frontiers in Systems Neuroscience

ORIGINAL RESEARCH
published: 31 October 2018
doi: 10.3389/fnsys.2018.00056

Restoration and Efficiency of the Neural Processing of Continuous Speech Are Promoted by Prior Knowledge

Francisco Cervantes Constantino[1†] *and Jonathan Z. Simon*[1,2,3,4*]

[1] *Program in Neuroscience and Cognitive Science, University of Maryland, College Park, College Park, MD, United States,* [2] *Department of Electrical and Computer Engineering, University of Maryland, College Park, College Park, MD, United States,* [3] *Department of Biology, University of Maryland, College Park, College Park, MD, United States,* [4] *Institute for Systems Research, University of Maryland, College Park, College Park, MD, United States*

Sufficiently noisy listening conditions can completely mask the acoustic signal of significant parts of a sentence, and yet listeners may still report the perception of hearing the masked speech. This occurs even when the speech signal is removed entirely, if the gap is filled with stationary noise, a phenomenon known as perceptual restoration. At the neural level, however, it is unclear the extent to which the neural representation of missing extended speech sequences is similar to the dynamic neural representation of ordinary continuous speech. Using auditory magnetoencephalography (MEG), we show that stimulus reconstruction, a technique developed for use with neural representations of ordinary speech, works also for the missing speech segments replaced by noise, even when spanning several phonemes and words. The reconstruction fidelity of the missing speech, up to 25% of what would be attained if present, depends however on listeners' familiarity with the missing segment. This same familiarity also speeds up the most prominent stage of the cortical processing of ordinary speech by approximately 5 ms. Both effects disappear when listeners have no or little prior experience with the speech segment. The results are consistent with adaptive expectation mechanisms that consolidate detailed representations about speech sounds as identifiable factors assisting automatic restoration over ecologically relevant timescales.

Keywords: speech processing, auditory cortex, magnetoencephalography, stimulus reconstruction, speech envelope

OPEN ACCESS

Edited by:
Maria V. Sanchez-Vives, Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Spain

Reviewed by:
Emili Balaguer-Ballester, Bournemouth University, United Kingdom
Anton Filipchuk, Unité de Neuroscience, Information et Complexité (UNIC), France

****Correspondence:***
Jonathan Z. Simon
jzsimon@umd.edu

†Present address:
Francisco Cervantes Constantino, Centro de Investigación Básica en Psicología, Universidad de la República, Montevideo, Uruguay

Received: *22 January 2018*
Accepted: *09 October 2018*
Published: *31 October 2018*

Citation:
Cervantes Constantino F and Simon JZ (2018) Restoration and Efficiency of the Neural Processing of Continuous Speech Are Promoted by Prior Knowledge. Front. Syst. Neurosci. 12:56. doi: 10.3389/fnsys.2018.00056

INTRODUCTION

The ability to correctly interpret speech despite disruptions masking a conversation is a hallmark of communication (Cherry, 1953). In many cases, contextual knowledge poses an informational advantage for a listener, so as to successfully disengage the masker and restore the intended template signal (Shahin et al., 2009; Riecke et al., 2012; van Wassenhove and Schroeder, 2012; Leonard et al., 2016; Cervantes Constantino and Simon, 2017). Usually, relevant information is available from multimodal sources and/or low-level auditory and higher level linguistic analyses, although it remains unclear how and which factors are most effective in assisting speech restoration under natural conditions. Recently, cortical network activity profiles consistent with phonemic

restoration, the effect where absent phonemes in a signal may nonetheless be heard (Samuel, 1981, 1996), have been identified in binary semantic decision tasks (Leonard et al., 2016), yet factors that bias into one or the other of two perceptual alternatives remain unclear. At the algorithmic level, there is evidence that such restorative processes may be influenced by contributions from audiovisual integration cues (Crosse et al., 2016), lexical priming (Sohoglu et al., 2012), and within the auditory domain, by predictive template matching (SanMiguel et al., 2013). At the computational level, proposals include the deployment of intentional expectations about temporal patterns in sound (Nozaradan et al., 2011; Tal et al., 2017), and the use of mental imagery as a weak form of perception (Pearson et al., 2015).

In order to affect ongoing speech percepts, outcomes from such mechanisms would have to be readily accessible before and during missing auditory input. These type of contributions potentially entail (i) generation of a provisional template of the forthcoming speech, (ii) that the template be stored in a compatible format with the internal representation of ongoing sound, and (iii) that they are later subject to point-wise matching – in what has been termed the *zip metaphor* (Grimm and Schröger, 2007; Tavano et al., 2012; Bendixen et al., 2014). In addition, the contribution by such putative mechanisms in enhancing the neural representation of speech may allow a speed up of cortical processing during integration (van Wassenhove et al., 2005).

Here, we test how a string of natural speech tokens spanning several words may be represented cortically, even if entirely removed and replaced by stationary masking noise – under different levels of informational gain provided by prior knowledge of the masked elements. Prior research has shown that information from missing consonants can be inferred from cortical activity sustained over brief (∼100 ms) noise probes, by their similarity to responses to the original consonant sounds, such as a single fricative (Leonard et al., 2016). We use the fact that the low-frequency envelope of speech spanning several words indexes the acoustic signal's slow changes over time and is known to phase-lock neural activity in auditory cortex, as measured by magnetoencephalography (MEG) and electroencephalography (EEG) (Giraud et al., 2000; Ding and Simon, 2012b; Zion Golumbic et al., 2013; Di Liberto et al., 2015). Because of its timescale, the low-frequency envelope of speech typically reveals attributes such as the patterns of syllabic lengths and loudness changes, as well as prosodic information including intonation, rhythm, and stress cues. We hypothesize that by repeating the strings of speech tokens, and controlling for the extent of repetition, one can manipulate listeners' cortical ability to develop detailed predictions about forthcoming elements in these long sentences. More repetitions would allow the generation of more detailed templates for those tokens, to serve for a point-wise matching when later, spontaneous maskers disrupt the same string of tokens. If neurally instantiated, then the process may be investigated by testing how well the missing speech token can be decoded from the cortical signals representing it, despite the lack of related acoustic input. Furthermore, because the template would be formed prior, one may also investigate the possibility that cortical representations of highly repeated speech stimuli are facilitated by accelerated processing times.

To address these hypotheses, we employ a pair of complementary systems-based neural analysis methods. In one case, we analyze neural responses to reconstruct the stimulus speech envelope (Mesgarani, 2014), an approach that has been successfully applied in auditory electrophysiology (Mesgarani et al., 2009; Ramirez et al., 2011), EEG/MEG (Ding and Simon, 2012a; O'Sullivan et al., 2015), electrocorticography (Pasley et al., 2012; Leonard et al., 2016), and fMRI (Naselaris et al., 2011). In the case of speech restoration, electrophysiological responses have been used to re-create the acoustic representation of the stimulus using a data-driven decoder that effectively recovers the spectrogram of a missing consonant (i.e., substituted by noise) from listeners' cortical activity (Leonard et al., 2016). Complementary to this reconstruction analysis, temporal response function (TRF) analysis uses an acoustic representation of the stimulus to predict neural responses. This forward model permits direct analysis of cortical latencies involved in natural speech processing, the most prominent of which occurs between 100 and 180 ms, consistent with the latency of the evoked response M100 component (Cervantes Constantino et al., 2017). We investigated the possibility of related adaptations, such as reduced cortical latencies, under the same prior knowledge conditions employed in the decoding analyses, since faster processing has been observed in situations where additional context facilitates perceptual integration of incoming speech (van Wassenhove et al., 2005; van Wassenhove and Schroeder, 2012). In addition, similar task-related cortical plasticity changes in stimulus-response mappings are often observed at the neuronal level (Fritz et al., 2003; David et al., 2012) and would represent a potential biophysical basis for restorative mechanisms given the present task demands. The forward model applied to the MEG was then used to address whether and how similar adaptations might be reflected at the whole brain level.

