The neural representation of missing speech and the influence of prior knowledge on cortical fidelity and latency

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

[1] Program in Neuroscience and Cognitive Science; University of Maryland, College Park; College Park, MD 20742, USA.

[2] Department of Electrical and Computer Engineering; University of Maryland, College Park; College Park, MD 20742, USA.

[3] Department of Biology; University of Maryland, College Park; College Park, MD 20742, USA.

[4] Institute for Systems Research; University of Maryland, College Park; College Park, MD 20742, USA.

[*] Corresponding Author: Jonathan Z. Simon, jzsimon@umd.edu

[†] Current address: Centro de Investigación Básica en Psicología, Universidad de la República, Montevideo, 11200, Uruguay

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

Running title: Cortical reconstruction of missing speech

Abstract

In naturally noisy listening conditions, for example at a cocktail party, noise disruptions may completely mask significant parts of a sentence, and yet listeners may still perceive the missing speech as being present. Here we demonstrate that dynamic speech-related auditory cortical activity, as measured by magnetoencephalography (MEG), which can ordinarily be used to directly reconstruct to the physical speech stimulus, can also be used to "reconstruct" acoustically missing speech. The extent to which this occurs depends on the extent that listeners are familiar with the missing speech, which is consistent with this neural activity being a dynamic representation of perceived speech even if acoustically absence. Our findings are two-fold: first, we find that when the speech is entirely acoustically absent, the acoustically absent speech can still be reconstructed with performance up to 25% of that of acoustically present speech without noise; and second, that this same expertise facilitates faster processing of natural speech by approximately 5 ms. Both effects disappear when listeners have no or very little prior experience with a given sentence. Our results suggest adaptive mechanisms of consolidation of detailed representations about speech, and the enabling of strong expectations this entails, as identifiable factors assisting automatic speech restoration over ecologically relevant timescales.

1 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 clear informational advantages 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). 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. For instance, while cortical network activity profiles have been identified that are consistent with phonemic restoration (the effect where absent phonemes in a signal may nonetheless be heard (Samuel, 1996, 1981)) in binary semantic decision tasks (Leonard et al., 2016), the factors that bias into one or the other of two perceptual alternatives remain unclear. 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) or even intentional expectations about temporal patterns in sound (Nozaradan et al., 2011; Tal et al., 2017).

In order to affect ongoing speech percepts, the potential outcomes from these mechanisms must be readily accessible before and during missing auditory input. These type of contributions might 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* (Bendixen et al., 2014; Grimm and Schröger, 2007; Tavano et al., 2012). 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. We use the fact that the low-frequency envelope of speech (i.e., 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) (Di Liberto et al., 2015; Ding and Simon, 2012a; Giraud et al., 2000; Zion Golumbic et al., 2013). 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, it becomes possible to manipulate listeners' ability to develop detailed predictions about forthcoming elements in these long sentences. More repetitions would allow the generation of a better template for those tokens, to serve for a point-wise matching when later, spontaneous maskers disrupt the same string of tokens. Availability of a temporally-detailed template of the absent speech may allow the missing speech to be decoded from cortical signals representing a token, despite the acoustic absence of the speech itself. Furthermore,

because the template would be formed in advance, we also addressed the possibility that cortical representations of highly repeated speech stimuli may be facilitated in terms of processing time for those same speech tokens, even when *not* absent.

To address these hypotheses, we employ complementary systems-based neural analysis methods. In one case, we analyze neural responses in a way that allows reconstruction of a 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, 2012b; O'Sullivan et al., 2015) , electrocorticography (Leonard et al., 2016; Pasley et al., 2012), and fMRI (Naselaris et al., 2011). The performance of this decoding method allows a quantitative assessment of the extent to which prior knowledge of absent speech may enhance endogenous representations involved in its perceptual restoration. In the other case we instead use the stimulus speech envelope to estimate the neural response (Di Liberto et al., 2015; Ding and Simon, 2012a), under normal (non-absent) speech conditions. In this forward model case, we analyze cortical latencies involved in natural speech processing under different prior knowledge conditions. The possibility of reduced cortical latencies is of particular interest since faster processing has been observed in situations where additional context facilitates integration of incoming speech (van Wassenhove et al., 2005; van Wassenhove and Schroeder, 2012). Additionally, similar task-related cortical plasticity changes in stimulus-response mappings are often observed at the neuronal level (David et al., 2012; Fritz et al., 2003) and represent a potential biophysical basis for restorative mechanisms given the present task demands.

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 more knowledge about that particular speech sequence. Overall, the results suggest that the formation of online templates representing low-level features of frequently experienced speech may facilitate more efficient neural representations, both by means of faster encoding and by improved access to endogenous dynamic neural speech representations, time-locked to expected but missing speech, thus assisting its restoration.

