HHS Public Access

Author manuscript

Neuroimage. Author manuscript; available in PMC 2021 July 20.

Published in final edited form as:

Neuroimage. 2021 March ; 228: 117699. doi:10.1016/j.neuroimage.2020.117699.

Pre- and post-target cortical processes predict speech-in-noise performance

Subong Kim[#a], **Adam T. Schwalje**[#b], **Andrew S. Liu**[b], **Phillip E. Gander**[c], **Bob McMurray**[b,d,e], **Timothy D. Griffiths**[f], **Inyong Choi**[b,d,*]

[a]Department of Speech, Language, and Hearing Sciences, Purdue University, West Lafayette, IN 47907, USA

[b]Department of Otolaryngology – Head and Neck Surgery, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA

[c]Department of Neurosurgery, University of Iowa Hospitals and Clinics, Iowa City, IA 52242, USA

[d]Department of Communication Sciences and Disorders, University of Iowa, Iowa City, IA 52242, USA

[e]Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, USA

[f]Biosciences Institute, Newcastle University, Newcastle upon Tyne NE1 7RU, UK

[#] These authors contributed equally to this work.

Abstract

Understanding speech in noise (SiN) is a complex task that recruits multiple cortical subsystems. There is a variance in individuals' ability to understand SiN that cannot be explained by simple hearing profiles, which suggests that central factors may underlie the variance in SiN ability. Here, we elucidated a few cortical functions involved during a SiN task and their contributions to individual variance using both within- and across-subject approaches. Through our within-subject analysis of source-localized electroencephalography, we investigated how acoustic signal-to-noise ratio (SNR) alters cortical evoked responses to a target word across the speech recognition areas, finding stronger responses in left supramarginal gyrus (SMG, BA40 the *dorsal lexicon* area) with quieter noise. Through an individual differences approach, we found that listeners show different neural sensitivity to the background noise and target speech, reflected in the amplitude ratio of earlier auditory-cortical responses to speech and noise, named as an *internal SNR*. Listeners with better *internal SNR* showed better SiN performance. Further, we found that the post-speech time SMG activity explains a further amount of variance in SiN performance that is not accounted for by *internal SNR*. This result demonstrates that at least two cortical processes contribute to SiN performance independently: pre-target time processing to attenuate neural representation of background noise and post-target time processing to extract information from speech sounds.

[*]Corresponding author at: Department of Communication Sciences and Disorders, University of Iowa, 250 Hawkins Dr., Iowa City, IA 52242, USA. inyong-choi@uiowa.edu (I. Choi).

Keywords

Speech-in-noise; Speech unmasking; Speech recognition; Individual differences; Electroencephalography; Supramarginal gyrus

1. Introduction

Understanding speech in noise (SiN) is essential for communication in social settings. Young adult listeners with normal hearing are remarkably adept at this. Even in challenging SiN conditions where the speech and noise have the same intensity (i.e., 0 dB signal-to-noise ratio: SNR) and overlapped frequency components, they often recognize nearly 90% of sentences correctly (Ohlenforst et al., 2017; Plomp and Mimpen, 1979). This suggests a surprising capacity of the auditory system to cope with noise. However, the ability to understand SiN degrades severely with increased background noise level (Ohlenforst et al., 2017), hearing loss (Harris and Swenson, 1990), and/or aging (Goossens et al., 2017; Nabelek, 1988).

Recent studies show that normal-hearing listeners show large individual differences in SiN performance (Liberman et al., 2016). The premise of this study is that by linking this variable ability for SiN perception to variation in cortical activity, we may be able to understand the neural mechanisms by which humans accomplish this ability, and this may shape our understanding of how best to remediate hearing loss.

Two broad neural mechanisms might give rise to better or worse SiN performance. First, listeners may vary in neural processing that separates the target auditory object from the mixture of sounds (i.e., similar to *external* selection processes in Strauss and Francis (2017)). Auditory scene analysis (Bregman, 1999) processes, when occurring in parallel with auditory selective attention (Shinn-Cunningham, 2020), can inhibit the neural representation of competing sounds and enhance the neural response to attended input. This process has been conceptualized as a form of sensory gain control (Hillyard et al., 1998) or neural filtering (Obleser and Erb, 2020). Its effectiveness is often quantified as attentional modulation index (AMI), the amplitude ratio of evoked responses to background noise and target (Dai and Shinn-Cunningham, 2016; O'Sullivan et al., 2019), or the degree of neural phase-locking to the attended speech (Etard and Reichenbach, 2019; Mesgarani and Chang, 2012; Viswanathan et al., 2019). A successful sensory gain control, indicated by a positive AMI, during a SiN task will unmask the target speech from maskers, which will enhance the effective SNR in the central auditory pathway (e.g., the primary and secondary auditory cortices in the superior temporal plane: STP and the posterior superior temporal gyrus: STG).

Second, listeners might vary in neural processes for the prompt extraction of information from a speech signal (i.e., similar to *internal* selection processes in Strauss and Francis (2017). An inherent challenge in recognizing speech (e.g., a spoken word) is the mapping between the incoming speech cues and higher-level units like words and meaning while speech unfolds rapidly over time [for review, see Weber and Scharenborg (2012), Dahan and Magnuson (2006), and Davis (2016) with references therein]. In quiet listening conditions,

average young normal-hearing listeners activate a range of lexical candidates immediately at the onset of the auditory stimulus (i.e., shown by works using eye-movements in the visual world paradigm: (Allopenna et al., 1998; Dahan and Gareth Gaskell, 2007; Magnuson et al., 2007). For example, after hearing the/ba/at the onset of *bakery*, listeners will immediately consider a range of words like *bacon, bathe,* or *base*, at both phonological and semantic levels. However, such rapid lexical processing develops slowly in children (Rigler et al., 2015); continuous differences in lexical processing are linked to differences in language ability (McMurray et al., 2010); and they differ in listeners with hearing loss or deteriorated acoustic-cue encoding (McMurray et al., 2019). These facts suggest that there can be individual differences in the prompt lexical processing even for *clean* speech signals.

It is as yet unclear the degree to which variation in SiN performance is related to variation in both processes (particularly in combination).

1.1. Assessing individual differences in speech unmasking

Individual differences in the speech unmasking pathway may arise from 1) the fidelity of encoding supra-threshold acoustic features and 2) cognitive control of the domain-general attentional network. Auditory scene analysis relies on the supra-threshold acoustic features that provide binding cues for auditory grouping (Darwin, 1997). These include the spectra (Lee et al., 2013), location (Frey et al., 2014; Goldberg et al., 2014), temporal coherence (Moore, 1990; Shamma et al., 2013; Teki et al., 2011), rhythm (Calderone et al., 2014; Golumbic et al., 2013; Herrmann et al., 2016; Obleser and Kayser, 2019), and timing (Lange, 2009) of the figure and ground. The fidelity of encoding such supra-threshold acoustic features may affect the separation of target speech from background noise. Supporting this idea, previous studies have correlated the fidelity of supra-threshold acoustic cue coding to SiN understanding (Anderson and Kraus, 2010; Anderson et al., 2013; Holmes and Griffiths, 2019; Hornickel et al., 2009; Liberman et al., 2016; Parbery-Clark et al., 2009; Song et al., 2011).

Individual differences also exist in how strongly selective attention modulates cortical evoked responses to sounds (Choi et al., 2014). This suggests there may be a correlation between top-down selective attention efficacy and SiN performance. Indeed, (Strait and Kraus, 2011) reported that reaction time during a selective attention task predicts SiN performance. Similarly, studies suggest that poor cognitive control of executive attentional network predicts auditory selective attention performance (Bressler et al., 2017; Dai et al., 2018), though this has not been extended to SiN.

We can obtain a measure of the overall function of these bottom-up and top-down neural processing for speech unmasking by quantifying the amplitude ratio of early cortical auditory evoked responses to noise and target speech (similarly to the AMI concept). Here we use the N1/P2 event-related potential (ERP) components which occur with 100–300 ms latency. Previous studies showed that such ERP components are strongly modulated by selective attention but only when auditory objects are successfully segregated (Choi et al., 2013; Choi et al., 2014; Kong et al., 2015). Since those early cortical ERP components originate from multiple regions across Heschl's gyrus (i.e., the primary auditory cortex) and its surrounding areas (e.g., posterior superior temporal gyrus) (Čeponien et al., 1998), an

efficient and collective way of indexing the neural efficiency in speech unmasking is using a scalp electroencephalographical (EEG) potential at the vertex [e.g., "Cz" of the international 10–10 system for EEG electrode montage: Koessler et al. (2009) within a limited time-window (e.g., 100–300 ms range after the stimulus onset].

1.2. Assessing individual differences in mapping speech to words and meaning

To assess individual differences related to the second neural mechanism – the downstream speech information processing – we must assess a larger range of cortical regions above the auditory brainstem and cortex. Current models of speech processing suggest two distinct cortical networks (i.e., dorsal and ventral stream) that are used in parallel (Gow, 2012; Hickok and Poeppel, 2007; Myers et al., 2009; Scott and Johnsrude, 2003). The ventral stream pathway including anterior STG and middle temporal gyrus (MTG) integrates speech-acoustic and semantic information progressively over time for the sound-to-meaning mapping (Davis and Johnsrude, 2003). The dorsal stream pathway comprising supramarginal gyrus (SMG, also known as tempo-parietal junction or TPJ) and pre- / post-central gyri mediates the mapping between sound and articulation (Rauschecker and Scott, 2009), while inferior frontal gyrus (IFG) interacts with both pathways for lexical decision-making processes (Gow, 2012).

Studies have compared cortical responses in these areas to spoken words against acoustically-matching non-word sounds. These highlight the SMG/TPJ, MTG, and IFG in the left hemisphere as three regions that tend to exhibit more activity for words than pseudo-words (Davis and Gaskell, 2009; Taylor et al., 2013). The dual-lexicon model suggested by Gow (2012) confirms the importance of those three regions by referring to left SMG and MTG as dorsal and ventral lexicons that communicate with left IFG for lexical decision making. While both SMG and MTG exhibit explicitly lexical representations, they may take complementary roles consistent with the dual stream pathway model (Gow, 2012; Hickok and Poeppel, 2007). Thus, the type of task (e.g., whether subjects are asked to make a phonological or semantic judgment) may influence the relative dominance between SMG and MTG activities during speech recognition.

Supporting the idea of broad cortical regions contributing to individual differences in speech processing, fMRI studies showed that SNR changes alter the level of neural activities across frontal, central, and temporo-parietal regions (Du et al., 2016; Vaden et al., 2015; Wong et al., 2009; Zekveld et al., 2006), while Du et al. (2016) reported the correlation between activities in fronto-central regions and speech recognition performance. However, these correlations could reflect earlier variation in speech unmasking – if auditory/attentional unmasking mechanisms are less efficient, then they could lead to differences in how strongly later regions (IFG, SMG, etc.) must work to recognize words or complete the task. Thus, it is crucial to evaluate both mechanisms simultaneously to isolate a potential role for later processes.

