of sentences during lexical processing.

Furthermore, the difference in N400 between high and low probability conditions was significantly larger in the single-talker masker condition than in the babble condition, $b = 0.985$, $t(342) = 2.928$, $p = 0.004$. This suggests that listeners increased reliance on the semantic context in sentences to cope with the single-talker masker. It appears that listeners paid greater attention to the semantic content of the entire sentence in the single-talker condition because having expectations about upcoming words may help the listener attend to the correct speech stream. This effect was not significantly different for English and Korean listeners, $p = 0.517$ (i.e., no significant three-way interaction of sentence type, masker type, and language). In addition, the main effect of masker (i.e., no masker vs masker and single-talker vs babble) was not significant, there was no significant main effect or interaction involving the spatial separation, and the contrast between high and low probability sentences was not significantly different between the no-masker and the masker conditions, $p > 0.05$.

In order to further examine the distribution of the context-related N400 variation and its latency, a nonparametric cluster-based permutation analysis (Maris and Oostenveld, 2007) was performed for each listener group in the 200–600-ms time window. Both listener groups had a significant difference in N400 amplitude between high and low cloze probability conditions. As shown in Fig. 5, the difference was found in a large significant cluster across the scalp between 200 and 600 ms for native listeners, $p < 0.001$. In contrast, a significant cluster was found between 400 and 600 ms for Korean listeners, $p < 0.001$, suggesting relatively delayed context-related N400 variation. Consistent with previous studies (Hahne and Friederici, 2001; Hahne, 2001), this result indicates that lexical processing for L2 speech can be slower than for L1. Moreover, the distribution of their response changed slightly over time from central scalp locations (400–450 ms) to wider, frontocentral locations (500–600 ms). It is difficult to know why this change occurred at the later time range within this N400 study, but it might reflect some late processes related to semantic processing, which have been found in non-native listeners in the right anterior-central electrode sites (Hahne and Friederici, 2001).

Another cluster-based permutation analysis was performed to further investigate the effect of masker type (single-talker vs babble maskers) on context-related N400 variation. This was conducted across the two listener groups because there was no significant interaction of language with sentence type and masker type in the mixed-effects analysis. The results (Fig. 5) demonstrated that the N400 difference between high- and low-predictability sentences significantly differed between single-talker and babble noise conditions in the 200–500-ms time range but in more frontal electrodes ($p < 0.001$). This was because the greater context-related N400 differences in the single-talker condition had a broader distribution across the scalp, whereas the N400 context variation in the babble condition had a more classic centro-parietal distribution, as displayed in the topographic plots in Fig. 4. A more frontal scalp distribution of