We provide evidence that the speech temporal envelope is better reconstructed when listeners have obtained more knowledge about a particular speech sequence, and, critically, that this effect applies even in the case where the speech itself is absent, having been replaced entirely with noise. The data also show that cortical latencies in the processing of clean speech can be reduced by several milliseconds when the listener has obtained detailed knowledge about that particular speech sequence. Overall, the results suggest improved efficiency in accessing dynamic neural representations of low-level features of frequently experienced speech, indicated by both faster stimulus encoding and endogenous restorative processes that reflect a neural representation of the missing speech.

MATERIALS AND METHODS

Participants

Thirty-five experimental subjects (19 women, 21.3 ± 2.9 years of age [mean ± SD]), with no history of neurological disorder or metal implants, participated in the study. Data from one

additional subject was not included, due to excessive artifacts caused by a poor fit with the MEG helmet. Each subject received monetary compensation proportional to the study duration (approximately 1.5 h). This study was carried out in accordance with the recommendations of the UMCP Institutional Review Board with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the UMCP Institutional Review Board.

Stimuli and Experimental Design

Sound stimuli were prepared with the MATLAB® software package (MathWorks, Natick, MA, United States) at a sampling rate of 22.05 kHz, and consisted of a recorded poem ("A Visit from St. Nicholas," Moore or Livingston, 1823) obtained from an online archive[1]. In addition to the narrated contents, the spoken verses transmit intonation, rhythm, and stress cues, all amenable for encoding as prosodic information units, and all predisposed to potential cortical restoration extending over multiple syllables. Each of the 14 verses (each verse being a quatrain of four lines) in the poem were separated and considered as individual stimuli. In order to probe the contribution of prior experience to cortical coding, the four stimulus blocks presented to each subject, each containing 64 stimuli (i.e., 256 lines), had some of stimuli repeated multiple times, as follows. For the first block, a verse/stimulus from the first half of the poem was chosen to be a "High" frequency stimulus, with sufficient repetitions to make up half of the block's stimulus presentations (32/64). Similarly, other verses were chosen as "Medium" and "Low" frequency stimuli, which were repeated for a quarter (16/64) and an eighth (8/64) of the block's stimulus presentations, respectively. The remainder of the block was filled with "Control" stimuli, namely the four remaining verses presented either one, two, or four times within the block. This category represents, for missing speech, the case for which the listener would have insufficient prior experience with the specific speech segment to promote restoration; for non-missing speech, forward model analysis of this case acts as a control for comparing latencies of more frequently presented stimuli.

Silent intervals (gaps) in the narration were reduced to approximately equalize stimuli durations (range: 13.1–13.6 s). Stimuli were randomized in order and concatenated in time. For the second block, the same procedure was followed using material from the second half of the poem. Blocks 3 and 4 consisted of the same stimuli used as in 1 and 2, respectively, but with a different randomized order and different placement of noise probes (see below). The procedure was recreated with different randomizations for each subject, resulting in a total of 35 different stimulus sets of about 1 h each in total duration. Importantly though, the usage of particular stimuli at a given repetition level was controlled across participants, resulting in seven groups of five listeners each that underwent the same "High," "Medium," "Low," and "Control" stimuli selection.

[1]https://archive.org/details/AVisitFromSt.Nicholas-ByClementClarkeMoore-NarratedByGrantRaymond

For each stimuli, two to four spectrally matched noise probes of 800 ms duration each were applied at pseudo-random times with a minimum 2.5 s between probe onsets. Noise onset times were selected from a pool of values indicating articulation onset times (e.g., syllables), obtained as the envelope rising slope maxima. Thus, 768 noise probe samples were presented per experiment, and each was individually constructed by randomizing phase values across the specific frequency-domain phase information contained in the underlying speech stimulus that would have occurred at the same time as the masker noise, yielding a noise with equal spectral amplitude characteristics (Prichard and Theiler, 1994). The original speech content occurring during the same time was removed entirely and substituted with this spectrally matched noise, at a power signal level matching that of the excised clean original. Subjects listened to the speech sounds while watching a silent film. To ensure attention to the auditory stimulus, after each probe, they were instructed to report via a button press whether they understood what the speaker meant to say during the noise. The button presses are not analyzed here.

Data Recording

We recorded neural responses using MEG, a non-invasive neuroimaging technique well-suited to measure dynamical neural activity from human cortex, and especially from auditory cortical areas. Such recordings typically demonstrate time-locked neural responses to speech low frequency modulations, especially of the acoustic energy envelope, with remarkable temporal fidelity (Ding and Simon, 2012b). MEG data were collected with a 160-channel system (Kanazawa Technology Institute, Kanazawa, Japan) inside a magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany). Sensors (15.5 mm diameter) were uniformly distributed inside a liquid He dewar, spaced ∼25 mm apart. Sensors were configured as first-order axial gradiometers with 50 mm separation and sensitivity >5 fT.Hz$^{-1/2}$ in the white noise region (>1 kHz). Three of the 160 sensors were magnetometers employed as environment reference channels. A 1-Hz high-pass filter, 200-Hz low-pass filter, and 60-Hz notch filter were applied before sampling at 1 kHz. Participants lay supine inside the magnetically shielded room under soft lighting, and were asked to minimize movement, particularly of the head.

Data Processing

Pre-processing and Sensor Rejection

The time series of raw recordings from the MEG sensor array were be submitted to a fast implementation of independent component analysis (Hyvärinen, 1999), from which two independent components were selected for their maximal proportion of broadband (0–500 Hz) power (because of the $\sim 1/f$ power spectrum of typical neural MEG signals, these components are dominated by non-neural artifacts). These independent components, combined with the physical reference channels, were treated as environmental noise sources arising from unwanted electrical signals not related to brain activity of interest, and were removed using time-shifted principal component analysis (TS-PCA) (de Cheveigné and Simon, 2007).

Sensor-specific sources of signals unrelated to brain activity were reduced by sensor noise suppression (SNS) (de Cheveigné and Simon, 2008b).

Data Analysis

To analyze low-frequency cortical activity, recordings were band-pass filtered between 1 and 8 Hz with an order-2 Butterworth filter, with correction for the group delay. A blind source separation technique, denoising source separation (DSS) (de Cheveigné and Simon, 2008a), was used to construct components (virtual channels constructed of linear combinations of the sensor channels), ranked in order of their trial-to-trial reproducibility, trained *only from clean speech presentations*, and used as described below.

Stimulus Reconstruction

The ability to reconstruct the speech stimulus envelope from recorded neural responses was used to measure the dynamical cortical representation of perceived speech. Decoders were separately estimated based on either reproducible or "reference" activity as ranked by DSS, with two such pairs of decoders: the first pair trained on responses to speech and the second pair trained on responses to noise, as described in the following. The first decoder of each pair was based upon the first three DSS components (i.e., with highest reproducibility across instances of listening to *clean speech*). These highly reproducible components were used to train an optimal linear decoder designed to reconstruct the envelope of the stimulus that was presented under normal speech listening conditions but absent (though perhaps expected) under noise listening conditions. That the reproducible neural activity is generated by auditory cortex is reflected in a DSS component topography consistent with auditory responses arising from temporal cortices (**Supplementary Figure S1**). The last three DSS components (with the lowest reproducibility from the same clean speech dataset), were similarly used to train the second linear decoder for each pair, used as a reference, i.e., to estimate baseline performance. Each decoding procedure produced a corresponding reconstructed stimulus time series whose similarity with the corresponding speech envelope was assessed via Pearson's r correlation coefficient. Each similarity score was, respectively, designated as reproducible (r_e) and reference (r_f). This referencing procedure is necessary to obtain a baseline in decoding performance since time series' lengths varied across conditions (as a result of the different repetition rates and verses involved); otherwise, there would be positive biases in r for shorter sequences, irrespective of underlying relationship to the stimulus. The appropriate pairs of decoders were applied separately to their respective neural responses, to clean speech and to noise. Noise edge and button-press-related segments were excluded from all analysis. To the extent that a noise-only response can be used to reconstruct an absent but expected stimulus (better than baseline performance) reflects the presence of neural activity consistent with a representation of the acoustically absent speech.