In addition to testing the loci of activities, it is also important to test the relative timing of activity in these pathways during speech processing. Functionally, studies using eye-movements in the Visual World Paradigm (VWP) have extensively characterized the time course of word recognition in both quiet (Allopenna et al., 1998; Dahan and Gareth Gaskell,

2007; Magnuson et al., 2007) and under challenging conditions such as noise or signal degradation (Ben-David et al., 2011; Brouwer and Bradlow, 2016; Huettig and Altmann, 2005; McMurray et al., 2017; McQueen and Huettig, 2012). Most VWP data show that, in quiet, the maximum lexical competition occurs at ~400 ms after the onset. These studies also report delayed processing under challenging conditions, but such delays do not exceed 250 ms even under the most severe degradation. This timing information can guide us when we interpret the functional implication of neural activity within certain regions; if the latency of neural activity is larger than ~400 ms, such activity may be less likely related to the online word recognition.

However, the timing of cortical activity within each pathway and the way this may be moderated by challenging listening conditions are largely unknown. This is true both within isolated regions (e.g., SMG or MTG) and across activation of broader regions (e.g., frontal lobe). This is because most of the work on speech in noise perception has been conducted with fMRI (Du et al., 2014, 2016; Wong et al., 2009; Wong et al., 2008) which has a poor temporal resolution. One study has examined evoked responses across speech processing regions using source localized EEG (Bidelman and Howell, 2016). This suggests an early response at roughly 100 ms post-stimulus in IFG. However, this study used non-sense syllables that cannot reveal lexical processes.

1.3. The present study

The central aim of this study is to investigate the simultaneous contributions of both speech unmasking and speech recognition processes to the individual differences in SiN understanding. Our main question is whether the early-stage speech *unmasking* and later-stage *recognition* processes independently predict SiN performance, or whether the latter variable is dependent on the former. We attempted to answer this question through the combination of within- and across-subject analyses using both sensor- and source-space evoked responses in an EEG paradigm.

Subjects performed a SiN noise task in which they heard isolated consonant-vowel-consonant (CVC) English words and selected which of four orthographically presented words matched the auditory stimuli. Noise began 1 second before the speech. Two noise conditions were used: a low SNR (−3 dB) or hard condition, and high SNR (+ 3 dB) or easy condition. EEG was recorded from 64 electrodes while subjects performed this task, using both source- and sensor-space analyses to quantify cortical activity.

Speech unmasking and speech recognition can be distinguished by the 1) timing and 2) regional differences of neural activities evoked by both noise and target speech. Thus, we used a trial structure comprised of clearly separated events (i.e., fixed onsets of background noise and target speech) while observing time-locked neural responses to such events with EEG. The degree of speech unmasking was quantified as the amplitude ratio of evoked responses to the onsets of noise and target speech measured at a vertex scalp electrode (the key scalp location for evoked responses from early auditory processes), henceforth referred to as *internal SNR*. Although the concept of *internal SNR* is similar to the attentional modulation index (Dai and Shinn-Cunningham, 2016; O'Sullivan et al., 2019), we named

the index rather phenomenologically to avoid limiting the mechanism underlying the index to selective attention.

The effectiveness of later speech *recognition* processing was quantified by measuring the amplitude of evoked responses within target cortical regions. As we described, prior work has implicated a number of such regions. However, it is unclear which may be relevant for our specific task. Thus, we take a data-driven approach first by asking which post-auditory regions show *greater* activity in the *higher* SNR (easier) listening condition. This would be suggestive of a region that conducts downstream analyses once the target speech is unmasked. We looked into two regions-of-interest (ROIs): left SMG and IFG. As reviewed above, those regions activate more strongly for speech than pseudo-speech sounds; we did not consider MTG (the ventral lexicon) as our task was single CVC word identification (matched to an orthographic response). In this task, phonological discrimination (indicating the dorsal stream), rather than semantic processing, is essential.

Having quantified the contributions of each pathway, we then conducted both timing and individual differences analyses to determine their relative contribution to SiN performance.

2. Material and methods

2.1. Participants

Twenty-six subjects between 19 and 31 years of age (mean = 22.42 years, SD = 2.97 years; median = 21.5 years; 8 (31%) male) were recruited from a population of students at the University of Iowa. All subjects were native speakers of American English, with normal hearing thresholds no worse than 20 dB HL at any frequency, tested in octaves from 250 to 8000 Hz. Written informed consent was obtained, and all work has been carried out in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki). All study procedures were reviewed and approved by the University of Iowa Institutional Review Board.

2.2. Task design and procedures

We aimed to simultaneously measure SiN performance and cortical neural activity in a short (15 minute) experimental session. Sessions were kept short to avoid confounding individual differences in irrelevant psychological factors – fatigue, level or engagement –with individual differences in performance and processing.

Each trial (Fig. 1) began with the presentation of a fixation cross (‘+’) on the screen. Listeners were asked to fix their gaze on this throughout the trial to minimize eye-movement artifacts. Next, they heard the cue phrase “check the word.” This enabled listeners to predict the timing of next acoustic event (the noise onset). After 700 ms of silence, the multi-talker babble noise began and continued for 2 seconds. Babble-noise excerpts were extracted from an original five-minute-long eight-talker babble noise from the Revised SPIN Test (a compact disc published by Auditec, Inc., St. Louis, Missouri). We used 8-talker babble because it provides relatively stationary spectro-temporal characteristics compared to fewer-speaker babbles, while it still provides both temporal and spectral cues for speech unmasking. One second after the noise onset, the target word was heard. Finally, 100 ms

after the composite auditory stimulus (noise + word) offset, four written choices appeared on the screen. The response options differed either in the initial or the final consonant (e.g., for target word *ban*, options were *than, van, ban*, and *pan;* for target word *hiss,* options included *hit, hip, hiss, hitch*). Subjects pressed a button on a keypad to indicate their choice, and no feedback was given. The next trial began 1 second after the button press.

Choice options were displayed after the offset of sounds to prevent the occurrence of visual, pre-motor, and motor artifacts during the sounds, from which we extracted neural activities. The timing and intervals of auditory stimuli (i.e., cue phrase, noise, and target) were designed to derive well-distinct cortical evoked responses to the onsets of background noise and target word.

Since we were particularly interested in SMG and IFG regions that are involved in phonological and lexical processing (Gow, 2012; Hickok and Poeppel, 2007), we used naturally spoken words, rather than non-sense speech tokens used by prior EEG studies (Bidelman and Howell, 2016; Parbery-Clark et al., 2009). Target words consisted of hundred monosyllabic CVC words from the California Consonant Test (Owens and Schubert, 1977), spoken by a male speaker with a General American accent.

Target words were always presented at 65 dB SPL. In each trial, the RMS level of noise was chosen randomly between 68 and 62 dB SPL to yield either −3 or + 3dB SNR (referred to as “low SNR” and “high SNR,” respectively). We fixed the target word-level between the two SNR conditions to prevent potential acoustic-level effects on the target word-evoked responses. Fifty words were presented at each SNR condition. The number of items in each of the three phonetic categories – affricate & fricative, plosive, and sonorant – was balanced within each SNR condition and matched between conditions (See Table A.1 in Appendix). −3 dB SNR was chosen from pilot experiments to emulate a condition yielding a mid-point performance (~65% correct) in the possible accuracy range (i.e., 25 – 100%), at which listening effort and individual differences in performance may be maximized (Ohlenforst et al., 2017). Thus, the SiN performance at −3 dB SNR condition was used as the dependent variable for the later correlational analysis in this study. + 3 dB SNR was chosen to emulate a less noisy condition from which the downstream speech recognition process will be measured for the correlational analysis.

The task was implemented using the Psychtoolbox 3 package (Brainard, 1997; Pelli, 1997) for Matlab (R2016b, The Mathworks). Participants were tested in a sound-treated, electrically shielded booth with a single loudspeaker (model #LOFT40, JBL) positioned at a 0° azimuth angle at a distance of 1.2 m. A computer monitor was located 0.5m in front of the subject at eye level. The auditory stimuli were presented at the same levels for all subjects.

2.3. EEG acquisition and preprocessing

Scalp electrical activity (EEG) was recorded during the SiN task using the BioSemi ActiveTwo system at a 2048 Hz sampling rate. Sixty-four active electrodes were placed according to the international 10–20 configuration. Trigger signals were sent from Matlab (R2016b, The Math-works) to the BioSemi ActiView acquisition software. The recorded

EEG data from each channel were bandpass filtered from 1 to 50 Hz using a 2048-point zero-phase FIR filter. Epochs were extracted from −500 ms to 3 s relative to stimulus onset. After baseline correction using the average voltage between −200 and 0 ms, epochs were down-sampled to 256 Hz. No re-referencing was done for the sensor-space analysis, as the data from the BioSemi ActiveTwo system is referenced to the Common Mode voltage (https://www.biosemi.com/faq/cms&drl.htm). EEG data were re-referenced to the all-channel average before the source-space analysis.

Since we were interested in the speech-evoked responses from frontal brain regions, we opted for a non-modifying approach to eye blink rejection: Trials that were contaminated by an eye blink artifact were rejected based on the voltage value of the Fp1 electrode (bandpass filtered between 1 and 20 Hz). Each subject's rejection threshold was chosen by visual inspections on the distribution of maximum voltage at Fp1 across trials, as suggested by Luck (2014). Rejection thresholds ranged from 35 to 120 μV (mean 62.7 μV, standard deviation 23.7 μV). After rejecting bad trials, averages for each electrode were calculated for the two conditions to extract evoked potentials. For analysis of speech-evoked responses, we repeated baseline correction using the average signal in the 300 ms preceding the word onset.

2.4. Sensor-space analysis

We performed sensor-space ERP analysis to investigate the effect of acoustic SNR on the representation of noise and speech in the auditory cortex and its individual differences. The other purpose of the sensor-space analysis was to ensure the quality of data in a more familiar form before running source-space analyses.

Cortical evoked responses time-locked to the target and noise were examined and compared between high- and low-SNR conditions. EEG data were bandpass filtered from 2 to 7 Hz to capture 3 – 5 Hz bands in which auditory N1 and P2 components fall. We used a zero-phase 128-point (at 256-Hz sampling rate) FIR filter that has symmetric non-causal impulse responses. The zero-phase filter was chosen to avoid delays in "true" transient neural responses (see Figure 13 in de Cheveigne et al. (2019)). ERP envelopes were obtained by applying the Hilbert transform to the bandpass-filtered ERPs and taking the absolute value. Mean activity levels were jackknifed prior to testing to assess the variance with clean ERP waveforms. In this approach, the relevant neural factors were computed for all subjects but one. This was repeated leaving out each subject in turn. Peak amplitude was found from each jackknifed (i.e., leave-one-out grand average) ERP envelope within a 50 – 400ms time window following each of the noise and target-word onset. Then, "internal SNR" was defined as the amplitude ratio of target word-evoked ERP envelope peak to noise-evoked ERP envelope peak in dB scale (Eq. (1)). We computed this index expecting to quantify a "neural" form of an individual's speech unmasking ability. The internal SNR is different for each subject, and is separate from the fixed external, or acoustic, SNR (here, ± 3 dB).

$$\text{Internal SNR} = 20\log_{10}\frac{\text{Target} - \text{evoked potential amplitude}}{\text{Noise} - \text{evoked potential amplitude}} \quad (1)$$

The resulting statistics were adjusted for jackknifing to reflect the fact that each data point reflects N-1 subjects (Luck, 2014).