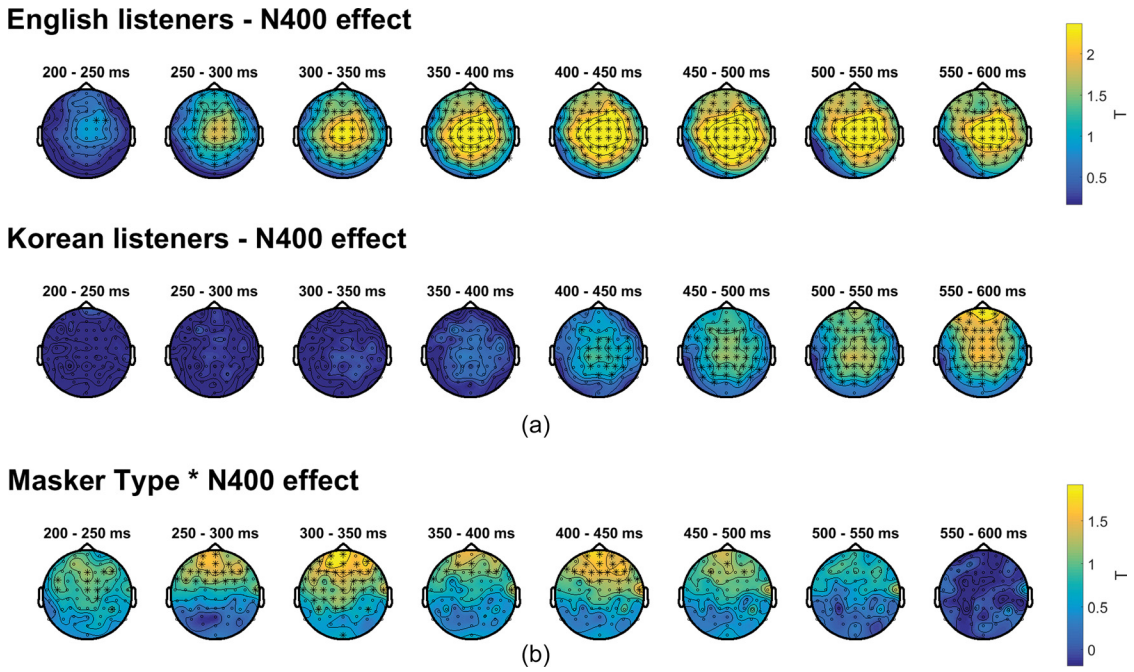

FIG. 5. (Color online) (a) Topographic maps showing *t*-values of each sample pair (time-electrode) in the nonparametric permutation analysis, which compared high vs low cloze probability conditions (i.e., N400 effect), for English and Korean listeners. The significant cluster is marked with "*." (b) Topographic maps showing *t*-values of each sample pair in the nonparametric permutation analysis, which compared HP-LP differences in single-talker vs babble conditions (i.e., interaction of masker * sentence type) for all listeners.

https://doi.org/10.1121/10.0001477

the N400 response may reflect the involvement of attention-related areas (Courchesne, 1990); a similar distribution has been previously found in response to foreign-accented speech in other N400 studies (e.g., Romero-Rivas *et al.*, 2015).

A further lexical processing analysis was conducted to evaluate whether lexical processing related to the single-talker distractors may have affected processing of the target. The single-talker distractors were read stories without the carefully controlled lexical probabilities of the target sentences. We examined the activation for words in the masker, using the computational model detailed in the method, which measured semantic similarity between words and used a mTRF analysis to extract neural activity related to semantic similarity variation (Broderick *et al.*, 2018; Broderick *et al.*, 2019). The results of this analysis were evaluated by calculating the correlation between the predicted and obtained semantic similarity functions. For the single-talker maskers, the correlation was close to zero with the average of $r = -0.004$ [standard deviation (sd), 0.016]; a one-sample t-test confirmed that the correlation values were not significantly different from zero, $t(39) = -1.654$, $p = 0.106$. That is, the analysis did not find that the semantic content of the distractors was represented in the EEG signals. For reference, the same analysis was performed on words in the attended sentences in the single-talker masker condition, even though this type of analysis is more suitable for continuous speech than for a series of unrelated sentences. The average correlation between the original and predicted semantic similarity signals was found to be much larger (mean, $r = 0.064$; sd, 0.052) as shown in the boxplot in Fig. 6. The correlation values were significantly greater than zero, $t(39) = 8.420$, $p < 0.001$, in a one-way t-test. The weights of the decoder used for this analysis are shown in Fig. 6, and they suggest that centro-parietal electrodes contributed most to the signal reconstruction at around a 300-ms time lag, which is similar to the distribution and latency of the N400 response (cf. see Broderick *et al.*, 2018, for weights of forward models). This contrasts with the weights of the distractor decoder (Fig. 6), which did not find any significant neural component related to semantic similarity. The results therefore verify that this newer lexical processing analysis is able to measure lexical activity that we knew must have occurred for the targets given the behavioral scores and N400 results, but it was unable to find similar activation for the single-target masker.

IV. DISCUSSION

The results illustrate the kinds of perceptual and cognitive adjustments that listeners make to cope with speech recognition under challenging conditions. That is, speech is processed very differently when there is a single-talker masker than when there is an unintelligible babble masker, particularly at a lexical level. The N400 was more modulated by semantic context with a single-talker masker (i.e., smaller for high-predictability sentences) than with a babble masker, suggesting that listeners paid greater attention to the meaning of the entire sentence in that condition. Moreover, N400 had a broader distribution across the scalp with a single-talker masker rather than a typical centro-parietal distribution, suggesting that more attentional resources were engaged at a lexical level. This corresponded with poorer auditory neural tracking as has been found previously for