To compute reconstruction effect sizes, each of the Pearson's r pairs (reproducible versus reference activity) were transformed to Cohen's effect size q (Cohen, 1988) by the transform $q = \frac{1}{2}\left(\ln\frac{1+r_e}{1-r_e} - \ln\frac{1+r_f}{1-r_f}\right)$ for both kinds of responses. Relative effect sizes (speech versus noise reconstruction) were computed by the fraction q_2/q_1 of reconstruction effect sizes given the stimulus presentation conditions above (expressed as percentages), where q_1 denotes the effect size obtained from reconstructions of clean speech from neural activity following clean speech, and q_2 the effect size from reconstructions of clean speech from neural activity arising from the noise probe (devoid of speech). Absolute effect sizes during noise presentations were used for statistical analysis.

Temporal Response Function of Stimulus Representation

The input–output relation between a representation $S(t)$ of auditory stimulus input and the evoked cortical response $r'(t)$ is modeled by a TRF. This linear model is formulated as:

$$r'_{\text{pred}}(t) = \sum_{\tau} TRF(\tau)S(t-\tau) + \varepsilon(t)$$

where $\varepsilon(t)$ is the residual contribution to the evoked response not explained by the linear model. As stimulus representation, the envelope was extracted by taking the instantaneous amplitude of each channel's analytic representation via the Hilbert transform (Bendat and Piersol, 2010), with sampling rates reduced to 1 kHz, transformed to dB scale. The response was chosen to be either the first or second DSS component (fixed for each subject; **Supplementary Figure S1**), according to which one produced a TRF with a more prominent $M100_{TRF}$, a strong negative peak with ∼100 ms latency (Ding and Simon, 2012a).

Statistical Analyses

For reconstructions, one-way repeated measures ANOVA were run across the four levels: "Control," and "Low," "Medium," and "High" repetitions, in order to examine differences between their related means overall. Cortical latency of the TRF was determined by the $M100_{TRF}$ latency. Peak delays with respect to control conditions were determined by cross-correlations of the TRF in the "Control" versus all other repetition conditions. The resulting peak delays were then submitted to a non-parametric one-tailed two-sample Kolmogorov–Smirnov test for differences in the underlying delay populations.

RESULTS

Neural Transformations Facilitated by Prior Experience Result in Improved Ability to Reconstruct Missing Speech From Noise

Connected syllables/words within a narrated poem were replaced by noise bursts of fixed duration. Each static noise probe was constructed with the same spectrum as the replaced speech segment (**Figure 1A**), and therefore lacked critical temporal modulations, e.g., in the low-frequency (2–8 Hz) envelope. Low-frequency fluctuations present in natural, unmasked speech typically generate time-locked

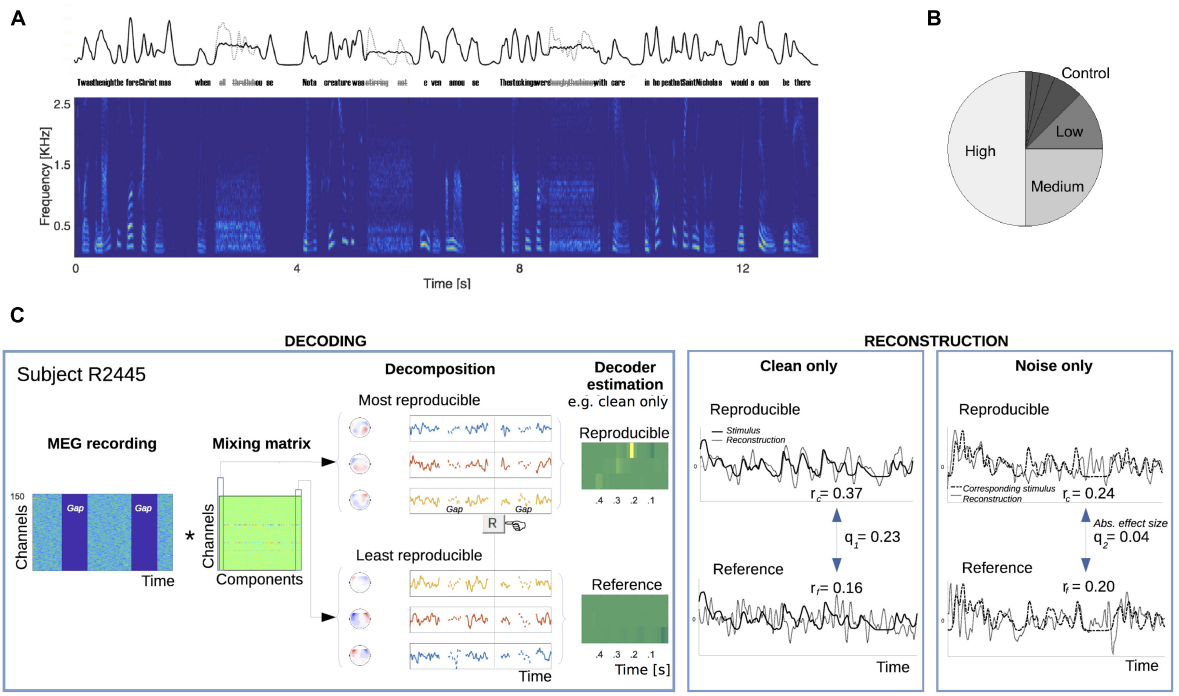

FIGURE 1 | Present/absent speech stimulus protocol and cortical speech envelope reconstruction algorithm procedure. **(A)** Speech material from a single poem was continuously presented to 35 listeners over the course of an hour, except that every 4–5 s, some speech was replaced with spectrally matched noise (0.8 s duration; three replacements shown). This manipulation removes all critical temporal modulation for that duration. Top: example of the slow envelope of continuous speech with occasional missing speech replaced by stationary noise in black; the envelope of the removed missing speech is in gray. Bottom: spectrogram of the continuous speech with missing speech replaced by stationary noise. **(B)** For any participant, a verse selection was chosen to be repeated more than any other (50% of all verse presentations) and denoted as having a "high" repetition rate. Another verse selection was similarly chosen to be repeated with a "medium" repetition rate (25% of all verse presentations). Additional verses were repeated less frequently: 12.5% ("low"), 6.25%, etc. Verses presented 6.25% or less often were grouped together as "Control." The frequency designation of verses was counterbalanced across subjects. **(C)** The stimulus protocol served to allow analysis of reconstruction performance for both types of response, to clean speech and to stationary noise without speech. In the *decoding* step, a subject's multichannel response was decomposed into components, ordered by trial-to-trial reproducibility during speech listening (see Section "Materials and Methods"). The top three waveform components (with corresponding topographies shown) served to train a *reproducible* neural activity-based decoder to estimate the clean speech low-frequency envelope. Correspondingly, the bottom three components trained a *reference* (baseline) decoder. Separate pairs were trained according to whether the speech was present (as shown in this panel) or absent and replaced by noise (not shown). In the *reconstruction* step, the appropriate pair of reproducible and reference decoders were applied to the neural responses to speech or noise, respectively. Results of the envelope reconstruction algorithm are shown for a representative subject.

auditory cortical activity recorded by MEG (Giraud et al., 2000; Ding and Simon, 2012b). Multichannel recordings thus contain information about dynamic neural representations of speech, and so may in turn serve as the basis to train decoding models that reconstruct dynamic features of the presented speech signal, e.g., its envelope. Linear decoders mapping from MEG responses to the stimulus speech envelope were estimated per subject, and their envelope reconstruction performance (in estimating the original speech signal) quantified. Crucially, decoders were either trained from recordings of clean speech presentation intervals, or, separately trained from intervals when only static noise was delivered. In all cases, performance was tested with respect to the corresponding clean speech signal at that part of the poem, whether actually delivered or not. To test whether acoustic presence is a necessary condition for reconstruction of continuous speech, listeners were exposed to extensive repetitions of some verses (each verse being a quatrain of four lines), and less frequent repetitions (or none at all) to the rest (**Figure 1B**). To limit confounding effects from specific properties of particular verses, counterbalanced subgroups heard different sets of verses for each repetition frequency, i.e., the five participants in each subgroup experienced the same verses with identical repetition frequencies, but the next subgroup heard a different set of verses for each repetition frequency.