We did not have a hypothesis on which specific ERP component best predicts an individual's speech unmasking ability. Rather, we intended to quantify the overall amplitude across the canonical auditory ERP components (i.e., P1, N1, or P2) using the ERP envelope.

Like we did not rely on a single ERP component for the Internal SNR computation, we did not hypothesize that a single cortical region would reflect cortical processes for speech unmasking. Rather, we believed that scalp electrodes at the central front site would capture composite responses from multiple auditory-related regions, including both left and right Heschl's gyri and superior temporal gyri. Thus, the internal SNR computation has been done in the sensor space.

2.5. Source analysis

The source-space analysis was based on minimum norm estimation (Gramfort et al., 2013; Gramfort et al., 2014) as a form of multiple sparse priors (Friston et al., 2008). After co-registration of average electrode positions to the reconstructed average head model MRI, the forward solution (a linear operator that transforms source-space signals to sensor space) was computed using a single-compartment boundary-element model (Hämäläinen, 1989). We used FreeSurfer's "fsaverage" template brain and head model (Fischl et al., 1999). The cortical current distribution was estimated assuming that the orientation of the source is perpendicular to the cortical mesh. EEG data were re-referenced to the all-channel average. Then across-channel EEG noise covariance, computed for each subject, was used to calculate the inverse operators. A noise-normalization procedure was used to obtain dynamic statistical parametric maps (dSPMs) as *z*-scores (Dale et al., 2000). The inverse solution with a regularization parameter of $1/3^2$ estimated the source-space time courses of event-related activity at each of 10,242 cortical voxels per hemisphere.

Following the whole-brain source estimation, we extracted representative source time courses from ROIs. In the present study, two predetermined ROIs were used based on the literature review (e.g., Davis (2016); Gow (2012)): (1) left SMG, and (2) left pars opercularis and pars triangularis of IFG. Destrieux Atlas of cortical parcellation (Fischl et al., 2004) was used to predetermine ROIs anatomically.

Since we did not have individual structural MRI head models, it was not ideal for taking the summed activity (mean or median) for all the voxels within ROIs. This is because individual difference in functional and anatomical structure of the brain may result in spatial blurring since current densities across adjacent voxels can overlap each other. Instead, representative voxels were identified within each ROI, for each SNR condition. We used a combination of previously-described methods to select voxels of interest that were used in fMRI studies (Tong et al., 2016). The voxel selection was performed in a particular sequence described below.

- From each voxel in an ROI, reconstructed source time courses were averaged across subjects.

- Voxels exhibiting greater-than-median max amplitude over the time period after the target word onset were selected using the grand-average source time courses in each ROI.
- A matrix of cross-correlation coefficients was obtained using the source time courses in remaining voxels, quantifying between-voxel similarities.
- The most representative voxel was determined based on the maximum mean correlation coefficient.

For the downstream statistical analyses, temporal envelopes were extracted from the within-ROI source time courses. This was done by applying a bandpass filter, then calculating the absolute value of the Hilbert transform. The source time course envelopes at different ROIs were based on slightly different frequency ranges: 2–7 Hz for SMG and 1–5 Hz for IFG. This was to reflect the differences between temporal and frontal lobes in their dominant neural oscillations which create evoked activities through phase coherence (Giraud and Poeppel, 2012).

2.6. Statistical approaches

Once the temporal envelope of the source time course from the most representative voxel was obtained for each SNR condition, mean activity levels were compared between the two SNR conditions using paired *t*-tests. Here, we also used a jackknifing approach, and test statistics were adjusted to account for the fact that each data-point represents N-1 participants. Finally, to identify timepoints that showed a significant difference between SNR conditions while addressing multiple comparison problem, the cluster-based permutation tests were conducted (Maris and Oostenveld, 2007).

In order to identify predictors of SiN performance, sensor and source space indices of activity were used in correlation/regression analysis with SiN performance (accuracy) as the dependent variable, and the peak magnitudes of the ERP envelopes from ROIs, and internal SNR as the predictor variables. The peak magnitudes of the ERP envelopes were obtained over timepoints that showed a significant difference between high and low SNR conditions identified in the ROI-based source analysis described above. After calculating the correlation between SiN performance and these predictors, a joint contribution was tested using linear regression analysis to simultaneously examine bottom-up and compensatory related SiN performance to three factors.

3. Results

Our analysis started by examining the effect of SNR on task performance (accuracy and reaction time). This was intended to document that noise manipulation had the expected effect. Next, we examined the effect of SNR on both the magnitude and timing of neural activity. This was done first in the sensor-space, using the auditory N1/P2 components and ERP envelopes to examine primary auditory pathways. Next, this was done in the source-space to examine the compensatory role of IFG, to evaluate whether IFG effects were early enough to play a role in speech perception, and to test hypotheses about SMG. Finally, we turn to our primary analysis: a regression testing the unique contributions of each pathway to

individual differences in SiN performance. Original raw and processed data of the present study are available at Mendeley Data (http://dx.doi.org/10.17632/jyvythkz5y.2).

3.1. SiN performance

There was a large variance in performance among participants. This was observed in both the high SNR condition (accuracy: mean = 80.6%, SD = 7.8%; reaction time: mean = 1.5 s, SD = 0.3 s) and the low SNR condition (accuracy: mean = 68.2%, SD = 8.9%; reaction time: mean = 1.7 s, SD = 0.4 s) where the chance level of accuracy was 25%. The paired *t*-tests were performed on the rationalized arcsine units. There was a significant effect of SNR on both accuracy ($t(25) = 7.00$, $p < 0.001$) and reaction time ($t(25) = -3.97$, $p < 0.001$) (Fig. 2 A). Reaction time and accuracy were correlated in the high SNR condition (Fig. 2 B, $r = -0.50$, $p = 0.0090$), but not in the low SNR condition (Fig. 2C, $r = -0.19$, $p = 0.34$). Results from further analyses on the accuracy variance across phonetic categories are shown in Table A.1 in the Appendix section.

As a whole, these results validate that the SNR manipulation was sufficient to create differences in speech perception. The negative correlation between the accuracy and reaction time may demonstrate the redundancy between individual differences of accuracy and difficulty.

3.2. Auditory evoked activity in the sensor space: at the noise and speech onsets

We next examined the peak amplitudes of evoked activity at both noise and target-word onsets using the ERP envelopes. This was done to estimate the contribution of primary auditory pathways and the effect of noise. These measures were compared between high- and low-SNR conditions using paired *t*-tests.

At the noise onset, as expected, the low SNR condition exhibited significantly larger peak envelope amplitude (i.e., at ~0.2s after noise onset: See Fig. 3. $t(25) = -3.95$, $p < 0.001$ from paired t-test). At the word onset, we saw larger peak envelope amplitude in the high SNR condition ($t(25) = 2.37$, $p = 0.026$) (1.2 – 1.3s in Fig. 3). Then we calculated the internal SNR, the amplitude ratio of the noise and target-word related ERP envelope peaks, and saw a greater internal SNR in the high SNR condition ($t(25) = 2.53$, $p = 0.018$).

The topographical layout of SNR effects is shown in Fig. 3 (top panels) at the time of the peak in envelopes for both noise- and word-evoked response. T-values from paired t-tests on peak envelope amplitudes at all electrodes between SNR conditions were also represented in topographies. These show a broad-based effect that is roughly centered at frontal-central channels for both the noise onset and speech onset. This justifies our use of frontal-central channels for sensor-space analyses. The significant difference in auditory ERP envelopes according to the noise level supports our use of the internal SNR as an index of individual ability to modulate representations of target speech relative to noise in the regression analysis.

3.3. The effect of SNR on cortical activity

Next, we conducted parallel analyses in source space to assess cortical activity through speech processing regions. We converted sensor-space EEG signals to whole-brain source time courses to localize the effects of SNR on evoked responses within targeted ROIs. Within left SMG, the cluster-based permutation test (Maris and Oostenveld, 2007) revealed that the high SNR condition evokes significantly greater activity than the low SNR condition from 270 to 340 ms ($p = 0.0020$) (Fig. 4 A **left**). High-SNR peak amplitude is found at 309 ms. Such a significant SNR effect was not found in the left IFG. The maximum magnitude of the grand average low-SNR evoked response (dSPM) is at 770 ms (Fig. 4 B **left**). Source time courses in the right SMG and IFG are shown in Fig. 4 A **and** B (the right panels) for visual comparisons.

Single time-point evoked-current estimates are shown for the peak SMG activity time (i.e., 309 ms, Fig. 4C) and IFG peak time (770 ms, Fig. 4 D) on the whole left-hemisphere cortical surface. At 309 ms, post-hoc paired *t*-tests on all the left-hemisphere voxels reveal an area near left SMG that shows greater evoked responses in the high SNR condition. This supports a significant role for SMG in SiN processing in this task. In contrast, at 770 ms, no voxel was found within IFG that shows significant differences between high- vs. low-SNR conditions. This confirms our timecourse analyses of IFG, suggesting it does not play a large role and that the trend that was observed is not broadly seen across voxels.

The webMAUS (Kisler et al., 2017) was used to identify the boundaries between the first and second and the second and third phonemes in each of the 100 stimuli to demonstrate the timing of SMG activity compared to the timing of phonological events. A histogram of these acoustic time points is shown in Fig. 5. This confirms that the peak of evoked activation in SMG (i.e., ~309 ms, denoted by a dashed vertical line) occurs before the end of target words.

3.4. Individual differences in internal SNR predict SiN performance

To address our primary research question, which was to evaluate the simultaneous contribution of speech unmasking and recognition processes to SiN performance, we conducted a linear regression analysis in which internal SNR and SMG activation were used as independent variables. SiN performance in the low SNR condition was used as the dependent variable. We extracted the internal SNR from the low SNR and the SMG activity from the high SNR condition, as we expected that the internal SNR captures how well listeners unmask speech from the noisy background while the SMG activity reflects the processing of relatively clean speech signal. As expected, those two metrics extracted from different trials did not show a correlation ($r = -0.040$, $p = 0.86$, the left panel of Fig. 6 D). Both internal SNR ($t(23) = 3.35$, $p = 0.0030$) and SMG activity ($t(23) = 2.29$, $p = 0.031$) were significant predictors of SiN performance in low SNR condition (Fig. 6 A). The linear combination of those predictors accounted for a large proportion of the variance ($r = 0.64$, $p = 0.00043$, Fig. 6 B). However, both internal SNR ($t(23) = 1.04$, $p = 0.31$) and SMG activity ($t(23) = 0.17$, $p = 0.87$) did not explain the variance in SiN performance in high SNR condition.

Fig. 6C **and** D show results from post-hoc correlational analyses. Internal SNR showed a significant correlation with accuracy in low SNR, while SMG activation did not (despite its significant contribution to the model). There was no correlation between internal SNR and SMG activation, as described above. A semi-partial correlation between SMG activation and the residual of accuracy after regressing out internal SNR was significant, which confirmed that the SMG activation accounted for an extra amount of variance in SiN performance in low SNR (the right panel of Fig. 6 D). This suggests that in order to identify the contribution of downstream recognition areas like SMG, models must account for the contribution of earlier upstream speech-unmasking processes.