2 Materials and methods

2.1 Participants. 35 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 hours). 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.

2.2 Stimuli and experimental design. Sound stimuli were prepared with the MATLAB® software package (MathWorks, Natick, 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 <http://archive.org/details/AVisitFromSt.Nicholas-ByClementClarkeMoore-NarratedByGrantRaymond>. Each of the fourteen verses (each verse being a quatrain of four lines) in the poem were separated and used as individual stimuli. Silence intervals (gaps) in the narration were reduced to approximately equalize stimuli durations (range: 13.1 – 13.6 s). Four stimulus blocks were presented in total, each containing 64 stimuli (i.e., 256 lines), with some stimuli repeated multiple times. For the first block, a verse from the first half of the poem was chosen as a 'High' frequency stimulus, repeated for half of the cases (32/64); similarly, different verses were chosen as 'Medium' and 'Low' frequency stimuli, which were repeated for a quarter (16/64) and an eighth (8/64) of the cases, respectively. The remainder of the block was filled with 'Control' stimuli, namely the four remaining verses presented either 1, 2 or 4 times within the block. 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 hour 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

5 listeners each that underwent the same ‘High’, ‘Medium’, ‘Low’, and ‘Control’ stimuli selection.

For each stimuli, 2–4 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.

2.3 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, 2012a).

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 $fT \cdot Hz^{-1/2}$ 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.

2.4 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 ~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, 2008a).

2.5 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, 2008b), was used to construct components (virtual channels constructed of linear combinations of the sensor channels), ranked in order of their trial-to-trial reproducibility, and used as described below.

2.5.1 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. The first three DSS components (i.e. with highest reproducibility) were used to train an optimal linear decoder, designed to reconstruct the envelope of the stimulus responsible for any particular response based on the reproducible aspects of the neural response under normal speech listening conditions. The last three DSS components (with the lowest reproducibility from the same dataset), were similarly used to train a separate linear decoder, used as a reference to estimate baseline. In each case, the decoding procedure produces a timeseries whose similarity with the original 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.

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 $q = \frac{1}{2}\left(\ln\frac{1+r_e}{1-r_e} - \ln\frac{1+r_f}{1-r_f}\right)$. Relative effect sizes (speech vs. 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).

2.5.2 Temporal response function of stimulus representation. The input-output relation between a representation $S(t)$ of auditory stimulus input and the evoked cortical response $\bar{r}(t)$ is modeled by a temporal response function (TRF). This linear model is formulated as:

$$\bar{r}_{\text{pred}}(t) = \sum_{\tau} TRF(\tau)\, S(t-\tau) + \epsilon(t)$$

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), according to which one produced a TRF with a more prominent $M100_{TRF}$, a strong negative peak with ~100 ms latency (Ding and Simon, 2012b).

2.5.3 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 temporal response function 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.