The hypothesis was tested using an index of speech reconstruction, Cohen's *q*, estimated using a two-step process (**Figure 1C**). For each subject, first, a data-driven response mixing matrix was obtained from responses to clean speech only, a procedure that serves to decompose multichannel timeseries into their most and least reproducible components relevant to natural speech processing. These most and least reproducible sets served to generate reconstruction models of the original speech signal, estimating optimal and baseline, respectively, reconstruction performance levels in the subsequent stage. Second, one pair of reproducible and reference decoders was trained from recording intervals where speech was delivered, and separately, another pair was similarly trained from intervals presenting only noise. In all cases, the reconstruction models targeted the envelope of

the normal speech stimulus, even when that speech was absent (but perhaps expected) under the noise listening conditions. In the final step of the reconstruction stage, the q index was computed from comparing the appropriate pair of reproducible versus reference decoding performances. A separate q index was computed for responses to speech and to noise, and for each repetition frequency condition.

Differences in the time-locked auditory cortical responses to clean speech and to noise intervals, by repetition frequency of the verse they belonged to, are illustrated in **Figure 2A** (for two different subject subgroups that experienced the same verses at two repetition frequencies). Over the span of several seconds, the neural waveforms in response to (aggregated) noise intervals appear more comparable to those in response to clean speech when the verse in question was frequently repeated in a subgroup (High; upper graph) than not (Control; middle graph). This is demonstrated by the timeseries' correlation coefficients between waveform pairs across the entire verse (**Figure 2B**). Because the repetition frequency of particular verses is counterbalanced across subjects, the analogous comparison for High versus Low repetition conditions (**Supplementary Figure S2**) uses different subgroups.

In terms of reconstruction performance, sentences that were maximally repeated (High repetition rate) over the hour-long session resulted in greatest relative performance in reconstruction of the envelope of the missing speech: approximately 25% of the performance for actual speech presented without any masking. Less exposure resulted in further reductions in relative performance (Medium: 21%, Low: 9%, and Control: 8%, respectively), down to the floor level in the case of masked speech with which the listener had little or no prior experience (**Figure 2C**; percentages inset). Because this measure is relative to clean speech reconstruction, a measure of reconstruction from noise alone was tested separately, using Cohen's q to quantify the effect size. Effect sizes in reconstruction of the missing speech envelope were confirmed to display a similar pattern as with relative performance (High: 0.079 ± 0.013; Medium: 0.060 ± 0.011; Low: 0.020 ± 0.013;

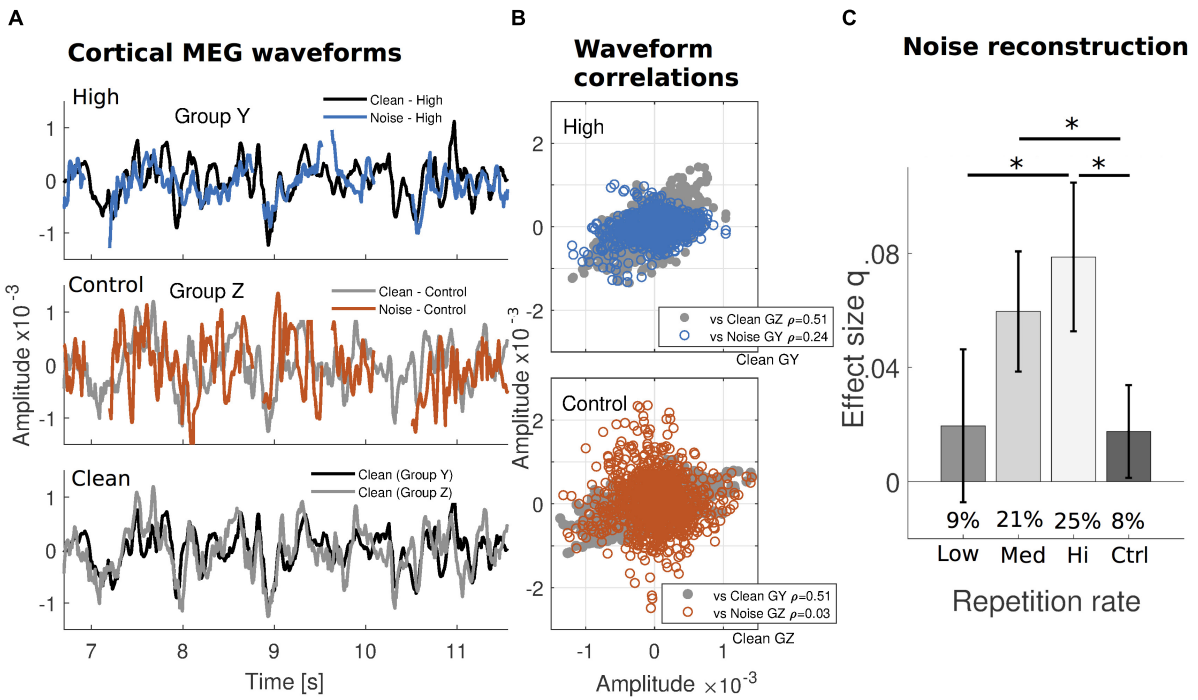

FIGURE 2 | Increased robustness of cortical reconstructability to missing speech under substantial prior exposure. **(A)** The median cortical MEG response waveforms from one cross-validation subject subgroup (group Y; N_Y = 5) exposed to a specific verse with High frequency, are contrasted with those from a different subgroup (group Z; N_Z = 5) exposed to the same verse but infrequently (Control). Within the High repetition rate subgroup (top), signals generated by clean speech (black) and static noise replacing the missing speech (blue) appear to covary more than for the Control repetition rate subgroup (middle: gray and red, respectively). The waveforms in response to clean speech from both subgroups are replotted together for comparison (bottom). (Gaps in the responses to noise are due to intervals when the speech was never replaced by noise, or during noise onset or button presses.) **(B)** Correlation coefficients indicate that, while clean speech responses are similar to each other for both subgroups (gray, same data in both plots), for noise intervals, they are much more similar to the clean responses only when the verse is frequently repeated (blue, top), but not when infrequently repeated (red, bottom). For a comparison from other subgroups using High and Low repetition rate conditions, see **Supplementary Figure S2**. Median waveforms use the dominant auditory response component (see **Supplementary Figure S1**) and are determined for each time point over the subject group. **(C)** The missing dynamic speech envelope may be reconstructed from responses to noise, with performance of about 25% of that obtained under clean conditions (percentages inset for each bar). As indicated by the absolute q index dependence on presentation frequency using the noise-trained decoders, this result is not a consequence of any clean-trained decoders dependence on frequency. Error bars indicate confidence intervals for the means (Bonferroni-corrected α level).

Control: 0.018 ± 0.008) (**Figure 2C**). A one-way repeated measures ANOVA with four repetition levels was performed to determine whether decoding success of the linear model of the envelope significantly changed across conditions. q index results for reconstructions exclusively using noise intervals showed that the sphericity condition was not violated [Mauchly test, $x^2(5) = 6.322$; $p = 0.276$]. The subsequent ANOVA resulted in a significant main effect of repetition frequency [$F(3,102) = 8.070$; $p = 7.1 \times 10^{-5}$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed that this increased exposure to speech significantly improved the stimulus reconstruction effect size from Low and Control repetition rate conditions to High ($p = 2.5 \times 10^{-3}$ and $p = 7.7 \times 10^{-4}$, respectively), and also from Control to Medium ($p = 7.7 \times 10^{-3}$).