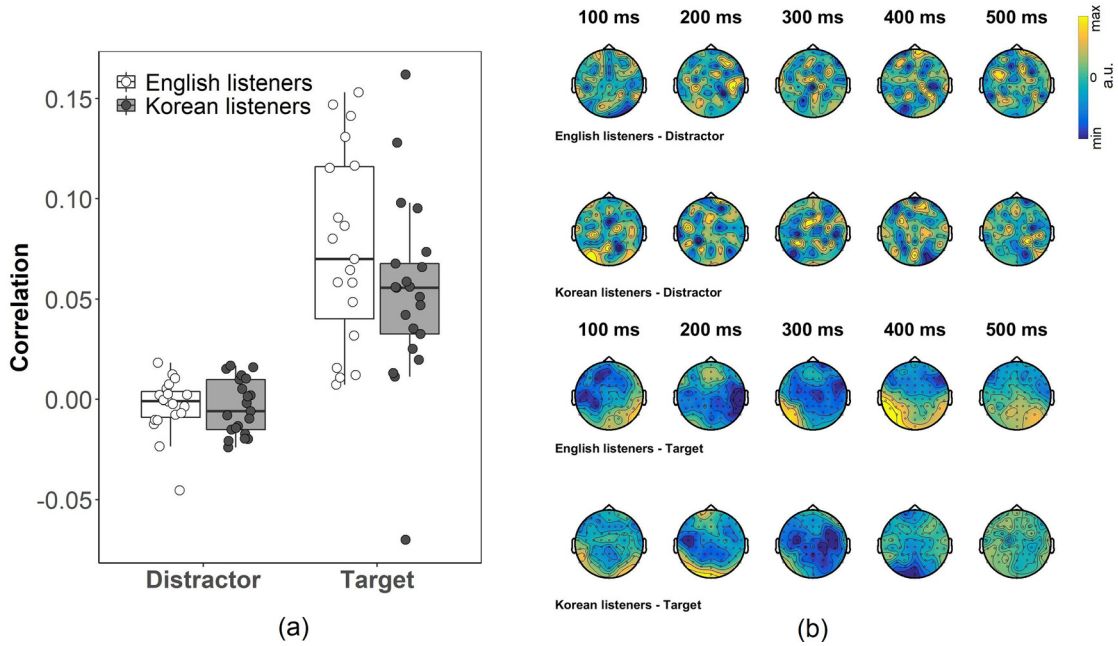

FIG. 6. (Color online) (a) Boxplot showing correlation coefficients between reconstructed and original semantic similarity metrics, divided by talker type (distractor, single-talker maskers; target, N400 sentences) and listener group. (b) Topographic maps showing weights of the decoder mapping EEG signals and semantic similarity metrics for words in the single-talker maskers and for target sentences. The maps are displayed over 100–500 ms time lags, separately for each listener group.

intelligible noise-vocoded maskers (Dai *et al.*, 2018). Some of these results have ambiguous interpretations on their own (e.g., whether decreased neural tracking is caused by greater perceptual difficulty or reduced effort), but they fit together to suggest that single-talker maskers make it more challenging at a lexical-semantic level to understand the target speech in a way that reduces the resources available to track speech acoustics. Furthermore, these processing differences can be found at relatively high SNR levels (i.e., small reductions in behavioral accuracy for L1 listeners) rather than only happening near threshold levels of understanding.

Previous work had suggested that intelligible maskers may be difficult because the maskers produce lexical activation even when they are meant to be ignored. Specifically, Brouwer and Bradlow (2016) found that when two simultaneous words are presented, the unattended word can affect eye movements in a visual-world paradigm (i.e., looking longer at items that match the distractor), and Aydelott *et al.* (2015) found that unattended words can produce semantic priming in lexical decision tasks. Much older dichotic listening experiments found that attention can shift when the listener's name is presented in the unattended ear (Moray, 1959). It thus seems likely that some monitoring of unattended speech takes place, at least some monitoring for salient word forms, but our present results suggest that unattended continuous speech does not produce the kinds of lexical search and semantic integration that are detected by our neural measures of lexical processing. That is, we found no neural component that tracked semantic similarity in the distractor speech. Moreover, if words in the masker had increased lexical competition for target words, the overall N400 amplitudes for target words would have increased regardless of context; we found no such main effect of masker type on N400 magnitude but found instead an interaction of masker type with lexical-semantic predictability.