3 Results

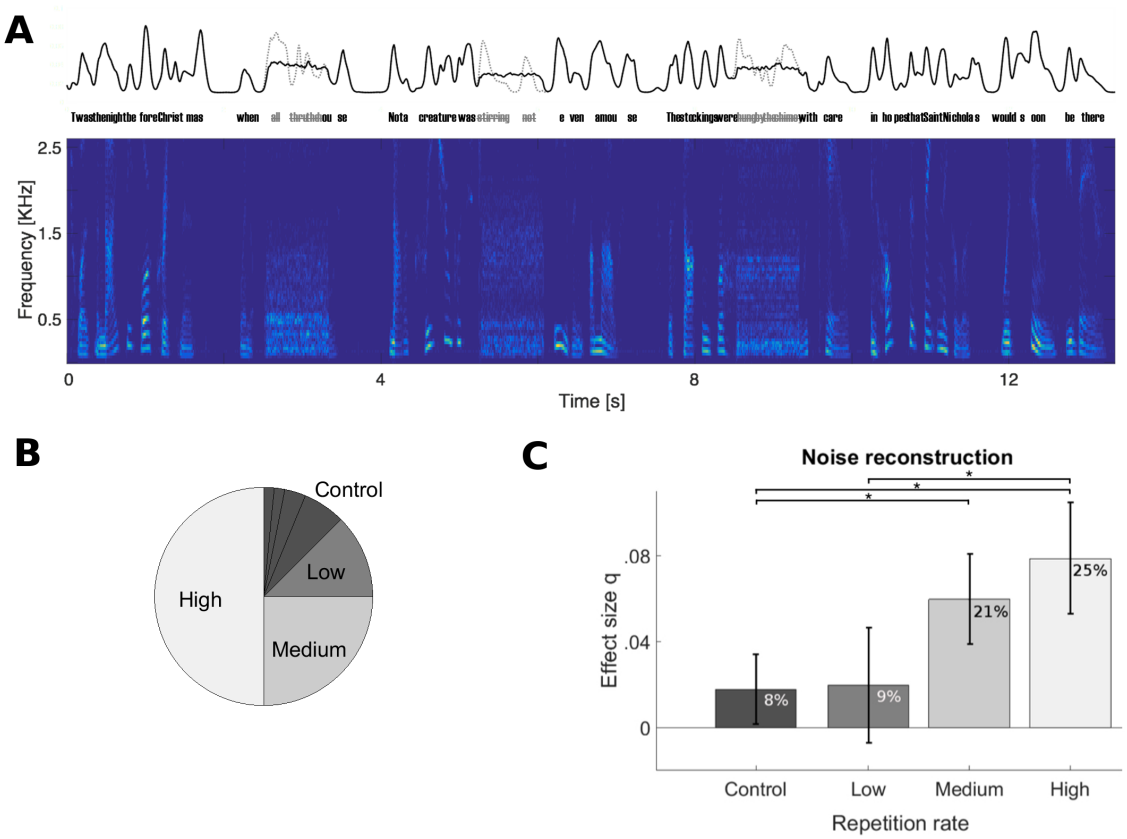

Figure 1. Cortical reconstruction of acoustically missing multi-word speech envelope from noise, as a function of repeated replays. (**A**) Speech material from a poem was repeatedly

presented to 35 listeners, but every 4-5 s some speech was replaced with spectrally-matched noise (0.8 s duration; three instances shown in spectrogram, bottom). This manipulation removes critical temporal modulation due to the missed words, as shown by the slow envelope (top). (**B**) Some verses were presented multiple times, taking up 50%, 25%, 12.5%, etc. of all verse presentations during the hour-long MEG recording session. (**C**) The missing dynamic speech envelope could nevertheless be reconstructed from responses to the static noise that replaced the missing speech, with performance up to 25% of that obtained under clean conditions (percentages inset within each bar). This effect was not an artifact of changing contributions from clean speech reconstructions, as indicated by an alternate measure of performance, i.e., normalized with respect to independent noise-trained decoders (scale on vertical axis, right). Error bars indicate confidence intervals for the means (Bonferroni-corrected α-level).

3.1 Reconstruction of missing speech from noise with context. Fixed-duration spectrally-matched static noise bursts were used to mask connected syllable/word sets within a narrated poem. Each noise probe was designed to have the same spectral composition over time as the replaced speech segment (Fig. 1A), without any supporting temporal modulations in the low-frequency (2-8 Hz) envelope (Ding and Simon, 2012a; Giraud et al., 2000). For natural speech without masking, these low-frequency fluctuations generate time-locked auditory cortical activity recorded by MEG and, given a suitable decoding model, can be used to reconstruct the envelope of the original speech signal. Such linear decoders were created to establish an optimal mapping from cortical activity to the original unmasked speech envelope. To test whether acoustic presence is a necessary condition for reconstruction of continuous speech, the 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 (Fig. 1B). 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 (Fig. 1C; percentages inset within each bar). Because this measure is relative to clean speech reconstruction, a measure of reconstruction from noise alone was also employed, 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; Control: 0.018 ± 0.008)(Fig. 1C). A one-way repeated measures ANOVA with four repetition levels was applied to determine whether decoding success of the linear model of the envelope significantly changed across conditions. Results for independent reconstructions using exclusively noise-derived *q* scores had shown that the sphericity condition was not violated (Mauchly test, $\chi^2(5)=6.322$; $p=0.276$). The subsequent ANOVA resulted in a significant main effect of repetition ($F(3,102)=8.070$; $p<0.001$). Post hoc pairwise comparisons using Bonferroni correction revealed that this increased exposure to speech significantly improved the stimulus reconstruction effect size from Control and Low repetition rate conditions to High ($p=0.002$ and $p=0.001$ respectively), and also from Control to Medium ($p=0.008$).