Expedited Auditory Cortical Processing of Natural Speech

The TRF is a linear model used to predict the dynamics of the neural response to sound input, given a representation of the stimulus such as the acoustic envelope. Its characteristic peaks, and especially their polarity and latencies, are indicative of the progression of neural processing stages akin to the distinct generators of evoked responses to simple sounds such as pure tones (Ding and Simon, 2012a,b; Cervantes Constantino et al., 2017), but with the advantage of being directly derived from the neural processing of continuous natural speech. We examined the effect of prior exposure on the TRF's temporal structure in general, and also for the most prominent peak, the $M100_{TRF}$, occurring 100–180 ms post envelope change (**Figure 3A**). When a given speech sequence was listened to repeatedly, a significant within-participant latency shift of 5.3 ± 2.2 ms earlier was observed for $M100_{TRF}^{High}$ versus $M100_{TRF}^{Control}$ peaks [$t(33) = 2.387$; $p = 0.023$], indicating expedited cortical processing cortical for more familiar stimuli (**Figure 3B**). Across participants, the differences between repeated (High, Medium, and Low) and baseline (Control) levels, in terms of maxima in their cross-correlation functions, were shown to arise from significantly different distributions ($D = 0.294$; $p = 0.043$) (**Figures 3C,D**), suggesting that prior experience by repeated presentations effectively speeds up cortical processing even as early as 100 ms latency.

DISCUSSION

We present evidence of dynamic envelope coding of missing natural speech, by means of stimulus reconstruction methods applied to auditory cortical responses. This occurs as long as when there has been a history of repeated, frequent exposure to the original missing speech, suggesting that prior experience facilitates access and maintenance of a detailed temporal representation of the stimulus even though absent as low-level input. In addition, we find that cortical processing dynamics timescales are reduced by about 5 ms under similar prior experience conditions, for natural speech processing.

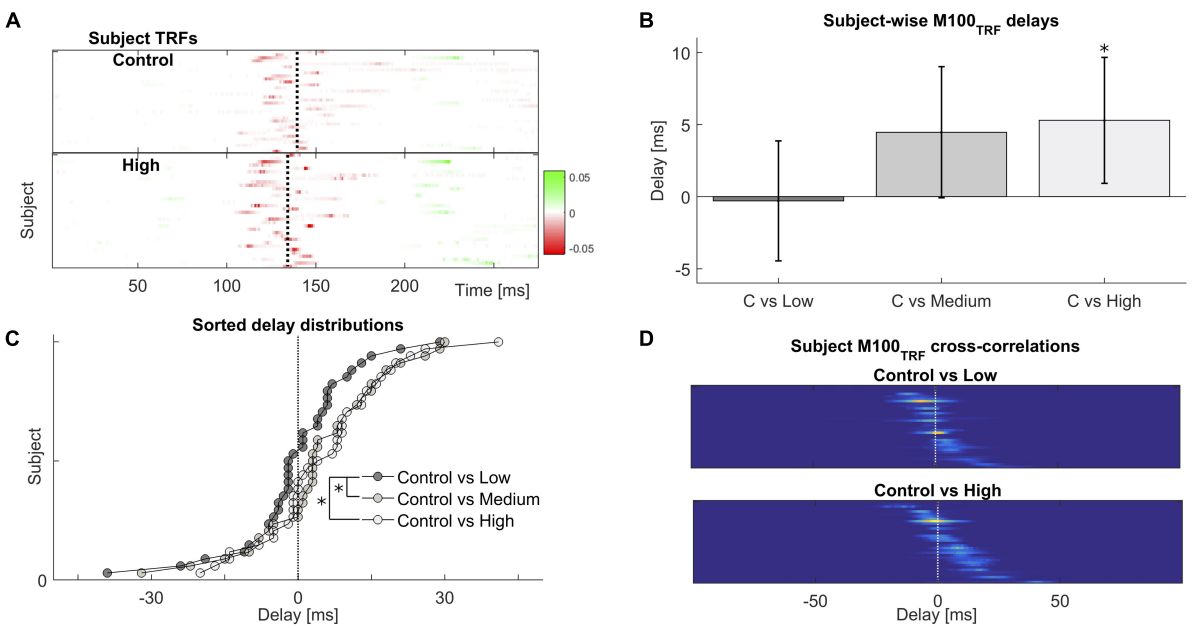

FIGURE 3 | Frequent repetitions of natural speech speed-up their cortical processing. **(A)** Temporal response functions across participants reveal a common cortical processing step, referred to as the $M100_{TRF}$, typically occurring about 100 ms after a speech envelope fluctuation (red colored features near the vertical dotted lines). **(B)** Depending on familiarity with the speech tokens, the same processing step may shift in time: processing of frequently repeated speech occurs about 5 ms earlier than for novel or sparsely presented sentences, within subjects. **(C)** Across subjects, the distribution of relative delays is consistently biased towards positive (earlier) values for the most extreme repetition conditions. **(D)** Illustration of how shifts within subjects were obtained, by cross-correlating individual $M100_{TRF}$ peak profiles obtained per condition in each subject.

Spectrogram reconstructability of noise-replaced phonemes (e.g., fricatives) has been demonstrated when subjects interpret the remainder of the single word accordingly (Leonard et al., 2016). Such endogenous activity may arise from top-down modulations of auditory cortical areas (Petkov et al., 2007; Petkov and Sutter, 2011) with the effect of modulating perceptual processing, including, the ability to entrain to speech signals (Ding et al., 2013), to optimize detection performance (Henry and Obleser, 2012), and to support auditory illusions (Riecke et al., 2009). Under the umbrella of *attractive temporal context effects* (Snyder et al., 2015), a group of facilitatory mechanisms including perceptual hysteresis and stabilization (cf. Kleinschmidt et al., 2002; Pearson and Brascamp, 2008; Schwiedrzik et al., 2014), auditory restoration effects improve perceptual invariance in the face of discontinuously fluctuating, broadly cluttered environments. The involvement of storage-based reactivation in perceptual processes, including attention, is an area of active research (Backer and Alain, 2012, 2014; Zimmermann et al., 2016). We therefore provide evidence for reactivation mechanisms based on prior learning and storage of speech information, at the level of its temporal structure.

Access and Format of Stored Auditory Representations: Hierarchical Models

Encoding of speech and other stimuli into sensory memory, the function of primary sensory areas that integrates analysis and storage of stimulus features by relevance (Cowan, 1984; Weinberger, 2004), has been argued to assist in the ability to restore missing fragments of a sound source, e.g., as an internal replay of the fragment during phonemic restoration (Shinn-Cunningham, 2008). Examples of implicit auditory memory in sensory and perceptual encoding (Snyder and Gregg, 2011) are observed in repetition-based improved detection of arbitrary noise constructs, and on neural covariates of this improvement (Agus et al., 2010; Andrillon et al., 2015).