It is more plausible, based on our results, that listeners followed the meaning of the target sentences as a streaming mechanism. That is, expectations listeners built up about upcoming words based on contextual information may have helped them to bind words together into a single stream and selectively attend to the target. This fits previous behavioural work (Kidd *et al.*, 2014); listeners are better at understanding words in a correct syntactic structure than words in a random order, but this difference is larger for single-talker maskers than for noise bursts. The general concept of streaming by meaning is an old one; dichotic listening work found that listeners automatically follow the meaning of target speech when it is switched between ears (Gray and Wedderburn, 1960; Treisman, 1960). The present results demonstrate how this "streaming by meaning" process affects lexical processing at the level measured by N400 and show that this mechanism becomes influential with single-talker maskers.

These modifications in lexical processing may reduce the resources available for neural tracking of the speech envelope. Mattys *et al.* (2009) and Mattys *et al.* (2014) have previously demonstrated trade-offs in lexical and auditory processing depending on signal clarity and cognitive load; degrading speech with unintelligible noise causes listeners to attend to the signal and be less influenced by lexical structure, whereas cognitive load caused by a simultaneous visual task decreases attention to acoustic detail and increases reliance on lexical structure. To some extent, our results are similar in that our condition that required additional lexical-semantic processing (single-talker masker) reduced auditory neural tracking, and our unintelligible babble masker caused listeners to rely less on semantic structure. It is likely that neural tracking of the target in the single-talker condition was suppressed because listeners had to recruit additional attentional resources at lexical levels, which may have depleted the resources needed for tracking the target speech (e.g., Molloy *et al.*, 2015).

Our original aim for this study was to examine differing listening conditions to explain why L2 speakers have greater neural tracking for target speech in two-talker conditions than do L1 listeners (Song and Iverson, 2018). We did not find a listening condition that made L1 listeners behave more like L2 listeners, but we replicated our original finding that L2 listeners can have higher entrainment under some conditions. Other studies have recently found enhanced speech tracking in older listeners and hearing-impaired listeners (Presacco *et al.*, 2016; Brodbeck *et al.*, 2018; Decruy *et al.*, 2020). To some extent, this might be a misleading similarity; hearing-impaired listeners have peripheral impairments that our L2 speakers did not have, and larger auditory responses in older listeners can also be related to other problems (e.g., changes in inhibitory neural mechanisms or temporal processing in the midbrain; Alain *et al.*, 2014; Presacco *et al.*, 2016). That being said, it is plausible that some deficit in the uptake of phonetic information—whether caused by hearing impairment or phonetic processing that is not well tuned to the language—or additional cognitive load—whether caused by cognitive decline or lexical processing that is less developed for the L2—produces similar changes in neural entrainment across all three populations.

Overall, we replicated previous findings that L2 speakers are more adversely affected by noise than are L1 listeners (e.g., see Lecumberri *et al.*, 2010, for a review), and L2 listeners tend to have N400 responses that are delayed and less affected by semantic context (Song and Iverson, 2018; Hahne, 2001; Stringer and Iverson, 2019a). These N400 findings may have been caused because L2 lexical processing activates additional lexical candidates through phonetic misperceptions, as well as activating words from their L1 (e.g., Weber and Cutler, 2004; Sebastián-Gallés *et al.*, 2005). The smaller context-related N400 difference also suggests that their semantic processing skills were not fully developed. Despite these speech processing difficulties, L2 listeners appeared to modulate lexical processing in much the same way in response to the single-talker condition (i.e., enhanced context-related N400 effect). It appears that the streaming by meaning mechanism is a general strategy that can also be adopted by L2 listeners to aid speech

segregation even if they do not benefit from contextual cues as well as L1 listeners.

The effects of our spatial manipulations were relatively minor. We replicated the behavioral finding in our L1 listeners that there is greater spatial release of masking for single-talker maskers than for more continuous noises (e.g., Freyman *et al.*, 1999). However, the effect of spatial location of the masker had only a minor effect on auditory neural tracking that we could not fully explain and had no statistically significant effect on the N400. It might be that the auditory organization processes that are enhanced by increased spatial separation are not strongly represented in these measures.