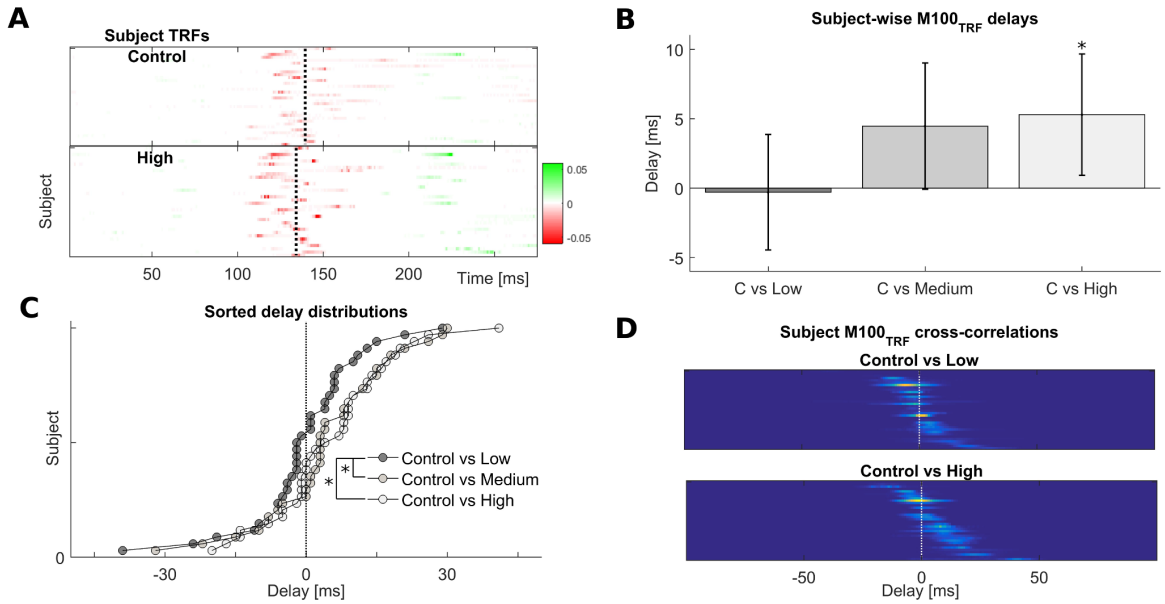

Figure 2. Frequent repetitions of natural speech speed-up their cortical processing. (**A**) Temporal response functions across participants reveal a common cortical processing step, referred to as the $M100_{TRF}$, typically occurring about 100 ms after a speech envelope fluctuation (red colored features 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.

3.2 Expedited auditory cortical processing of frequent natural speech replays. The temporal response function (TRF) is a functionally informative statistic, derived from a linear model, that predicts the neural response to sound stimuli, via a representation of the stimulus such as the acoustic envelope. Its characteristic peaks, and especially their polarity and latencies, are indicative of distinct neural processing stages, akin to the

distinct generators of evoked responses to simple sounds such as pure tones, but directly derived from the neural processing of continuous speech (Cervantes Constantino et al., 2017; Ding and Simon, 2012a, 2012b). We examined the effect of prior exposure on the TRF's temporal structure in general, and also for a specific peak, the $M100_{TRF}$, occurring 100-200 ms post envelope change (Fig. 2A). 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 $\mathrm{M100}_{TRF}^{\mathrm{High}}$ versus $\mathrm{M100}_{TRF}^{\mathrm{Control}}$ peaks ($t(33)=2.387$; $p=0.023$), indicating expedited cortical processing cortical for more familiar stimuli (Fig. 2B). 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 ($D=0.294$; $p=0.043$), suggesting that prior experience by repeated presentations effectively speeds up cortical processing even as early as 100 ms latency.

4 Discussion

The phenomenon of sensory restoration relies on inference regarding elements missing from a sensory signal. The results here demonstrate that auditory cortical activity measured by MEG contains information to reconstruct the missing sequences of speech replaced by noise, provided that a listener was previously and repeatedly exposed to the missing speech. Results therefore suggest that prior experience enables access and maintenance of a detailed representation of the stimulus, in a template format compatible with the dynamical acoustic envelope; a process that may in addition be related to speed up of cortical processing time. Together, these results point to the generation of a time-

locked, internally generated neural activity pattern consistent with the expected but absent sensory input. These findings complement those from related experiments investigating restoration at the phoneme-duration scale (e.g., disruptions lasting < 200 ms), which show that the acoustic presence of a specific sound pattern is not necessary for spectrogram reconstructability when speech is replaced by noise (Leonard et al., 2016) – as long as the immediate acoustic context is consistent with the restored phoneme. These results imply that the corresponding neural activity must rely on endogenous processes, possibly as top-down context-based modulations of auditory cortex populations (Petkov et al., 2007; Petkov and Sutter, 2011). The results here are consistent with the notion that this activity can be influenced by prior learning and storage of speech information, even at the level of its explicit temporal structure. Under this interpretation, enhanced listeners' expectations about forthcoming speech tokens may predispose them to restorative encoding, in contrast to the case when contextual information is poor or insufficient, where endogenous neural dynamics may fail to adhere to or predict the missing stimulus representation. Spontaneous neural background activity known to influence perceptual processing in general, includes the ability to entrain to a complex, natural signals such as speech (Ding et al., 2013), to optimize behavioral performance of detection tasks (Henry and Obleser, 2012), or even to increase the robustness of certain auditory illusory experiences (Riecke et al., 2009).