To visualize the contribution of internal SNR to SiN performance, Fig. 6 E showed evoked response differences between good and poor performers (based on a median split on the low-SNR condition accuracy). This reveals dramatic differences in the magnitude of noise onset-related potentials: despite the same physical noise level for each group, good performers exhibited less strong evoked response to noise onset, measured by the envelope peak magnitude within N1-P2 time range in the frontal-central channels ($t(24) = -2.60$, $p = 0.016$, two-sample *t*-test). In contrast, the word-evoked ERP envelope did not show a significant difference between the two groups ($t(24) = 0.21$, $p = 0.84$). This suggests that the neural mechanism underlying the internal SNR variance is the suppression of noise (rather than the enhancement of the target). The ERP envelopes to the cue phrase were investigated as a control condition and showed no difference between the two groups ($t(24) = -0.38$, $p = 0.71$), suggesting that the effect was directly related to the speech task (Fig. A.1).

4. Discussion

We investigated the neural correlates of SiN performance in young normal-hearing adults. Previous correlational studies focused on the contributions of either acoustic encoding fidelity (Anderson and Kraus, 2010; Anderson et al., 2013; Holmes and Griffiths, 2019; Hornickel et al., 2009; Liberman et al., 2016; Parbery-Clark et al., 2009; Song et al., 2011) or the degree of speech/language network recruitment (Du et al., 2016) to the SiN performance. However, the relative importance of each process has remained unclear. We showed that 1) how well the listener suppresses background noise *before* hearing the target speech and 2) how strongly the listener recruits temporo-parietal network *while* the speech signal is received contribute to the SiN performance independently. Combining those two factors explained about 40% of the variance in SiN performance.

Our results have both theoretical and clinical implications. Theoretically, our individual difference approach revealed at least two neural subsystems involve during SiN processing: sensory gain control and post-auditory speech recognition processing. Clinically, our results suggest that a relatively short (~15 minutes) SiN-EEG paradigm can assess crucial neural processes for SiN understanding.

4.1. Internal SNR: a measure of pre-speech processing for speech unmasking

The first among the two crucial processes – how well the listener suppresses background noise – was indexed as “internal SNR,” the ratio of noiseto target word-evoked cortical responses. This process can be understood as a pre-target cortical activity, appearing as an

enhanced neural representation of the target sound (the speech) and suppressed neural representation of ignored stimuli (the noise).

4.2. What is the source of variation in the internal SNR?

Such responses could reflect auditory selective attention, which shows a similar pattern in previous studies (Hillyard et al., 1973; Hillyard et al., 1998; Mesgarani and Chang, 2012). In the present study, good performers showed significantly weaker noise-evoked responses at frontal-central channels (around Cz), compared with poor performers, approximately 200 ms after the noise onset (Fig. 6 E). Decreased auditory responses to background noise in good performers are compatible with the presence of a sensory gain control mechanism (Hillyard et al., 1998) which may happen in multiple sub-regions in STP and posterior STG. The variation in the sensory gain control may originate from multiple factors. It may reflect the acuity of encoding spectro-temporal acoustic cues from speech and noise or grouping of such acoustic cues for auditory object formation (Moore, 1990; Shamma et al., 2013; Teki et al., 2011). How robustly the low-frequency neural oscillations (e.g., theta and delta) are phase-locked to the acoustic temporal structure of the stimuli (Etard and Reichenbach, 2019) may also contribute to the variation, as the neural phase-locking relies on the encoding of acoustic cues (Ding et al., 2014) and the prediction of temporal structure in speech rhythm (Ding et al., 2016). Since our experiment provided fixed timing of noise and target word onsets, the neural phase-locking based on predicted timing (See Fig. 1 of (Arnal and Giraud, 2012)) could occur and contribute to the internal SNR. Indeed, studies show that neural entrainment to the speech rhythm is one of the possible mechanisms underlying cocktail party listening (Golumbic et al., 2013). Studies about the relationship between musical ability and speech in noise performance also revealed that the sensitivity to “rhythm” exhibits the strongest correlation with SiN performance among different aspects of musical abilities (Yates et al., 2019).

The fact that evoked activity at the noise onset dominantly contributed to internal SNR makes it difficult to interpret that the internal SNR is driven by a “grouping” mechanism. To test whether the ERP amplitude to noise onset is driven by inherent differences in subjects, we observed ERPs to a more neutral event “check the word.” See Fig. 7 for the comparison between good vs. poor performers at their evoked responses to the carrier phrase “check the word,” which did not exhibit significant differences. This result may imply that the individual differences in internal SNR may not be driven by individuals’ inherent differences in auditory evoked potentials.

The variation may also reflect endogenous mechanisms for active suppression of background sounds along with neural enhancement of foreground sounds (Shinn-Cunningham and Best, 2008). It was not our goal to disentangle the sources of variation in sensory gain control. Rather, we aimed to quantify the effectiveness of sensory gain control by our unique trial structure that enables clear distinction of evoked responses to noise and target speech, and test how the internal SNR predicts later speech processes and behavioral accuracy. In this regard, we found a significant correlation between accuracy and the relative magnitude of the word- and noise-evoked potentials.

4.3. Evoked amplitude in SMG: the neural marker of effective and prompt lexical processing

While the computation of internal SNR was pre-specified, we had an open plan for extracting a representative neural factor to capture post-auditory speech recognition. To explore such neural markers, we added a 6-dB higher SNR condition and asked which region or regions showed increased activity within a reasonable (200 – 500 ms) time range. We investigated two ROIs: left SMG and left IFG. As left SMG showed increased evoked response to target speech in the less noisy condition at ~300 ms after the target onset, the peak evoked amplitude in left SMG measured in the high SNR condition was used as the second independent variable in the regression analysis.

4.4. Functional interpretation of SMG activity

Previous studies have suggested that spoken-word recognition occurs via a process of dynamic lexical competition as speech unfolds over time. The VWP studies reported that, for many words, this competition maximizes around 3–400 ms after word onset (Farris-Trimble and McMurray, 2013; Huettig and Altmann, 2005). In significantly challenging conditions (high noise), however, lexical processing can be delayed about 250 ms until most of the word has been heard (Farris-Trimble et al., 2014; McMurray et al., 2017), which may minimize competition. The latency of SMG activity that lied between the second and the third phonemes (see Fig. 5) in the high SNR condition aligns well with the timing of lexical competition found from the VWP studies, which may suggest that the SMG activity makes a neural substrate of immediate lexical access (Farris-Trimble et al., 2014; McMurray et al., 2017), consistent with Gow (2012). This immediacy was observed when speech sounds were relatively clean (high SNR), and it does not appear in previous EEG studies using non-word synthesized phonemes (Bidelman and Dexter, 2015; Bidelman and Howell, 2016).

After the contribution of speech unmasking (i.e., internal SNR) is regressed out, the SMG evoked amplitude in the cleaner condition predicted the residual of SiN performance (the right panel of Fig. 6 D). This indicates that changes in SMG activity may be an independent factor predicting speech recognition performance, rather than the outcome of pre-speech sensory gain control processing.

It can be seen surprising that the overall level of activity is much lower within SMG while the performance in the low SNR condition is still much above the chance level. One potential interpretation is, only when they are clearly heard, spoken words recruit “time-locked” online lexical processes that recruit SMG. When sounds are ambiguous (e.g., due to background noise), such “time-locked” activity weakens while the speech recognition routes through an alternative pathway, reflecting a different cognitive strategy. If such an alternative cognitive strategy for a poorer SNR condition works, accuracy can be preserved while the neural processing to achieve the performance is largely altered.

4.5. Limitation of the current study

In the present paper, we exhibited evoked responses only. Although our results demonstrated how this simple and traditional EEG analysis successfully predicted SiN performance, future

studies may pursue further understanding of SiN mechanisms by adopting extended analyses such as induced oscillation (e.g., Choi et al. (2020)) and connectivity analyses.

Our correlational result is limited to young normal-hearing listeners where variance does not come from hearing deterioration and aging. This study does not predict how much of the variance can be explained by the combination of internal SNR and SMG activity in the population with larger age ranges. For example, Tune et al. (2020) reported that no correlation is found between neural attention filters and behavioral success in a large cohort of aged listeners, although other studies revealed neural factors contributing to speech-in-noise performance (Decruy et al., 2019; Presacco et al., 2019).

We chose −3 dB as the main SNR from which the dependent variable for the regression analysis was extracted. Although we claim that the −3 dB SNR provides the most representative condition where the individual difference in performance is maximized, we do not claim that our main findings can be generalized to other SNR conditions.

The speech-rhythm entrainment hypothesis can be more directly tested using continuous speech stimuli (sentences) using established temporal response function methods introduced by recent papers (Ding and Simon, 2012; Lalor and Foxe, 2010; Vanthornhout et al., 2018)

Even for normal-hearing listeners, ERPs are strongly determined by sensation level. It would make sense to as a control include an audiogram-based predictor (such as a pure-tone average) in the model you use to predict speech intelligibility. However, in this study, subjects were “screened” for “normal hearing” (i.e., = < 20 dB HL across tested frequencies: 250Hz to 8000Hz), but their thresholds were not measured. Our future study will further investigate peripheral factors contributing to the individual differences in cortical process and speech-in-noise accuracy.

4.6. Methodological advances and justifications for source time course analysis

Our approach to identifying a single voxel within an ROI deserves a particular discussion. Identification of the representative voxel of an ROI is a problem common to EEG source analysis, fMRI, and other functional brain imaging studies. Many relevant neuroimaging analysis approaches have been described, including univariate, multivariate, and machine learning; however, most of these are intended for the identification of regions of interest or functional connections from a whole-brain map. Drawbacks of this type of whole-brain analysis include the need for strict multiple comparisons correction and, therefore, decreased statistical power. Using strong a priori hypotheses to generate regions of interest allowed us to circumvent these issues, but still requires identification of representative voxels within our regions of interest. Favored approaches generally require the identification of peak activity within an ROI (Tong et al., 2016). However, to avoid assuming that choosing peak activity implies, we opted instead to choose the voxel with the maximum average correlation to every other voxel within the ROI. In the present study, we chose not to constrain the location of the voxel of interest within an ROI for each condition. Because our anatomic resolution is unlikely to be at the voxel level, we elected to choose a different representative voxel for each condition, unconstrained by the location of the representative voxel from other conditions.

5. Conclusion

We found that better speech unmasking in good performers modulated the ratio of cortical evoked responses to the background noise and target sound, which effectively changed SNR internally, resulting in better performance. We also found that clean, intelligible speech elicits early processing at SMG, which explained an extra amount of variance in SiN performance. These findings may collectively form a neural substrate of individual differences in SiN understanding ability; the variance in SiN perception may be a matter of both primary processes that extract the signal from noise and later speech recognition processes to extract lexical information from speech signals promptly.

Acknowledgments

This work was supported by the Department of Defense Hearing Restoration and Rehabilitation Program grant awarded to Choi (W81XWH1910637), NIH T32 (5T32DC000040-24) awarded to Schwalje, and NIDCD P50 (DC000242 31) awarded to Griffiths and McMurray. The authors declare no competing financial interests.

Data and code availability statement

The data that support the findings of the present study are openly available in Mendeley Data at http://dx.doi.org/10.17632/jyvythkz5y.2.