Foreknowledge of acoustic features allows adaptation to a likely communication source, as shown by, e.g., facilitation with advance notice regarding the identity of a forthcoming instrument (Crowder, 1989), and by preferential activation in auditory association areas specific to speaker familiarity (Birkett et al., 2007). Such differential activation, given variable rates of sensory update, suggests that prior experience history of a dynamic sound pattern may influence its later representation: with few initial updates, storage at short intervals is associated with posterior superior temporal cortex, but over time, activation may take place at inferior frontal cortex (Buchsbaum et al., 2011). This progression is consistent, in memory terms, with readout from sensory buffers taking place at high temporal resolution under low-level representation formats; coarser temporal resolutions are instead attained at stores that operate under categorical, higher order feature codes (cf. Durlach and Braida, 1969; Winkler and Cowan, 2005). For perception, progression hierarchies are core features of models such as reverse hierarchy theory, which proposes that fast perception (e.g., when understanding speech in noise) is by default based on high-level cortical representations (Ahissar et al., 2009), except for specific conditions as systematic stimulus repeats, where information about fine temporal detail may then also be utilized (Nahum et al., 2008). Hierarchical models can be useful inasmuch they identify stages by which feed-forward general stimulus template extraction steps are completed, and they specify roles for feedback activity from higher areas (Kumar et al., 2007). In hearing their application includes, e.g., pitch and spectral envelope analyses, where top-down information serves to adapt effective processing time constants over lower areas that encode more temporally refined information (Kumar et al., 2007; Balaguer-Ballester et al., 2009).

Thus, a general prospect of these models is to determine the extent to which the natural hierarchy in sensory input might map to the anatomical hierarchy of the brain. In temporal terms, another application refers to an interesting distinction between "percept" versus "concept" representations of an environmental variable, namely, transient versus enduring representations (Kiebel et al., 2008). Because only the enduring (e.g., >1 s) representations have the capacity to shape how lower level representations may evolve, in the language of dynamical systems, they are seen as control parameters: consolidation of a "concept" automatically constrains where the trajectories of representations at subordinate processing levels may unfold autonomously (Kiebel et al., 2008). The role of prior knowledge in the cortical hierarchy of speech perception and representation, in particular with regard to the acoustic envelope, is a matter of current research interest (Sohoglu and Davis, 2016; Di Liberto et al., 2018). Hierarchical approaches therefore appear as a suitable framework to bridge findings of low-level endogenous representations with a mechanistic account of phonemic and speech restoration.

Phonological Structure and the Role of Auditory Retrieval Processes in Noise Listening

A tenet of speech restoration phenomena posits the use of prior abstractions or "schemata" that remain represented online, e.g., when an expected stimulus fails to occur, and are better resolved with increased familiarity (Hubbard, 2010). From the operational perspective, these are based upon the phonological structure of natural speech: representations first involve recognition of phonological structure (what is being heard), and second, that words be stored into verbal working memory using phonological code stores (Wagner and Torgesen, 1987). For efficiency reasons, the codes for lexicon storage and for the later retrieval probing process itself may be both the same, with phonologically similar words grouped together in the lexicon in "neighborhoods," as indicated by "phonological awareness" models from the developmental literature (e.g., Nittrouer, 2002; Lewis et al., 2010). In our results, this framework would indicate a hypothetical process for words replaced by noise probes where (1) each high-frequency word has been established as a competitive representative of its respective neighborhood store, (2) retrieval is however constrained to operate endogenously, primarily from available contextual information, and (3) the tempo of retrieval would be coordinated at a timescale superordinate to that of

phonological units, e.g., by prosodic and sentential information. In this scenario, the temporal envelope of missing speech, learned by prior experience, may coordinate the probe-triggering process.

Besides phonological structure, identification of auditory persistence processes unprompted by sensory input (Intons-Peterson, 2014) may involve self-directed imagery tasks (Bailes, 2007; Meyer et al., 2007), where activation levels in the planum temporale correlate with self-reported levels of imagery engagement and perceived vividness (Zatorre et al., 2009); both auditory imagery and rehearsal may be subserved by auditory association cortex areas in general (Meyer et al., 2007; Hubbard, 2010; Martin et al., 2014). Furthermore, in a bimodal stimulation study, transient activity from superior temporal cortex was shown to be critical at the beginning of the auditory retrieval process, but sustained planum temporale activity was involved overall (Buchsbaum et al., 2005). This is consistent with the interpretation that in such retrieval processes, task-relevant stimulus features may be maintained at (re)activated domains within the sensory representational space (Kaiser, 2015). The notion that both representation and maintenance involve overlapping processes (Hubbard, 2010) is supported by findings of reactivation, at retrieval, of sensory regions active during perception (Wheeler et al., 2000), and with auditory verbal imagery (McGuire et al., 1996; Shergill et al., 2001). The emerging, increasingly multimodal field of neural representations sustained during mental imagery may further support crucial clinical applications (Pearson et al., 2015).

Adaptive Dynamics of Speech Encoding and Representation During Masking

As related to speech, two main mechanisms indicate a correspondence of results with generative neural models (Pouget et al., 2013). First, the finding that cortical processing is sped up, under the same circumstances that promote restoration of speech-related neural activity, suggests that active, task-related endogenous processes directly optimize low-level speech processing with experience. A plausible mechanism for this is in promoting increased excitability of higher level neural populations. Second, our results indirectly support the suggestion that auditory "image" formation may entail activity consistent with that elicited by the original sound input (Janata, 2001; Martin et al., 2017), and whose temporal precision, and other related feature properties, may vary depending on factors such as context and experience (Janata and Paroo, 2006). The effect of frequent "refreshing" seen here may relate to the auditory memory reactivation hypothesis (Winkler and Cowan, 2005), where individual sound features can be effectively stored, along with neighboring sound patterns and sequences, when represented altogether by the auditory system as regularities. Over the course of presentation, high-level verse regularities may be continually learned, represented, and accessed, serving as referents. The current findings suggest that masker noise occurrences may be translated into missing values in the same low-level feature format as the low frequency envelope. While this does not preclude other dynamic features of speech to contribute to reactivation processes, such as higher order linguistic elements (e.g., Näätänen and Winkler, 1999; van Wassenhove and Schroeder, 2012; Di Liberto et al., 2015; Kayser et al., 2015), the key neural property of natural sound encoding via temporally based acoustic representations is underscored by its active maintenance during noise gaps as a function of prior experience.

AUTHOR CONTRIBUTIONS

FCC conceived, designed, and performed the experiments, analyzed the data, and prepared the manuscript figures. JZS supervised the research. Both authors wrote the manuscript text.

FUNDING

This study was funded by the National Institutes of Health (R01-DC-014085).

ACKNOWLEDGMENTS

We thank Anna Namyst for excellent technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnsys.2018.00056/full#supplementary-material

REFERENCES

Agus, T. R., Thorpe, S. J., and Pressnitzer, D. (2010). Rapid formation of robust auditory memories: insights from noise. *Neuron* 66, 610–618. doi: 10.1016/j.neuron.2010.04.014

Ahissar, M., Nahum, M., Nelken, I., and Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philos. Trans. R. Soc. B Biol. Sci.* 364, 285–299. doi: 10.1098/rstb.2008.0253

Andrillon, T., Kouider, S., Agus, T., and Pressnitzer, D. (2015). Perceptual learning of acoustic noise generates memory-evoked potentials. *Curr. Biol.* 25, 2823–2829. doi: 10.1016/j.cub.2015.09.027

Backer, K. C., and Alain, C. (2012). Orienting attention to sound object representations attenuates change deafness. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1554–1566. doi: 10.1037/a0027858

Backer, K. C., and Alain, C. (2014). Attention to memory: orienting attention to sound object representations. *Psychol. Res.* 78, 439–452. doi: 10.1007/s00426-013-0531-7

Bailes, F. (2007). The prevalence and nature of imagined music in the everyday lives of music students. *Psychol. Music* 35, 555–570. doi: 10.1177/0305735607077834

Balaguer-Ballester, E., Clark, N. R., Coath, M., Krumbholz, K., and Denham, S. L. (2009). Understanding pitch perception as a hierarchical process with top-down modulation. *PLoS Comput. Biol.* 5:e1000301. doi: 10.1371/journal.pcbi.1000301

Bendat, J. S., and Piersol, A. G. (2010). *The Hilbert Transform, Random Data: Analysis and Measurement Procedures*. Hoboken, NJ: John Wiley & Sons, 473–503. doi: 10.1002/9781118032428