4.1 Plausibility of auditory memory involvement in context effects. The auditory restoration effect investigated here may be considered part of the multimodal class of *attractive temporal context effects* (Snyder et al., 2015), a group of facilitatory

mechanisms including perceptual hysteresis (Kleinschmidt et al., 2002; Schwiedrzik et al., 2014) and perceptual stabilization (Pearson and Brascamp, 2008) in the vision literature. These are considered critical for improving perceptual invariance in the face of external demands imposed by discontinuously fluctuating, broadly cluttered environments. Conceptually, this class stands opposite to that of *contrastive* temporal context effects, which are mainly suppressive, habituation or fatigue-based biases that discount neural activity after repetitions, and effectively favor perceptual alternatives for which neural activity has not yet been adapted (Schwiedrzik et al., 2014; Snyder et al., 2015). These may include semantic satiation effects, i.e., the subjective experience of increasingly meaningless words after fast and prolonged repeats (Kounios et al., 2000; Pilotti et al., 1997). Some conceptual frameworks for the organization of auditory cortical areas integrate neural coding functions with cognitive and adaptive functions such as relevance analyses of sound features, and their storage, directly in primary cortical areas (Weinberger, 2004). Storage of present connected speech sequences into sensory memory would then require retention of memory traces over the span of a few seconds, as well as past completion of stimuli resolution rendered by composite collections of features that are more efficient for long term storage (Cowan, 1984). Sensory memory 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), and the involvement of memory-based reactivation in perceptual processes, including attention, is an area of active research (Backer and Alain, 2012, 2014; Zimmermann et al., 2016).

4.2 Access and format of stored auditory representations. Over the course of acoustic stimulus repetitions, attractive contextual effects may rely on implicit auditory memory, which is considered to regularly intervene in sensory and perceptual encoding (Snyder and Gregg, 2011). One such example is the improved detection of arbitrary noise structures after sequential presentations, and the time-locked potential sensory covariates of this improvement (Agus et al., 2010; Andrillon et al., 2015). Foreknowledge of acoustic features may allow listeners to adapt to a likely communication source, as demonstrated by perceptual facilitation when advance notice about the identity of a forthcoming instrument play is given (Crowder, 1989), and by preferential activation in auditory association areas specific to speaker familiarity (Birkett et al., 2007). The notion that strong expectations of a dynamic sound pattern influence the level of detail accessible in sensory representations is supported by findings of differential activation in implicit memory tasks with varying rates of sensory update: initially, short storage intervals may be associated with activation of posterior superior temporal areas, and over time, activity can be mediated by structures in inferior frontal cortex (Buchsbaum et al., 2011). Evidence from these studies is consistent with the hypothesis of transformation of memory trace representation formats, where readout from sensory buffers is at high temporal resolution under low-level representation formats, while coarser temporal resolutions may occur instead at stores that encode categorical higher-order input features (cf. Durlach and Braida, 1969; Winkler and Cowan, 2005).

4.3 The role of auditory imagery and related retrieval processes in listening in noise. Perceptual restoration phenomena, including phonemic restoration, may be related to

auditory imagery defined as the persistence of an auditory experience without prompting by direct sensory input (Intons-Peterson, 2014). During stimulus masking, sensory imagery is postulated to involve 'schemata' or prior abstractions actively formed with perceptual input that become better resolved with increased familiarity, and which may remain online while an expected stimulus fails to occur (Hubbard, 2010). The implication, for methodological purposes, is that the occurrence of auditory imagery processes can be judged either by subjective reports or by using tasks hypothesized to involve imagery with reasonable probability (Hubbard, 2010). This latter approach employs familiarity of prior experience as a condition for stimuli to automatically evoke auditory imagery of original natural sound pieces (Bailes, 2007; Meyer et al., 2007). Neurally, the planum temporale is a major computational hub for which activation levels may correlate with self-reported levels of engagement with imagery, or with perceived vividness by listeners (Zatorre et al., 2009), and auditory imagery and (related) rehearsal of natural complex sounds may be subserved by auditory association cortex areas therein (Hubbard, 2010; Martin et al., 2014). There is also evidence for a dual format of representations sustained during active rehearsing, under both auditory-specific (sometimes termed 'echoic memory') and modality-general codes; these two coding schemes have been indicated over distinct locations each on superior temporal cortical areas, with distinct timescales as transient (< 5 s) versus sustained phases respectively (Buchsbaum et al., 2005; Meyer et al., 2007). The present data are thus consistent with a common theme in auditory retrieval processes, for which task-relevant stimuli and/or features may rely on maintenance of (re)activated domains within the sensory representational space (Kaiser, 2015). This is also supported by findings of retrieval