Appendix

We further looked into the SNR effect on accuracy in three phonetic categories: “affricate and fricative,” “plosive,” and “sonorant.” As shown in Table A.1, affricates and fricatives had the lowest mean accuracy (67.4% in the + 3dB SNR condition, 52.9% in the −3dB SNR) with ~30 (in % correct) standard deviation across participants. Sonorants had the highest accuracy and smaller standard deviation across participants. Importantly, the following table also shows that the number of items in each phonetic category matched well between conditions in our study.

In order to demonstrate the cortical map in Fig. 4C more clearly, a larger size figure is attached (Fig. A.1). It is observed that the dominant activity is found across SMG and inferior post-central sulcus, which is just underneath the SMG if we imagine the original folded brain structure. About 40% of the area remains in the anatomical SMG boundary.

Table A.1

Target Consonant	SNR: +3dB		SNR: −3dB	
	Accuracy (%)	Number of Items	Accuracy (%)	Number of Items
Affricate & Fricative	67.4 (std 27.4)	17 (34%)	52.9 (std 31.6)	16 (32%)
Plosive	83.6 (std 25.0)	16 (32%)	70.2 (std 25.9)	17 (34%)
Sonorant	89.6 (std 16.0)	17 (32%)	81.9 (std 20.4)	17 (34%)

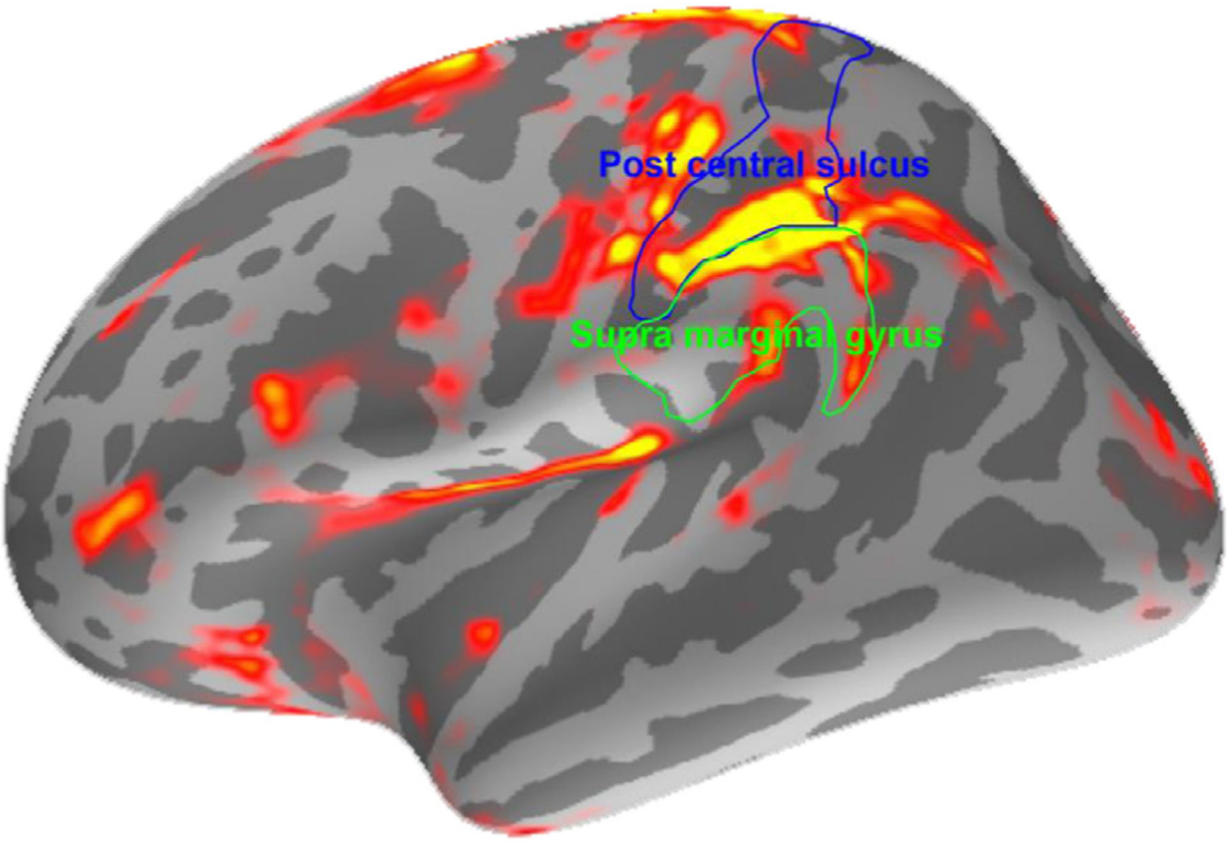

Fig. A.1.

Cortical activity at 309ms after target word onset in the high SNR condition. A dominant activity is found across supramarginal gyrus and inferior post-central sulcus.

References

Allopenna PD, Magnuson JS, Tanenhaus MK, 1998. Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. J. Mem. Lang 38, 419–439.

Anderson S, Kraus N, 2010. Objective neural indices of speech-in-noise perception. Trends Amplif. 14, 73–83. [PubMed: 20724355]

Anderson S, Parbery-Clark A, White-Schwoch T, Kraus N, 2013. Auditory brainstem response to complex sounds predicts self-reported speech-in-noise Performance. J. Speech, Lang. Hear. Res 56, 31–43. [PubMed: 22761320]

Arnal LH, Giraud AL, 2012. Cortical oscillations and sensory predictions. Trends Cogn. Sci 16, 390–398. [PubMed: 22682813]

Ben-David BM, Chambers CG, Daneman M, Pichora-Fuller MK, Reingold EM, Schneider BA, 2011. Effects of aging and noise on real-time spoken word recognition: evidence from eye movements. J. Speech Lang. Hear. Res 54, 243–262. [PubMed: 20689026]

Bidelman GM, Dexter L, 2015. Bilinguals at the “cocktail party”: dissociable neural activity in auditory-linguistic brain regions reveals neurobiological basis for nonnative listeners’ speech-in-noise recognition deficits. Brain Lang. 143, 32–41. [PubMed: 25747886]

Bidelman GM, Howell M, 2016. Functional changes in inter- and intra-hemispheric cortical processing underlying degraded speech perception. Neuroimage 124, 581–590. [PubMed: 26386346]

Brainard DH, 1997. The psychophysics toolbox. Spat. Vis 10, 433–436. [PubMed: 9176952]

Bregman AS, 1999. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge, Mass.

Bressler S, Goldberg H, Shinn-Cunningham B, 2017. Sensory coding and cognitive processing of sound in Veterans with blast exposure. Hear. Res 349, 98–110. [PubMed: 27815131]

Brouwer S, Bradlow AR, 2016. The temporal dynamics of spoken word recognition in adverse listening conditions. J. Psycholinguist Res 45, 1151–1160. [PubMed: 26420754]

Calderone DJ, Lakatos P, Butler PD, Castellanos FX, 2014. Entrainment of neural oscillations as a modifiable substrate of attention. Trends Cognit. Sci 18, 300–309. [PubMed: 24630166]

Čeponien R, Cheour M, Näätänen R, 1998. Interstimulus interval and auditory event-related potentials in children: evidence for multiple generators. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section 108, 345–354.

Choi I, Kim S, Choi I, Schwalje A, 2020. Cortical dynamics of speech-in-noise understanding. Acoust. Sci. Technol 41, 400–403.

Choi I, Rajaram S, Varghese LA, Shinn-Cunningham BG, 2013. Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. Front. Hum. Neurosci 7, 115. [PubMed: 23576968]

Choi I, Wang L, Bharadwaj H, Shinn-Cunningham B, 2014. Individual differences in attentional modulation of cortical responses correlate with selective attention performance. Hear. Res 314, 10–19. [PubMed: 24821552]

Dahan D, Gareth Gaskell M, 2007. The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. J. Mem. Lang 57, 483–501. [PubMed: 18071581]

Dahan D, Magnuson JS, 2006. Spoken Word Recognition. Handbook of psycholinguistics. Elsevier, pp. 249–283.

Dai L, Shinn-Cunningham BG, 2016. Contributions of sensory coding and attentional control to individual differences in performance in spatial auditory selective attention tasks. Front. Hum. Neurosci 10, 530. [PubMed: 27812330]

Dai L, Shinn-Cunningham BG, Best V, 2018. Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. Proc. Natl. Acad. Sci. U. S. A 115, E3286–E3295. [PubMed: 29555752]

Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E, 2000. Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. Neuron 26, 55–67. [PubMed: 10798392]

Darwin CJ, 1997. Auditory grouping. Trends Cogn. Sci 1, 327–333. [PubMed: 21223942]

Davis MH, 2016. The neurobiology of lexical access. In: Neurobiology of Language. Elsevier, pp. 541–555.

Davis MH, Gaskell MG, 2009. A complementary systems account of word learning: neural and behavioural evidence. Philos. Trans. R. Soc. Lond. Ser. B, Biol. Sci 364, 3773. [PubMed: 19933145]

Davis MH, Johnsrude IS, 2003. Hierarchical processing in spoken language comprehension. J. Neurosci 23, 3423–3431. [PubMed: 12716950]

de Cheveigne A, de Cheveigne A, de Cheveigne A, Nelken I, 2019. Filters: when, why, and how (Not) to use them. Neuron 102, 280–293. [PubMed: 30998899]

Decruy L, Vanthornhout J, Francart T, 2019. Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. J. Neurophysiol 122, 601–615. [PubMed: 31141449]

Ding N, Melloni L, Zhang H, Tian X, Poeppel D, 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci 19, 158–164. [PubMed: 26642090]

Ding N, Simon JZ, 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. procnatiacadscie. Proc. Natl. Acad. Sci. USA 109, 11854–11859. [PubMed: 22753470]

Ding N, Simon JZ, Chatterjee M, 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage 88, 41–46. [PubMed: 24188816]

Du Y, Buchsbaum BR, Grady CL, Alain C, 2014. Noise differentially impacts phoneme representations in the auditory and speech motor systems. Proc. Natl. Acad. Sci. U S A 111, 7126–7131. [PubMed: 24778251]

Du Y, Buchsbaum BR, Grady CL, Alain C, 2016. Increased activity in frontal motor cortex compensates impaired speech perception in older adults. Nat. Commun 7, 12241. [PubMed: 27483187]

Etard O, Reichenbach T, 2019. Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. J. Neurosci 39, 5750–5759. [PubMed: 31109963]

Farris-Trimble A, McMurray B, 2013. Test-retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. J. Speech Lang. Hear. Res 56, 1328–1345. [PubMed: 23926331]

Farris-Trimble A, McMurray B, Cigrand N, Tomblin JB, 2014. The process of spoken word recognition in the face of signal degradation. J. Exp. Psychol. Hum. Percept. Perform 40, 308–327. [PubMed: 24041330]

Fischl B, Sereno MI, Tootell RBH, Dale AM, 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp 8, 272–284. [PubMed: 10619420]

Fischl B, van der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, Busa E, Seidman LJ, Goldstein J, Kennedy D, Caviness V, Makris N, Rosen B, Dale AM, 2004. Automatically parcellating the human cerebral cortex. Cereb. Cortex 14, 11–22. [PubMed: 14654453]

Frey JN, Mainy N, Lachaux JP, Muller N, Bertrand O, Weisz N, 2014. Selective modulation of auditory cortical alpha activity in an audiovisual spatial attention task. J. Neurosci 34, 6634. [PubMed: 24806688]

Friston K, Harrison L, Daunizeau J, Kiebel S, Phillips C, Trujillo-Barreto N, Henson R, Flandin G, Mattout J, 2008. Multiple sparse priors for the M/EEG inverse problem. Neuroimage 39, 1104–1120. [PubMed: 17997111]

Giraud AL, Poeppel D, 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci 15, 511–517. [PubMed: 22426255]

Goldberg HR, Choi I, Varghese LA, Bharadwaj H, Shinn-Cunningham BG, 2014. Auditory attention in a dynamic scene: Behavioral and electrophysiological correlates. J. Acoust. Soc. Am 135, 2415.