One of the ironies of this study, using modern computational methods and EEG, is that some of our findings fit with what was found in 1960 using tape recorders (e.g., using meaning to track an utterance; Gray and Wedderburn, 1960; Treisman, 1960). That being said, the present investigation more directly uncovers the mechanisms involved, using analyses that allow us to assess both auditory and lexical processing within a relatively naturalistic connected speech task. Speech recognition involves modulating processing and effort at both levels, and further investigation is required to better understand how this depends on the listening condition and the language background of the listener.

ACKNOWLEDGMENTS

This study was supported by the Economic and Social Research Council of the United Kingdom.

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., and Merzenich, M. M. (**2001**). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," Proc. Natl. Acad. Sci. U.S.A. **98**, 13367–13372.

Alain, C., Roye, A., and Salloum, C. (**2014**). "Effects of age-related hearing loss and background noise on neuromagnetic activity from auditory cortex," Front. Syst. Neurosci. **8**, 1–12.

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (**2001**). "The CIPIC HRTF database," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 99–102.

Arbogast, T. L., Mason, C. R., and Kidd, G. (**2005**). "The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **117**, 2169–2180.

Aydelott, J., Jamaluddin, Z., and Nixon Pearce, S. (**2015**). "Semantic processing of unattended speech in dichotic listening," J. Acoust. Soc. Am. **138**, 964–975.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2014**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Black, J. W., and Hast, M. H. (**1962**). "Speech reception with altering signal," J. Speech Hear. Res. **5**, 70–75.

Blauert J. (**1983**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).

Brodbeck, C., Presacco, A., Anderson, S., and Simon, J. Z. (**2018**). "Over-representation of speech in older adults originates from early response in higher order auditory cortex," Acta Acust. Acust. **104**, 774–777.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (**2018**). "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," Curr. Biol. **28**, 803–809.

Broderick, M. P., Anderson, A. J., and Lalor, E. C. (**2019**). "Semantic context enhances the early auditory encoding of natural speech," J. Neurosci. **39**, 7564–7575.

Brouwer, S., and Bradlow, A. R. (**2016**). "The temporal dynamics of spoken word recognition in adverse listening conditions," J. Psycholinguist. Res. **45**, 1151–1160.

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

Brown, C., and Hagoort, P. (**1993**). "The processing nature of the N400: Evidence from masked priming," J. Cogn. Neurosci. **5**, 34–44.

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

Burnett, F. H. (**1909**). *The Secret Garden* (Heineman, London, England), available at http://etc.usf.edu/lit2go/163/the-secret-garden/ (Last viewed October 6, 2017).

Cooke, M., Garcia Lecumberri, M. L., and Barker, J. (**2008**). "The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception," J. Acoust. Soc. Am. **123**, 414–427.

Courchesne, E. (**1990**). "Chronology of postnatal human brain development: Event-related potential, positron emission tomography, myelinogenesis, and synaptogenesis studies," in *Event-Related Brain Potentials: Basic Issues and Applications*, edited by J. W. Rohrbaugh, R. Parasuraman, and R. Johnson, Jr. (Oxford University Press, New York), pp. 210–241.

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (**2016**). "The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," Front. Hum. Neurosci. **10**, 1–14.

Dai, B., McQueen, J. M., Terporten, R., Hagoort, P., and Kösem, A. (**2018**). "Distracting linguistic information impairs neural entrainment to attended speech," bioRxiv, https://doi.org/10.1101/364042.

David, S. V., Mesgarani, N., and Shamma, S. A. (**2007**). "Estimating sparse spectro-temporal receptive fields with natural stimuli," Netw. Comput. Neural Syst. **18**, 191–212.

Decruy, L., Vanthornhout, J., and Francart, T. (**2020**). "Hearing impairment is associated with enhanced neural tracking of the speech envelope," Hear. Res. **393**, 1–13.

Ding, N., Chatterjee, M., and Simon, J. Z. (**2014**). "Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure," Neuroimage **88**, 41–46.

Ding, N., and Simon, J. Z. (**2012**). "Emergence of neural encoding of auditory objects while listening to competing speakers," Proc. Natl. Acad. Sci. U.S.A. **109**, 11854–11859.

Ding, N., and Simon, J. Z. (**2014**). "Cortical entrainment to continuous speech: Functional roles and interpretations," Front. Hum. Neurosci. **8**, 1–7.