processes in vision and hearing that involve reactivation of sensory regions active during perception (Wheeler et al., 2000), something also found with auditory verbal imagery (McGuire et al., 1996; Shergill et al., 2001), overall pointing to the notion that both involve overlapping processes (Hubbard, 2010).

4.4 Adaptive dynamics of speech encoding and representation during masking. The brain's utilization of a neural model of speech input, used dynamically to infer the content of bottom-up sensory information (Pouget et al., 2013), indicates two separate but related strategies. First, the finding that cortical processing is sped up under the same circumstances that promote neural restoration of speech-coherent neural activity suggests that active, task-related endogenous processes directly optimize low-level speech processing with relevant experience. One plausible mechanism is increased excitability in a population which normally only becomes active at later stages of speech processing. Determining conditions under which this occurs may in the future provide real-time noninvasive indices of the subjective states by which a person maintains in register a template auditory pattern. Second, our results are consistent with the suggestion that auditory 'image' formation entails activity consistent with that elicited by original sound input (Janata, 2001; Martin et al., 2017), where preservation of the temporal acuity (and related properties) of the original stimulus may deteriorate depending on factors such as context and experience (Janata and Paroo, 2006). The latter appears related to the different success rates in reconstruction of missing speech found here, which decreased for increasingly unfamiliar stimuli. A need for frequent "refreshing" then echoes the auditory memory reactivation hypothesis where storage of individual sound features is embedded in the context of those neighboring patterns and sequences representable by

the auditory system as regularities. Reactivation here denotes the automatic process where variable sound input is matched to constancies extracted previously; likelihood of storage is then increased by proximity between a prior rule and current update tokens (Winkler and Cowan, 2005). This description, originating from oddball sequence studies, can be considered to apply in the present study across its verse stimulus structure: e.g., dynamic acoustic features of speech preceding a masker may serve as referents for a listener, enabling the process of translation of verse regularities learned and represented over the course of the experiment, into specific values in the same feature format (Winkler and Cowan, 2005). While this does not preclude additional dynamic stimulus features also contributing, including higher-order linguistic elements (e.g. Di Liberto et al., 2015; Kayser et al., 2015; Näätänen and Winkler, 1999; Wassenhove and Schroeder, 2012), the suggestion that a key neural property of natural sound encoding is via temporally-based acoustic representations is underscored by their active maintenance during noise gaps, based on prior experience.

Conflict of Interest

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.

Funding

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

Acknowledgments

We thank Anna Namyst for excellent technical assistance.

Data availability All relevant data are available to all interested parties in a public repository at <http://hdl.handle.net/1903/20259>.

References

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

Andrillon, T., Kouider, S., Agus, T., Pressnitzer, D., 2015. Perceptual Learning of Acoustic Noise Generates Memory-Evoked Potentials. Curr. Biol. 25, 2823–2829. https://doi.org/10.1016/j.cub.2015.09.027

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

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

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

Bendat, J.S., Piersol, A.G., 2010. The Hilbert Transform, in: Random Data: Analysis and Measurement Procedures. John Wiley & Sons, Inc., pp. 473–503.

Bendixen, A., Scharinger, M., Strauß, A., Obleser, J., 2014. Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments. Cortex 53, 9–26. https://doi.org/10.1016/j.cortex.2014.01.001

Birkett, P.B., Hunter, M.D., Parks, R.W., Farrow, T.F., Lowe, H., Wilkinson, I.D., Woodruff, P.W., 2007. Voice familiarity engages auditory cortex. Neuroreport

18, 1375–1378. https://doi.org/10.1097/WNR.0b013e3282aa43a3

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

Buchsbaum, B.R., Padmanabhan, A., 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. https://doi.org/10.1162/jocn.2010.21496

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

Cervantes Constantino, F., Villafañe-Delgado, M., Camenga, E., Dombrowski, K., Walsh, B., Simon, J.Z., 2017. Functional significance of spectrotemporal response functions obtained using magnetoencephalography. bioRxiv 168997. https://doi.org/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. https://doi.org/10.1121/1.1907229

Cohen, J., 1988. Statistical power analysis for the behavioral sciences. L. Erlbaum Associates, Hillsdale, N.J. :