Golumbic EMZ, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Good-man RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE, 2013. Mechanisms underlying selective neuronal tracking of attended speech at a cocktail party. Neuron 77, 980–991. [PubMed: 23473326]

Goossens T, Vercammen C, Wouters J, van Wieringen A, 2017. Masked speech perception across the adult lifespan: impact of age and hearing impairment. Hear. Res 344, 109–124. [PubMed: 27845259]

Gow DW Jr., 2012. The cortical organization of lexical knowledge: a dual lexicon model of spoken language processing. Brain Lang. 121, 273–288. [PubMed: 22498237]

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hamalainen M, 2013. MEG and EEG data analysis with MNE-Python. Front. Neurosci 7, 267. [PubMed: 24431986]

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hamalainen MS, 2014. MNE software for processing MEG and EEG data. Neuroimage 86, 446–460. [PubMed: 24161808]

Hämäläinen MS, Sarvas J, 1989. Realistic conductivity geometry model of the human head for interpretation of neuromagnetic data. IEEE Trans. Biomed. Eng 36, 165–171. [PubMed: 2917762]

Harris RW, Swenson DW, 1990. Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment. Audiology 29, 314–321. [PubMed: 2275646]

Herrmann B.r., Henry MJ, Haegens S, Obleser J, 2016. Temporal expectations and neural amplitude fluctuations in auditory cortex interactively influence perception. NeuroImage 124, 487–497. [PubMed: 26386347]

Hickok G, Poeppel D, 2007. The cortical organization of speech processing. Nat. Rev. Neurosci 8, 393–402. [PubMed: 17431404]

Hillyard SA, Hink RF, Schwent VL, Picton TW, 1973. Electrical signs of selective attention in the human brain. Science 182, 177–180. [PubMed: 4730062]

Hillyard SA, Vogel EK, Luck SJ, 1998. Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philos. Trans. R. Soc. Lond. B Biol. Sci 353, 1257–1270. [PubMed: 9770220]

Holmes E, Griffiths TD, 2019. Normal' hearing thresholds and fundamental auditory grouping processes predict difficulties with speech-in-noise perception. Sci. Rep 9.

Hornickel J, Skoe E, Nicol T, Zecker S, Kraus N, 2009. Subcortical differentiation of stop consonants relates to reading and speech-in-noise perception. Proc. Natl. Acad. Sci. U S A 106, 13022–13027. [PubMed: 19617560]

Huettig F, Altmann GT, 2005. Word meaning and the control of eye fixation: semantic competitor effects and the visual world paradigm. Cognition 96, B23–B32. [PubMed: 15833303]

Kisler T, Reichel UD, Sciel F, 2017. Multilingual processing of speech via web services. Comput. Speech Lang 45, 326–347.

Koessler L, Maillard L, Benhadid A, Vignal JP, Felblinger J, Vespignani H, Braun M, 2009. Automated cortical projection of EEG sensors: Anatomical correlation via the international 10–10 system. NeuroImage 46, 64–72. [PubMed: 19233295]

Kong YY, Somarowthu A, Ding N, 2015. Effects of spectral degradation on attentional modulation of cortical auditory responses to continuous speech. JARO -NEW YORK-16, 783–796.

Lalor EC, Foxe JJ, 2010. TECHNICAL SPOTLIGHT: Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci 31, 189–193. [PubMed: 20092565]

Lange K, 2009. Brain correlates of early auditory processing are attenuated by expectations for time and pitch. Brain Cognit. 69, 127 [PubMed: 18644669]

Lee AKC, Rajaram S, Xia J, Bharadwaj H, Larson E, Hämäläinen MS, Shinn-Cunningham BG, 2013. Auditory selective attention reveals preparatory activity in different cortical regions for selection based on source location and source pitch. Front. Neurosci 6, 190. [PubMed: 23335874]

Liberman MC, Epstein MJ, Cleveland SS, Wang H, Maison SF, 2016. Toward a differential diagnosis of hidden hearing loss in humans. PLoS One 11, 1–16.

Luck SJ, 2014. An Introduction to the Event-Related Potential Technique, Second edition The MIT Press, Cambridge, Massachusetts.

Magnuson JS, Dixon JA, Tanenhaus MK, Aslin RN, 2007. The dynamics of lexical competition during spoken word recognition. Cogn. Sci 31, 133–156. [PubMed: 21635290]

Maris E, Oostenveld R, 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. [PubMed: 17517438]

McMurray B, Ellis TP, Apfelbaum KS, 2019. How do you deal with uncertainty? Cochlear implant users differ in the dynamics of lexical processing of noncanonical inputs. Ear Hear. 40, 961–980. [PubMed: 30531260]

McMurray B, Farris-Trimble A, Rigler H, 2017. Waiting for lexical access: cochlear implants or severely degraded input lead listeners to process speech less incrementally. Cognition 169, 147–164. [PubMed: 28917133]

McMurray B, Samelson VM, Lee SH, Bruce Tomblin J, 2010. Individual differences in online spoken word recognition: Implications for SLI. Cognit. Psychol. Cognit. Psychol 60, 1–39. [PubMed: 19836014]

McQueen JM, Huettig F, 2012. Changing only the probability that spoken words will be distorted changes how they are recognized. J. Acoust. Soc. Am 131, 509–517. [PubMed: 22280612]

Mesgarani N, Chang EF, 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236. [PubMed: 22522927]

Moore BC, 1990. Co-modulation masking release: spectro-temporal pattern analysis in hearing. Br. J. Audiol 24, 131–137. [PubMed: 2190655]

Myers EB, Blumstein SE, Walsh E, Eliassen J, 2009. Inferior frontal regions underlie the perception of phonetic category invariance. PSCI Psychol. Sci 20, 895–903.

Nabelek AK, 1988. Identification of vowels in quiet, noise, and reverberation: relationships with age and hearing loss. J. Acoust. Soc. Am 84, 476–484. [PubMed: 3170940]

O'Sullivan J, Mesgarani N, Smith E, Schevon C, McKhann GM, Sheth SA, Herrero J, Mehta AD, Smith E, Sheth SA, 2019. Hierarchical encoding of attended auditory objects in multi-talker speech perception. Neuron 104, 1195–120 9e1193. [PubMed: 31648900]

Obleser J, Erb J, 2020. Neural Filters for Challenging Listening Situations. The Cognitive Neurosciences MIT Press.

Obleser J, Kayser C, 2019. Neural entrainment and attentional selection in the listening brain. Trends Cogn. Sci 23, 913–926. [PubMed: 31606386]

Ohlenforst B, Zekveld AA, Lunner T, Wendt D, Naylor G, Wang Y, Versfeld NJ, Kramer SE, 2017. Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. Hear. Res 351, 68–79. [PubMed: 28622894]

Owens E, Schubert ED, 1977. Development of the California consonant test. J. Speech Lang. Hear. Res 20, 463–474.

Parbery-Clark A, Skoe E, Kraus N, 2009. Musical experience limits the degradative effects of background noise on the neural processing of sound. J. Neurosci 29, 14100–14107. [PubMed: 19906958]

Pelli DG, 1997. The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat. Vis 10, 437–442. [PubMed: 9176953]

Plomp R, Mimpen AM, 1979. Speech-reception threshold for sentences as a function of age and noise level. J. Acoust. Soc. Am 66, 1333–1342. [PubMed: 500971]

Presacco A, Simon JZ, Anderson S, 2019. Speech-in-noise representation in the aging midbrain and cortex: Effects of hearing loss. PLoS One 14, e0213899. [PubMed: 30865718]

Rauschecker JP, Scott SK, 2009. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat. Neurosci 12, 718–724. [PubMed: 19471271]

Rigler H, Farris-Trimble A, Greiner L, Walker J, Tomblin JB, McMurray B, 2015. The slow developmental timecourse of real-time spoken word recognition. Dev. Psychol 51, 1690–1703. [PubMed: 26479544]

Scott SK, Johnsrude IS, 2003. The neuroanatomical and functional organization of speech perception. Trends Neurosci. 26, 100–107. [PubMed: 12536133]

Shamma S, Elhilali M, Ma L, Micheyl C, Oxenham AJ, Pressnitzer D, Yin P, Xu Y, 2013. Temporal coherence and the streaming of complex sounds. Adv. Exp. Med. Biol 787, 535–543. [PubMed: 23716261]

Shinn-Cunningham BG, 2020. 14 Brain mechanisms of auditory scene analysis. Cognit. Neurosci 159–166.

Shinn-Cunningham BG, Best V, 2008. Selective attention in normal and impaired hearing. Trends Amplif. 12, 283–299. [PubMed: 18974202]

Song JH, Skoe E, Banai K, Kraus N, 2011. Perception of speech in noise: neural correlates. J. Cognit. Neurosci 23, 2268–2279. [PubMed: 20681749]

Strait DL, Kraus N, 2011. Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. Front. Psychol 2, 113 113. [PubMed: 21716636]

Strauss DJ, Francis AL, 2017. Toward a taxonomic model of attention in effortful listening. Cognit. Affect. Behav. Neurosci 17, 809–825. [PubMed: 28567568]

Taylor JSH, Rastle K, Davis MH, 2013. Can Cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. Psychol. Bull 139, 766–791. [PubMed: 23046391]

Teki S, Kumar S, Von Kriegstein K, Griffiths TD, Chait M, 2011. Brain bases for auditory stimulus-driven figure-ground segregation. J. Neurosci 31, 164–171. [PubMed: 21209201]

Tong Y, Chen Q, Nichols TE, Rasetti R, Callicott JH, Berman KF, Wein-berger DR, Mattay VS, 2016. Seeking optimal region-of-interest (ROI) single-value summary measures for fMRI studies in imaging genetics. PLoS One 11, 1–20.

Tune S, Alavash M, Fiedler L, Obleser J, 2020. Neural Attention Filters Do Not Predict Behavioral Success in a Large Cohort of Aging Listeners bioRxiv, 2020.2005.2020.105874.

Vaden KI Jr., Kuchinsky SE, Ahlstrom JB, Dubno JR, Eckert MA, 2015. Cortical activity predicts which older adults recognize speech in noise and when. J. Neurosci 35, 3929–3937. [PubMed: 25740521]

Vanthornhout J, Decruy L, Wouters J, Simon JZ, Francart T, 2018. Speech intelligibility predicted from neural entrainment of the speech envelope. JARO J. Assoc. Res. Otolaryngol 19, 181–191. [PubMed: 29464412]

Viswanathan V, Bharadwaj HM, Shinn-Cunningham BG, 2019. Electroencephalographic signatures of the neural representation of speech during selective attention. eNeuro 6.