Bendixen, A., Scharinger, M., Strauß, A., and Obleser, J. (2014). Prediction in the service of comprehension: modulated early brain responses to

omitted speech segments. *Cortex* 53, 9–26. doi: 10.1016/j.cortex.2014.01.001

Birkett, P. B., Hunter, M. D., Parks, R. W., Farrow, T. F., Lowe, H., Wilkinson, I. D., et al. (2007). Voice familiarity engages auditory cortex. *Neuroreport* 18, 1375–1378. doi: 10.1097/WNR.0b013e3282aa43a3

Buchsbaum, B. R., Olsen, R. K., Koch, P., and Berman, K. F. (2005). Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory. *Neuron* 48, 687–697. doi: 10.1016/j.neuron.2005.09.029

Buchsbaum, B. R., Padmanabhan, A., and Berman, K. F. (2011). The neural substrates of recognition memory for verbal information: spanning the divide between short- and long-term memory. *J. Cogn. Neurosci.* 23, 978–991. doi: 10.1162/jocn.2010.21496

Cervantes Constantino, F., and Simon, J. Z. (2017). Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm. *Sci. Rep.* 7:17536. doi: 10.1038/s41598-017-17063-0

Cervantes Constantino, F., Villafañe-Delgado, M., Camenga, E., Dombrowski, K., Walsh, B., and Simon, J. Z. (2017). Functional significance of spectrotemporal response functions obtained using magnetoencephalography. *bioRxiv* [Preprint]. doi: 10.1101/168997

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* 25, 975–979. doi: 10.1121/1.1907229

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Cowan, N. (1984). On short and long auditory stores. *Psychol. Bull.* 96, 341–370. doi: 10.1037/0033-2909.96.2.341

Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. *J. Neurosci.* 36, 9888–9895. doi: 10.1523/JNEUROSCI.1396-16.2016

Crowder, R. G. (1989). Imagery for musical timbre. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 472–478. doi: 10.1037/0096-1523.15.3.472

David, S. V., Fritz, J. B., and Shamma, S. A. (2012). Task reward structure shapes rapid receptive field plasticity in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 109, 2144–2149. doi: 10.1073/pnas.1117717109

de Cheveigné, A., and Simon, J. Z. (2007). Denoising based on time-shift PCA. *J. Neurosci. Methods* 165, 297–305. doi: 10.1016/j.jneumeth.2007.06.003

de Cheveigné, A., and Simon, J. Z. (2008a). Denoising based on spatial filtering. *J. Neurosci. Methods* 171, 331–339. doi: 10.1016/j.jneumeth.2008.03.015

de Cheveigné, A., and Simon, J. Z. (2008b). Sensor noise suppression. *J. Neurosci. Methods* 168, 195–202. doi: 10.1016/j.jneumeth.2007.09.012

Di Liberto, G. M., Lalor, E. C., and Millman, R. E. (2018). Causal cortical dynamics of a predictive enhancement of speech intelligibility. *Neuroimage* 166, 247–258. doi: 10.1016/j.neuroimage.2017.10.066

Di Liberto, G. M., O'Sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Ding, N., Chatterjee, M., and Simon, J. Z. (2013). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88, 41–46. doi: 10.1016/j.neuroimage.2013.10.054

Ding, N., and Simon, J. Z. (2012a). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U.S.A.* 109, 11854–11859. doi: 10.1073/pnas.1205381109

Ding, N., and Simon, J. Z. (2012b). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *J. Neurophysiol.* 107, 78–89. doi: 10.1152/jn.00297.2011

Durlach, N. I., and Braida, L. D. (1969). Intensity perception. I. preliminary theory of intensity resolution. *J. Acoust. Soc. Am.* 46, 372–383. doi: 10.1121/1.1911699

Fritz, J., Shamma, S., Elhilali, M., and Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat. Neurosci.* 6, 1216–1223. doi: 10.1038/nn1141

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., et al. (2000). Representation of the temporal envelope of sounds in the human brain. *J. Neurophysiol.* 84, 1588–1598. doi: 10.1152/jn.2000.84.3.1588

Grimm, S., and Schröger, E. (2007). The processing of frequency deviations within sounds: evidence for the predictive nature of the Mismatch Negativity (MMN) system. *Restor. Neurol. Neurosci.* 25, 241–249.

Henry, M. J., and Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proc. Natl. Acad. Sci. U.S.A.* 109, 20095–20100. doi: 10.1073/pnas.1213390109

Hubbard, T. L. (2010). Auditory imagery: empirical findings. *Psychol. Bull.* 136, 302–329. doi: 10.1037/a0018436

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.* 10, 626–634. doi: 10.1109/72.761722

Intons-Peterson, M. J. (2014). "Components of auditory imagery," in *Auditory Imagery*, ed. D. Reisberg (Hove: Psychology Press), 45–72.

Janata, P. (2001). Brain electrical activity evoked by mental formation of auditory expectations and images. *Brain Topogr.* 13, 169–193. doi: 10.1023/A:1007803102254

Janata, P., and Paroo, K. (2006). Acuity of auditory images in pitch and time. *Percept. Psychophys.* 68, 829–844. doi: 10.3758/BF03193705

Kaiser, J. (2015). Dynamics of auditory working memory. *Front. Psychol.* 6:613. doi: 10.3389/fpsyg.2015.00613

Kayser, S. J., Ince, R. A. A., Gross, J., and Kayser, C. (2015). Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *J. Neurosci.* 35, 14691–14701. doi: 10.1523/JNEUROSCI.2243-15.2015

Kiebel, S. J., Daunizeau, J., and Friston, K. J. (2008). A hierarchy of time-scales and the brain. *PLoS Comput. Biol.* 4:e1000209. doi: 10.1371/journal.pcbi.1000209

Kleinschmidt, A., Büchel, C., Hutton, C., Friston, K. J., and Frackowiak, R. S. J. (2002). The neural structures expressing perceptual hysteresis in visual letter recognition. *Neuron* 34, 659–666. doi: 10.1016/S0896-6273(02)00694-3

Kumar, S., Stephan, K. E., Warren, J. D., Friston, K. J., and Griffiths, T. D. (2007). Hierarchical processing of auditory objects in humans. *PLoS Comput. Biol.* 3:e100. doi: 10.1371/journal.pcbi.0030100

Leonard, M. K., Baud, M. O., Sjerps, M. J., and Chang, E. F. (2016). Perceptual restoration of masked speech in human cortex. *Nat. Commun.* 7:13619. doi: 10.1038/ncomms13619

Lewis, D., Hoover, B., Choi, S., and Stelmachowicz, P. (2010). The relationship between speech perception in noise and phonological awareness skills for children with normal hearing. *Ear Hear.* 31, 761–768. doi: 10.1097/AUD.0b013e3181e5d188

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N. E., Rieger, J., et al. (2014). Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* 7:14. doi: 10.3389/fneng.2014.00014

Martin, S., Mikutta, C., Leonard, M. K., Hungate, D., Koelsch, S., Shamma, S., et al. (2017). Neural encoding of auditory features during music perception and imagery. *Cereb. Cortex* doi: 10.1093/cercor/bhx277 [Epub ahead of print].

McGuire, P. K., Silbersweig, D. A., Murray, R. M., David, A. S., Frackowiak, R. S., and Frith, C. D., (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychol. Med.* 26, 29–38. doi: 10.1017/S0033291700033699

Mesgarani, N. (2014). "Stimulus reconstruction from cortical responses," in *Encyclopedia of Computational Neuroscience*, eds D. Jaeger and R. Jung (New York, NY: Springer), 1–3. doi: 10.1007/978-1-4614-7320-6_108-1

Mesgarani, N., David, S. V., Fritz, J. B., and Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J. Neurophysiol.* 102, 3329–3339. doi: 10.1152/jn.91128.2008

Meyer, M., Elmer, S., Baumann, S., and Jancke, L. (2007). Short-term plasticity in the auditory system: differential neural responses to perception and imagery of speech and music. *Restor. Neurol. Neurosci.* 25, 411–431.