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

Crosse, M.J., Di Liberto, G.M., 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. Off. J. Soc. Neurosci. 36, 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

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

David, S.V., Fritz, J.B., 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. https://doi.org/10.1073/pnas.1117717109

de Cheveigné, A., Simon, J.Z., 2008a. Sensor noise suppression. J. Neurosci. Methods 168, 195–202. https://doi.org/10.1016/j.jneumeth.2007.09.012

de Cheveigné, A., Simon, J.Z., 2008b. Denoising based on spatial filtering. J. Neurosci. Methods 171, 331–339. https://doi.org/10.1016/j.jneumeth.2008.03.015

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

Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C., 2015. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr. Biol. 25, 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., Chatterjee, M., Simon, J.Z., 2013. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage. https://doi.org/10.1016/j.neuroimage.2013.10.054

Ding, N., Simon, J.Z., 2012a. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107, 78–89.

https://doi.org/10.1152/jn.00297.2011

Ding, N., Simon, J.Z., 2012b. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. 109, 11854–11859. https://doi.org/10.1073/pnas.1205381109

Durlach, N.I., Braida, L.D., 1969. Intensity Perception. I. Preliminary Theory of Intensity Resolution. J. Acoust. Soc. Am. 46, 372–383. https://doi.org/10.1121/1.1911699

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

Giraud, A.-L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the Temporal Envelope of Sounds in the Human Brain. J. Neurophysiol. 84, 1588–1598.

Giraud, A.L., Lorenzi, C., Ashburner, J., Wable, J., Johnsrude, I., Frackowiak, R., Kleinschmidt, A., 2000. Representation of the temporal envelope of sounds in the human brain. J. Neurophysiol. 84, 1588–1598.

Grimm, S., 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., Obleser, J., 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proc. Natl. Acad. Sci. 109, 20095–20100. https://doi.org/10.1073/pnas.1213390109

Hubbard, T.L., 2010. Auditory imagery: Empirical findings. Psychol. Bull. 136, 302–

329. https://doi.org/10.1037/a0018436

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

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

Janata, P., 2001. Brain electrical activity evoked by mental formation of auditory expectations and images. Brain Topogr. 13, 169–193.

Janata, P., Paroo, K., 2006. Acuity of auditory images in pitch and time. Percept. Psychophys. 68, 829–844.

Kaiser, J., 2015. Dynamics of auditory working memory. Front. Psychol. 6. https://doi.org/10.3389/fpsyg.2015.00613

Kayser, S.J., Ince, R.A.A., Gross, J., Kayser, C., 2015. Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. J. Neurosci. Off. J. Soc. Neurosci. 35, 14691–14701. https://doi.org/10.1523/JNEUROSCI.2243-15.2015

Kleinschmidt, A., Büchel, C., Hutton, C., Friston, K.J., Frackowiak, R.S.J., 2002. The Neural Structures Expressing Perceptual Hysteresis in Visual Letter Recognition. Neuron 34, 659–666. https://doi.org/10.1016/S0896-6273(02)00694-3

Kounios, J., Kotz, S.A., Holcomb, P.J., 2000. On the locus of the semantic satiation effect: evidence from event-related brain potentials. Mem. Cognit. 28, 1366–1377.

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

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N.E., Rieger, J., Schalk, G., Knight, R.T., Pasley, B.N., 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroengineering 7, 14. https://doi.org/10.3389/fneng.2014.00014

Martin, S., Mikutta, C., Leonard, M.K., Hungate, D., Koelsch, S., Shamma, S., Chang, E.F., Millán, J.D.R., Knight, R.T., Pasley, B.N., 2017. Neural Encoding of Auditory Features during Music Perception and Imagery. Cereb. Cortex N. Y. N 1991 1–12. https://doi.org/10.1093/cercor/bhx277

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

Mesgarani, N., 2014. Stimulus Reconstruction from Cortical Responses, in: Jaeger, D., Jung, R. (Eds.), Encyclopedia of Computational Neuroscience. Springer New York, pp. 1–3. https://doi.org/10.1007/978-1-4614-7320-6_108-1

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

Meyer, M., Elmer, S., Baumann, S., 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., Winkler, I., 1999. The concept of auditory stimulus representation in cognitive neuroscience. Psychol. Bull. 125, 826–859.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage, Multivariate Decoding and Brain Reading 56, 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nozaradan, S., Peretz, I., Missal, M., Mouraux, A., 2011. Tagging the Neuronal Entrainment to Beat and Meter. J. Neurosci. 31, 10234–10240. https://doi.org/10.1523/JNEUROSCI.0411-11.2011