Weber A, Scharenborg O, 2012. Models of spoken-word recognition Models of word recognition. WIREs Cogn. Sci 3, 387–401.

Wong PC, Jin JX, Gunasekera GM, Abel R, Lee ER, Dhar S, 2009. Aging and cortical mechanisms of speech perception in noise. Neuropsychologia 47, 693–703. [PubMed: 19124032]

Wong PC, Uppunda AK, Parrish TB, Dhar S, 2008. Cortical mechanisms of speech perception in noise. J. Speech Lang. Hear. Res 51, 1026–1041.

Yates KM, Moore DR, Amitay S, Barry JG, 2019. Sensitivity to melody, rhythm, and beat in supporting speech-in-noise perception in young adults. Ear Hear. 40, 358–367. [PubMed: 29965864]

Zekveld AA, Heslenfeld DJ, Festen JM, Schoonhoven R, 2006. Top-down and bottom-up processes in speech comprehension. Neuroimage 32, 1826–1836. [PubMed: 16781167]

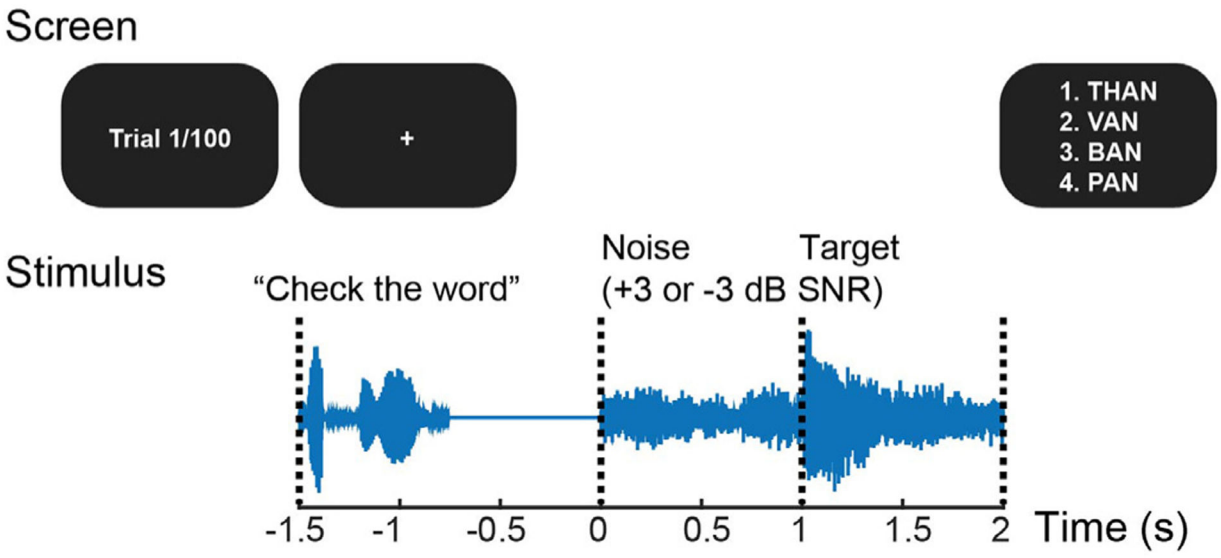

Fig. 1.

Trial and stimulus structure. Every trial starts with the cue phrase "check the word." A target word starts 1 second after the noise onset. Four choices are given after the word ends; subjects select the correct answer with a keypad. No feedback is given. The noise level is manipulated to create high (+ 3 dB) and low (−3 dB) SNR conditions. Subjects complete 50 trials for each condition.

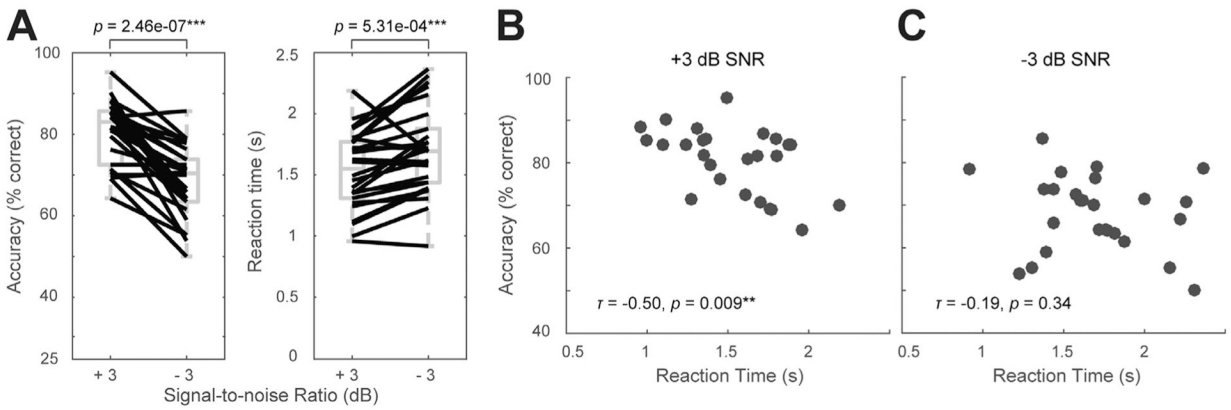

Fig. 2.

Behavioral results. **A.** Summary of behavioral performance for the two conditions (+ 3 and −3 dB SNR). Boxes denote the 25^{th} – 75^{th} percentile range; the horizontal bars in the center denote the median; the ranges are indicated by vertical dashed lines. Solid lines connect points for the same subject in different conditions. The bottom boundary line of the accuracy panel represents chance-level performance (i.e., 25%). **B.** Average accuracy as a function of reaction time in + 3 dB SNR condition. **C.** Average accuracy and reaction time in −3 dB SNR condition.

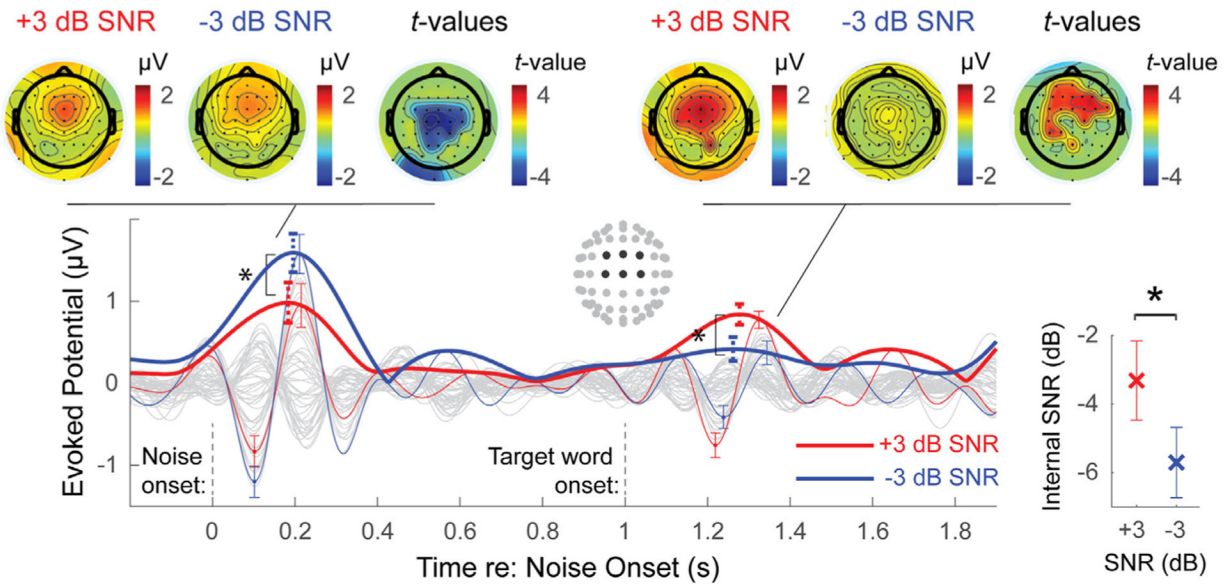

Fig. 3.

Sensor-level event-related potential (ERP) with the topographical layout and cortical maps. The time course of the auditory ERP and its envelope, with the standard error of the mean (± 1 SEM) at the peak amplitude (red color: + 3 dB SNR, blue color: −3 dB SNR). Thin colored lines are grand-average ERP waveforms averaged across front-central channels (FC1, FCz, FC2, C1, Cz, and C2: denoted as block dots in the electrode position legend on top of the waveforms). Thick colored lines are their temporal envelopes. Gray lines are grand-average waveforms at other electrodes. Asterisks represent significant differences in the amplitude between + 3 and −3 dB SNR conditions (paired *t*-test). Top panels show peak P2 amplitudes of all electrodes in topographical layouts. The *t*-values from paired *t*-tests between two SNR conditions are also shown as topographies.

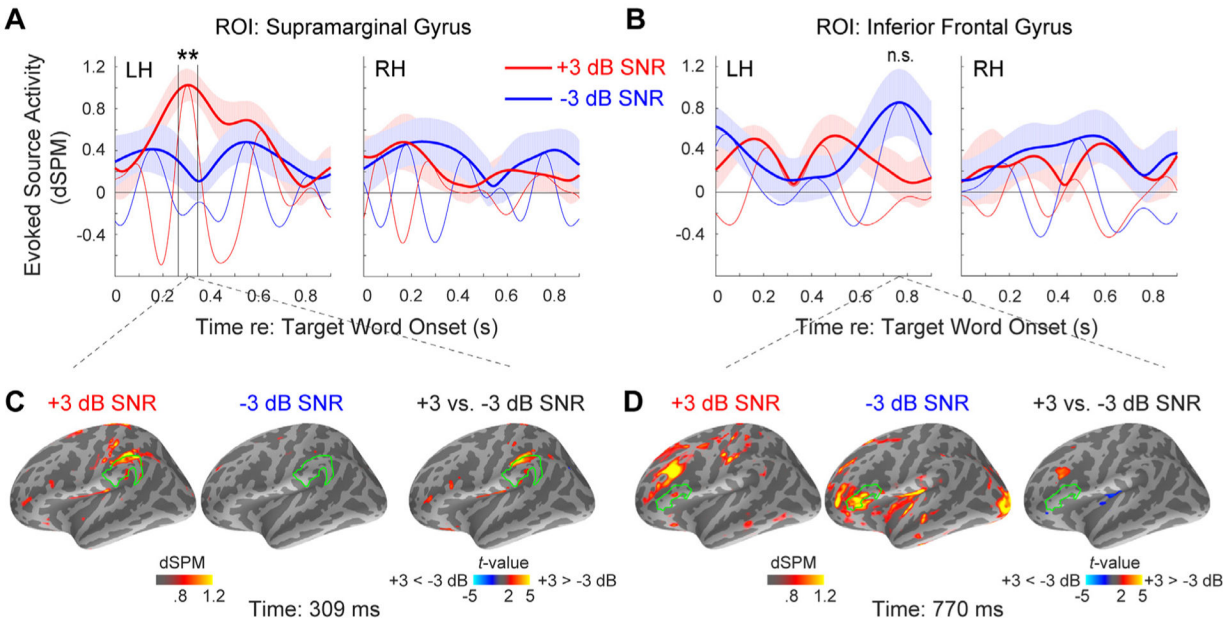

Fig. 4.