Federmeier, K. D. (**2007**). "Thinking ahead: The role and roots of prediction in language comprehension," Psychophysiology **44**, 491–505.

Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (**2004**). "Effect of number of masking talkers and auditory priming on informational masking in speech recognition," J. Acoust. Soc. Am. **115**, 2246–2256.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Fuglsang, S. A., Dau, T., and Hjortkjær, J. (**2017**). "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," Neuroimage **156**, 435–444.

Goslin, J., Duffy, H., and Floccia, C. (**2012**). "An ERP investigation of regional and foreign accent processing," Brain Lang. **122**, 92–102.

Gray, J. A., and Wedderburn, A. A. I. (**1960**). "Grouping strategies with simultaneous stimuli," Q. J. Exp. Psychol. **12**, 180–184.

Hagoort, P. (**2008**). "The fractionation of spoken language understanding by measuring electrical and magnetic brain signals," Philos. Trans. R. Soc. Lond. B. Biol. Sci. **363**, 1055–1069.

Hahne, A. (**2001**). "What's different in second-language processing? Evidence from event-related brain potentials," J. Psycholinguist. Res. **30**, 251–266.

Hahne, A., and Friederici, A. D. (**2001**). "Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials," Biling. Lang. Cogn. **4**, 123–141.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., and Bießmann, F. (**2014**). "On the interpretation of weight vectors of linear models in multivariate neuroimaging," Neuroimage **87**, 96–110.
Hothorn, T., Bretz, F., Westfall, P., Heiberger, R. M., Schuetzenmeister, A., and Scheibe, S. (**2019**). "Multcomp package (Simultaneous inference in general parametric models)," pp. 1–36, available at http://multcomp.r-forge.r-project.org (Last viewed November 3, 2019).
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.
Kahneman, D. (**1973**). in *Attention and effort* (Prentice-Hall, Englewood Cliffs, NJ).
Kerlin, J. R., Shahin, A. J., and Miller, L. M. (**2010**). "Attentional gain control of ongoing cortical speech representations in a 'cocktail party,'" J. Neurosci. **30**, 620–628.
Kidd, G., Mason, C. R., and Best, V. (**2014**). "The role of syntax in maintaining the integrity of streams of speech," J. Acoust. Soc. Am. **135**, 766–777.
Kutas, M., and Federmeier, K. D. (**2000**). "Electrophysiology reveals semantic memory use in language comprehension," Trends Cogn. Sci. **12**, 463–470.
Kutas, M., and Hillyard, S. A. (**1980**). "Reading senseless sentences: Brain potentials reflect semantic incongruity," Science **207**, 203–205.
Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2017**). "lmerTest Package: Tests in linear mixed effects models," J. Stat. Softw. **82**, 1–26.
Lau, E. F., Phillips, C., and Poeppel, D. (**2008**). "A cortical network for semantics: (De)constructing the N400," Nat. Rev. Neurosci. **9**, 920–933.
Lecumberri, M. L. G., Cooke, M., and Cutler, A. (**2010**). "Non-native speech perception in adverse conditions: A review," Speech Commun. **52**, 864–886.
Luo, H., and Poeppel, D. (**2007**). "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," Neuron **54**, 1001–1010.
Maris, E., and Oostenveld, R. (**2007**). "Nonparametric statistical testing of EEG- and MEG-data," J. Neurosci. Methods **164**, 177–190.
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.
McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., and Wingfield, A. (**2005**). "Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech," Q. J. Exp. Psychol. Sect. A Hum. Exp. Psychol. **58**, 22–33.
Mikolov, T., Chen, K., Corrado, G., and Dean, J. (**2013**). "Efficient estimation of word representations in vector space," available at http://arxiv.org/abs/1301.3781 (Last viewed February 8, 2019).
Molloy, K., Griffiths, T. D., Chait, M., and Lavie, N. (**2015**). "Inattentional deafness: Visual load leads to time-specific suppression of auditory evoked responses," J. Neurosci. **35**, 16046–16054.
Moray, N. (**1959**). "Attention in dichotic listening: Affective cues and the influence of instructions," Q. J. Exp. Psychol. **11**, 56–60.
Obleser, J., and Kotz, S. A. (**2011**). "Multiple brain signatures of integration in the comprehension of degraded speech," Neuroimage **55**, 713–723.