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A., Lalor, E.C., 2015. Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cereb. Cortex N. Y. N 1991 25, 1697–1706. https://doi.org/10.1093/cercor/bht355

Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F., 2012. Reconstructing speech from human auditory cortex. PLoS Biol. 10, e1001251. https://doi.org/10.1371/journal.pbio.1001251

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

Petkov, C.I., O'Connor, K.N., Sutter, M.L., 2007. Encoding of Illusory Continuity in Primary Auditory Cortex. Neuron 54, 153–165. https://doi.org/10.1016/j.neuron.2007.02.031

Petkov, C.I., Sutter, M.L., 2011. Evolutionary conservation and neuronal mechanisms of auditory perceptual restoration. Hear. Res., Auditory Cortex: Current

Concepts in Human and Animal Research 271, 54–65. https://doi.org/10.1016/j.heares.2010.05.011

Pilotti, M., Antrobus, J.S., Duff, M., 1997. The effect of presemantic acoustic adaptation on semantic "satiation." Mem. Cognit. 25, 305–312.

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

Prichard, D., Theiler, J., 1994. Generating surrogate data for time series with several simultaneously measured variables. Phys. Rev. Lett. 73, 951.

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

Riecke, L., Esposito, F., Bonte, M., Formisano, E., 2009. Hearing Illusory Sounds in Noise: The Timing of Sensory-Perceptual Transformations in Auditory Cortex. Neuron 64, 550–561. https://doi.org/10.1016/j.neuron.2009.10.016

Riecke, L., Vanbussel, M., Hausfeld, L., Başkent, D., Formisano, E., Esposito, F., 2012. Hearing an Illusory Vowel in Noise: Suppression of Auditory Cortical Activity. J. Neurosci. 32, 8024–8034. https://doi.org/10.1523/JNEUROSCI.0440-12.2012

Samuel, A., 1996. Phoneme Restoration. Lang. Cogn. Process. 11, 647–654. https://doi.org/10.1080/016909696387051

Samuel, A.G., 1981. Phonemic restoration: Insights from a new methodology. J. Exp.

Psychol. Gen. 110, 474–494. https://doi.org/10.1037/0096-3445.110.4.474

SanMiguel, I., Widmann, A., Bendixen, A., Trujillo-Barreto, N., Schröger, E., 2013. Hearing Silences: Human Auditory Processing Relies on Preactivation of Sound-Specific Brain Activity Patterns. J. Neurosci. 33, 8633–8639. https://doi.org/10.1523/JNEUROSCI.5821-12.2013

Schwiedrzik, C.M., Ruff, C.C., Lazar, A., Leitner, F.C., Singer, W., Melloni, L., 2014. Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks. Cereb. Cortex N. Y. N 1991 24, 1152–1164. https://doi.org/10.1093/cercor/bhs396

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

Shergill, S.S., Bullmore, E.T., Brammer, M.J., Williams, S.C., Murray, R.M., McGuire, P.K., 2001. A functional study of auditory verbal imagery. Psychol. Med. 31, 241–253.

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

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

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

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

Tal, I., Large, E.W., Rabinovitch, E., Wei, Y., Schroeder, C.E., Poeppel, D., Zion Golumbic, E., 2017. Neural Entrainment to the Beat: The "Missing-Pulse" Phenomenon. J. Neurosci. Off. J. Soc. Neurosci. 37, 6331–6341. https://doi.org/10.1523/JNEUROSCI.2500-16.2017

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

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

van Wassenhove, V., Schroeder, C.E., 2012. Multisensory Role of Human Auditory Cortex, in: Poeppel, D., Overath, T., Popper, A.N., Fay, R.R. (Eds.), The Human Auditory Cortex, Springer Handbook of Auditory Research. Springer New York, pp. 295–331. https://doi.org/10.1007/978-1-4614-2314-0_11

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

Wheeler, M.E., Petersen, S.E., Buckner, R.L., 2000. Memory's echo: vivid remembering reactivates sensory-specific cortex. Proc. Natl. Acad. Sci. U. S. A.

97, 11125–11129.

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

Zatorre, R.J., Halpern, A.R., Bouffard, M., 2009. Mental Reversal of Imagined Melodies: A Role for the Posterior Parietal Cortex. J. Cogn. Neurosci. 22, 775–789. https://doi.org/10.1162/jocn.2009.21239

Zimmermann, J.F., Moscovitch, M., Alain, C., 2016. Attending to auditory memory. Brain Res., Auditory Working Memory 1640, Part B, 208–221. https://doi.org/10.1016/j.brainres.2015.11.032

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." Neuron 77, 980–991. https://doi.org/10.1016/j.neuron.2012.12.037