Region-of-interest (ROI) based source analysis. **A and B.** The time course of the evoked source activity and its envelope, with the standard error of the mean (± 1 SEM), obtained at representative voxels for two ROIs in the left hemisphere ("LH"): supramarginal gyrus (SMG), and the pars opercularis and triangularis of the inferior frontal gyrus (IFG), respectively, in each SNR condition (red color: + 3 dB SNR, blue color: −3 dB SNR). The time courses in the corresponding regions in the right hemisphere ("RH") are also shown for visual comparisons. Asterisks show the timing of a significant difference between + 3 and −3 dB SNR conditions (cluster-based permutation test, $p = 0.0020$) at the left SMG. **C and D.** Whole-brain maps showing statistical contrasts (t-values obtained from post-hoc paired t-tests between the two SNR conditions) of source activation at each voxel at the peak timepoint of the grand-average source time course of each ROI. Across all the panels, dSPM stands for dynamic statistical parametric maps, a noise-normalized source current strength in ratio (z-scores) (Dale et al., 2000).

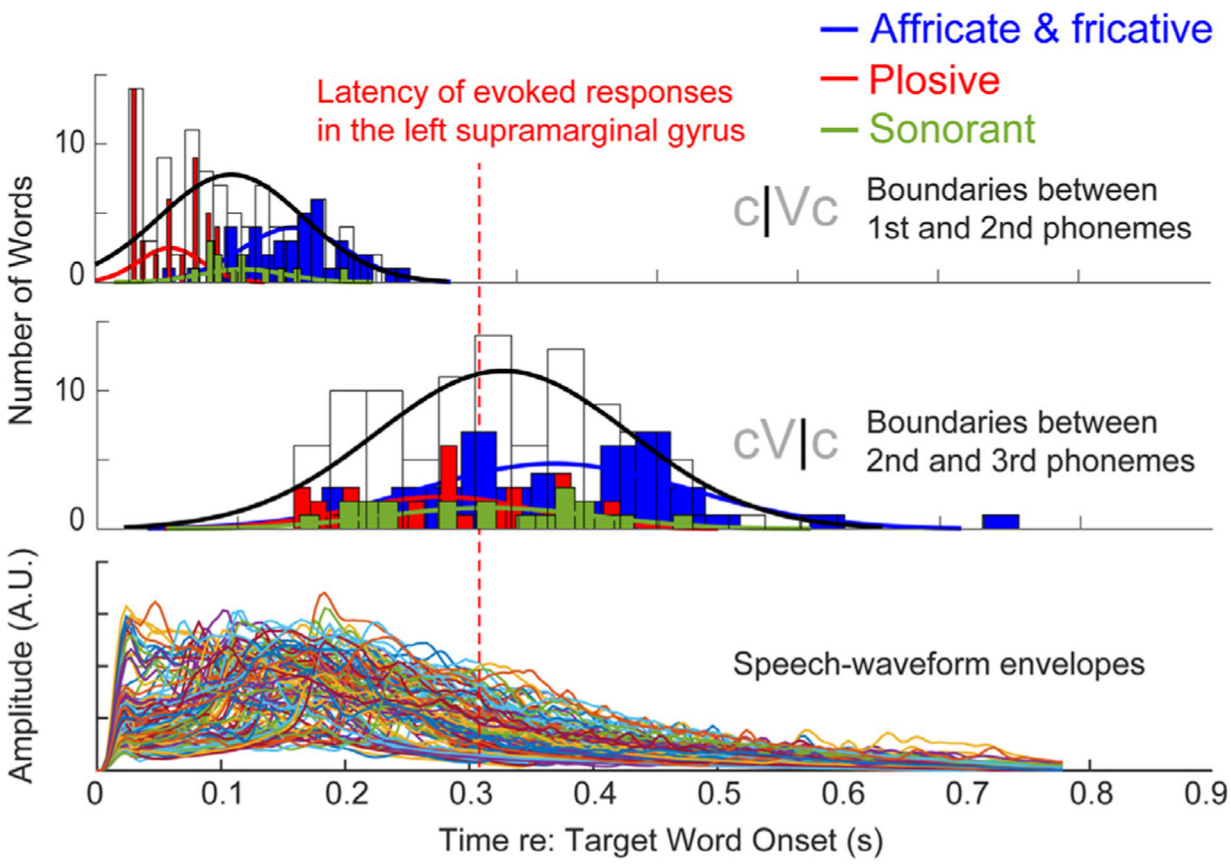

Fig. 5.

Top and second panel show histograms of boundaries between phonemes of each stimulus. The third panel shows superimposed temporal envelopes extracted from waveforms of the 100 words.

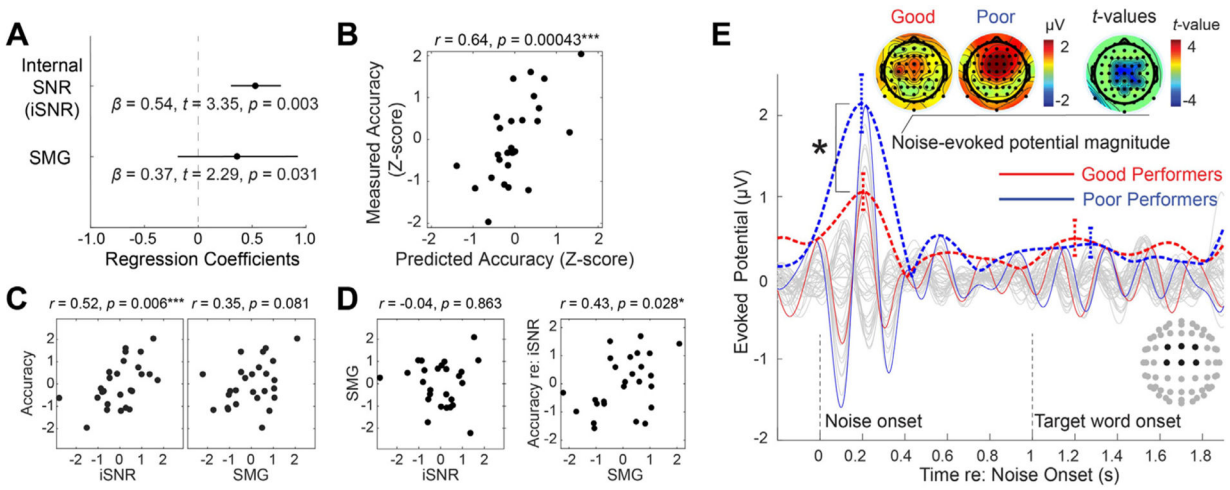

Fig. 6.

Individual differences in speech-in-noise processing. **A**. Regression coefficients and their standard errors. **B**. A scatter plot showing the relationship between predicted and measured accuracy in −3 dB SNR condition. **C**. Post-hoc correlation analyses: Raw correlations between each independent variable and the dependent variable. **D**. Left: Relationship between independent variables shows no correlation between internal SNR and evoked source current at the left supramarginal gyrus (SMG). Right: Semi-partial correlation between SMG evoked source current and the residual of accuracy after regressing out internal SNR. **E**. The time course of the auditory event-related potential and its envelope, with the standard error of the mean (± 1 SEM) at the peak magnitude in −3 dB SNR condition (red color: good performers, blue color: poor performers). An asterisk shows a significant difference in the magnitude between two groups (two-sample *t*-test).

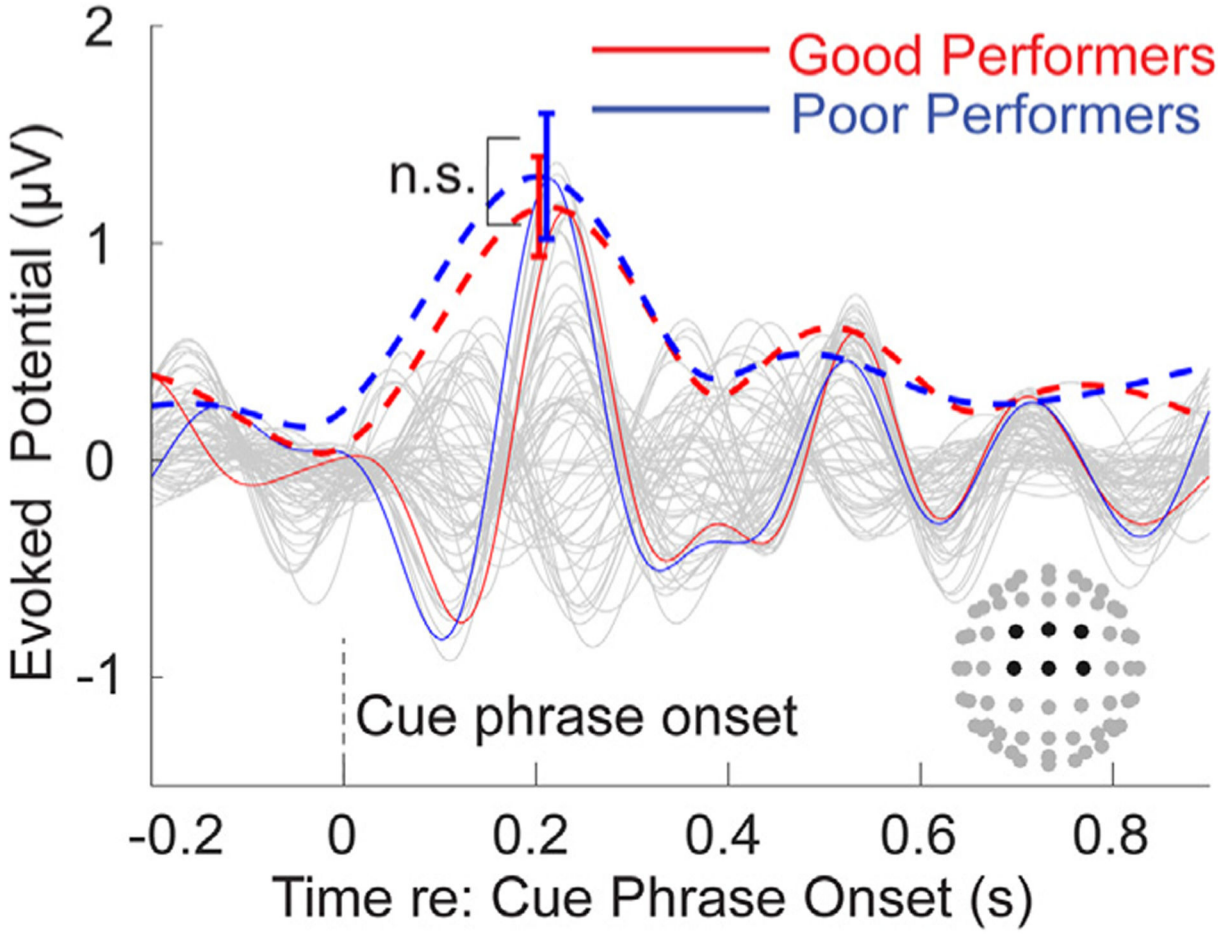

Fig. 7.

ERPs evoked by the carrier phrase “check the word” compared between listeners with good and poor performance in the low SNR condition. Colored lines represent ERPs averaged over front central electrodes. Gray lines represent ERPs at all other electrodes.