Näätänen, R., and Winkler, I. (1999). The concept of auditory stimulus representation in cognitive neuroscience. *Psychol. Bull.* 125, 826–859. doi: 10.1037/0033-2909.125.6.826

Nahum, M., Nelken, I., and Ahissar, M. (2008). Low-level information and high-level perception: the case of speech in noise. *PLoS Biol.* 6:e126. doi: 10.1371/journal.pbio.0060126

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage* 56, 400–410. doi: 10.1016/j.neuroimage.2010.07.073

Nittrouer, S. (2002). From ear to cortex: a perspective on what clinicians need to understand about speech perception and language processing. *Lang. Speech Hear. Serv. Sch.* 33, 237–252. doi: 10.1044/0161-1461(2002/020)

Nozaradan, S., Peretz, I., Missal, M., and Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *J. Neurosci.* 31, 10234–10240. doi: 10.1523/JNEUROSCI.0411-11.2011

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., et al. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cereb. Cortex* 25, 1697–1706. doi: 10.1093/cercor/bht355

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., et al. (2012). Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251. doi: 10.1371/journal.pbio.1001251

Pearson, J., and Brascamp, J. (2008). Sensory memory for ambiguous vision. *Trends Cogn. Sci.* 12, 334–341. doi: 10.1016/j.tics.2008.05.006

Pearson, J., Naselaris, T., Holmes, E. A., and Kosslyn, S. M. (2015). Mental imagery: functional mechanisms and clinical applications. *Trends Cogn. Sci.* 19, 590–602. doi: 10.1016/j.tics.2015.08.003

Petkov, C. I., O'Connor, K. N., and Sutter, M. L. (2007). Encoding of illusory continuity in primary auditory cortex. *Neuron* 54, 153–165. doi: 10.1016/j.neuron.2007.02.031

Petkov, C. I., and Sutter, M. L. (2011). Evolutionary conservation and neuronal mechanisms of auditory perceptual restoration. *Hear. Res.* 271, 54–65. doi: 10.1016/j.heares.2010.05.011

Pouget, A., Beck, J. M., Ma, W. J., and Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* 16, 1170–1178. doi: 10.1038/nn.3495

Prichard, D., and Theiler, J. (1994). Generating surrogate data for time series with several simultaneously measured variables. *Phys. Rev. Lett.* 73, 951–954. doi: 10.1103/PhysRevLett.73.951

Ramirez, A. D., Ahmadian, Y., Schumacher, J., Schneider, D., Woolley, S. M. N., and Paninski, L. (2011). Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. *J. Neurosci.* 31, 3828–3842. doi: 10.1523/JNEUROSCI.3256-10.2011

Riecke, L., Esposito, F., Bonte, M., and Formisano, E. (2009). Hearing illusory sounds in noise: the timing of sensory-perceptual transformations in auditory cortex. *Neuron* 64, 550–561. doi: 10.1016/j.neuron.2009.10.016

Riecke, L., Vanbussel, M., Hausfeld, L., Başkent, D., Formisano, E., and Esposito, F. (2012). Hearing an illusory vowel in noise: suppression of auditory cortical activity. *J. Neurosci.* 32, 8024–8034. doi: 10.1523/JNEUROSCI.0440-12.2012

Samuel, A. (1996). Phoneme restoration. *Lang. Cogn. Process.* 11, 647–654. doi: 10.1080/016909696387051

Samuel, A. G. (1981). Phonemic restoration: insights from a new methodology. *J. Exp. Psychol. Gen.* 110, 474–494. doi: 10.1037/0096-3445.110.4.474

SanMiguel, I., Widmann, A., Bendixen, A., Trujillo-Barreto, N., and Schröger, E. (2013). Hearing silences: human auditory processing relies on preactivation of sound-specific brain activity patterns. *J. Neurosci.* 33, 8633–8639. doi: 10.1523/JNEUROSCI.5821-12.2013

Schwiedrzik, C. M., Ruff, C. C., Lazar, A., Leitner, F. C., Singer, W., and Melloni, L. (2014). Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. *Cereb. Cortex* 24, 1152–1164. doi: 10.1093/cercor/bhs396

Shahin, A. J., Bishop, C. W., and Miller, L. M. (2009). Neural mechanisms for illusory filling-in of degraded speech. *Neuroimage* 44, 1133–1143. doi: 10.1016/j.neuroimage.2008.09.045

Shergill, S. S., Bullmore, E. T., Brammer, M. J., Williams, S. C., Murray, R. M., and McGuire, P. K. (2001). A functional study of auditory verbal imagery. *Psychol. Med.* 31, 241–253. doi: 10.1017/S003329170100335X

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182–186. doi: 10.1016/j.tics.2008.02.003

Snyder, J. S., and Gregg, M. K. (2011). Memory for sound, with an ear toward hearing in complex auditory scenes. *Atten. Percept. Psychophys.* 73, 1993–2007. doi: 10.3758/s13414-011-0189-4

Snyder, J. S., Schwiedrzik, C. M., Vitela, A. D., and Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Front. Hum. Neurosci.* 9:594. doi: 10.3389/fnhum.2015.00594

Sohoglu, E., and Davis, M. H. (2016). Perceptual learning of degraded speech by minimizing prediction error. *Proc. Natl. Acad. Sci. U.S.A.* 113, E1747–E1756. doi: 10.1073/pnas.1523266113

Sohoglu, E., Peelle, J. E., Carlyon, R. P., and Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *J. Neurosci.* 32, 8443–8453. doi: 10.1523/JNEUROSCI.5069-11.2012

Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., et al. (2017). Neural entrainment to the beat: the "missing-pulse" phenomenon. *J. Neurosci.* 37, 6331–6341. doi: 10.1523/JNEUROSCI.2500-16.2017

Tavano, A., Grimm, S., Costa-Faidella, J., Slabu, L., Schröger, E., and Escera, C. (2012). Spectrotemporal processing drives fast access to memory traces for spoken words. *Neuroimage* 60, 2300–2308. doi: 10.1016/j.neuroimage.2012.02.041

van Wassenhove, V., Grant, K. W., and Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proc. Natl. Acad. Sci. U.S.A.* 102, 1181–1186. doi: 10.1073/pnas.0408949102

van Wassenhove, V., and Schroeder, C. E. (2012). "Multisensory role of human auditory cortex," in *The Human Auditory Cortex, Springer Handbook of Auditory Research*, eds D. Poeppel, T. Overath, A. N. Popper, and R. R. Fay (New York, NY: Springer), 295–331. doi: 10.1007/978-1-4614-2314-0_11

Wagner, R., and Torgesen, J. (1987). The nature of phonological processing and its causal role in the acquisition of reading-skills. *Psychol. Bull.* 101, 192–212. doi: 10.1037/0033-2909.101.2.192

Weinberger, N. M. (2004). Specific long-term memory traces in primary auditory cortex. *Nat. Rev. Neurosci.* 5, 279–290. doi: 10.1038/nrn1366

Wheeler, M. E., Petersen, S. E., and Buckner, R. L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11125–11129. doi: 10.1073/pnas.97.20.11125

Winkler, I., and Cowan, N. (2005). From sensory to long-term memory: evidence from auditory memory reactivation studies. *Exp. Psychol.* 52, 3–20. doi: 10.1027/1618-3169.52.1.3

Zatorre, R. J., Halpern, A. R., and Bouffard, M. (2009). Mental reversal of imagined melodies: a role for the posterior parietal cortex. *J. Cogn. Neurosci.* 22, 775–789. doi: 10.1162/jocn.2009.21239

Zimmermann, J. F., Moscovitch, M., and Alain, C. (2016). Attending to auditory memory. *Brain Res.* 1640(Part B), 208–221. doi: 10.1016/j.brainres.2015.11.032

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., et al. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron* 77, 980–991. doi: 10.1016/j.neuron.2012.12.037

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.