Obleser, J., Wise, R. J. S., Dresner, M. A., and Scott, S. K. (**2007**). "Functional integration across brain regions improves speech perception under adverse listening conditions," J. Neurosci. **27**, 2283–2289.
Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (**2011**). "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," Comput. Intell. Neurosci. **2011**, 156869.
Otsuka, S., and Kawaguchi, J. (**2007**). "Divided attention modulates semantic activation: Evidence from a nonletter-level prime task," Mem. Cogn. **35**, 2001–2011.
Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (**1995**). "How young and old adults listen to and remember speech in noise," J. Acoust. Soc. Am. **97**, 593–608.
Presacco, A., Simon, J. Z., and Anderson, S. (**2016**). "Evidence of degraded representation of speech in noise, in the aging midbrain and cortex," J. Neurophysiol. **116**, 2346–2355.
Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (**2005**). "Release from informational masking by time reversal of native and non-native interfering speech," J. Acoust. Soc. Am. **118**, 1274–1277.
Rimmele, J. M., Zion Golumbic, E., Schröger, E., and Poeppel, D. (**2015**). "The effects of selective attention and speech acoustics on neural speech-tracking in a multi-talker scene," Cortex **68**, 144–154.
Romero-Rivas, C., Martin, C. D., and Costa, A. (**2015**). "Processing changes when listening to foreign-accented speech," Front. Hum. Neurosci. **9**, 167.
Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (**2013**). "Listening to speech in a background of other talkers: Effects of talker number and noise vocoding," J. Acoust. Soc. Am. **133**, 2431–2443..
Ross, T. (**1985**). *Lazy Jack* (Andersen, London), available at http://www.storynory.com/ (Last viewed October 6, 2017).
Schmidt, J., Scharenborg, O., and Janse, E. (**2015**). "Semantic processing of spoken words under cognitive load in older listeners," in *Proc. 18th Int. Congr. Phonetic Sci. (ICPhS 2015)*.
Sebastián-Gallés, N., Echeverría, S., and Bosch, L. (**2005**). "The influence of initial exposure on lexical representation: Comparing early and simultaneous bilinguals," J. Mem. Lang. **52**, 240–255.
Shinn-Cunningham, B. G. (**2005**). "Influences of spatial cues on grouping and understanding sound," in Proc. Forum Acusticum, Vol. 355, pp. 1539–1544.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Smith, M. E., and Halgren, E. (**1987**). "Event-related potentials during lexical decision: Effects of repetition, word frequency, pronounce-ability, and concreteness," Electroencephalogr. Clin. Neurophysiol. Suppl. **40**, 417–421.
Song, J., and Iverson, P. (**2018**). "Listening effort during speech perception enhances auditory and lexical processing for non-native listeners and accents," Cognition **179**, 163–170.
Strauß, A., Kotz, S. A., and Obleser, J. (**2013**). "Narrowed expectancies under degraded speech: Revisiting the N400," J. Cogn. Neurosci. **25**(8), 1383–1395.
Stringer, L., and Iverson, P. (**2019a**). "Accent intelligibility differences in noise across native and nonnative accents: Effects of talker–listener pairing at acoustic–phonetic and lexical levels," J. Speech, Lang. Hear. Res. **62**, 2213–2226.
Stringer, L., and Iverson, P. (**2019b**). "Non-native speech recognition sentences: A new materials set for non-native speech perception research," Behav. Res. Methods **52**, 561–571.
Treisman, A. M. (**1960**). "Contextual cues in selective listening," Q. J. Exp. Psychol. **12**, 242–248.
Van Engen, K. J., and Bradlow, A. R. (**2007**). "Sentence recognition in native- and foreign-language multi-talker background noise," J. Acoust. Soc. Am. **121**, 519–526.
Weber, A., and Cutler, A. (**2004**). "Lexical competition in non-native spoken-word recognition," J. Mem. Lang. **50**, 1–25.
Winsler, K., Midgley, K. J., Grainger, J., and Holcomb, P. J. (**2018**). "An electrophysiological megastudy of spoken word recognition," Lang. Cogn. Neurosci. **33**, 1063–1082.
Wöstmann, M., Lim, S. J., and Obleser, J. (**2017**). "The human neural alpha response to speech is a proxy of attentional control," Cereb. Cortex **27**, 3307–3317.