ABSTRACT

Title of dissertation: SENSORY AND PERCEPTUAL CODES IN CORTICAL AUDITORY PROCESSING

Francisco Israel Cervantes Constantino, Doctor of Philosophy, 2017

Dissertation directed by: Professor Jonathan Z. Simon
Department of Electrical and Computer Engineering
Department of Biology
Institute for Systems Research

A key aspect of human auditory cognition is establishing efficient and reliable representations about the acoustic environment, especially at the level of auditory cortex. Since the inception of encoding models that relate sound to neural response, three longstanding questions remain open. First, on the apparently insurmountable problem of fundamental changes to cortical responses depending on certain categories of sound (e.g. simple tones versus environmental sound). Second, on how to integrate inner or subjective perceptual experiences into sound encoding models, given that they presuppose existing, direct physical stimulation which is sometimes missed. And third, on how does context and learning fine-tune these encoding rules, as adaptive changes to improve impoverished conditions particularly important for communication sounds.

In this series, each question is addressed by analysis of mappings from sound stimuli delivered-to and/or perceived-by a listener, to large-scale cortically-sourced response time series from magnetoencephalography. It is first shown that the divergent, categorical modes of sensory coding may unify by exploring alternative

acoustic representations other than the traditional spectrogram, such as temporal transient maps. Encoding models of either of artificial random tones, music, or speech stimulus classes, were substantially matched in their structure when represented from acoustic energy increases –consistent with the existence of a domain-general common baseline processing stage.

Separately, the matter of the perceptual experience of sound via cortical responses is addressed via stereotyped rhythmic patterns normally entraining cortical responses with equal periodicity. Here, it is shown that under conditions of perceptual restoration, namely cases where a listener reports hearing a specific sound pattern in the midst of noise nonetheless, one may access such endogenous representations in the form of evoked cortical oscillations at the same rhythmic rate.

Finally, with regards to natural speech, it is shown that extensive prior experience over repeated listening of the same sentence materials may facilitate the ability to reconstruct the original stimulus even where noise replaces it, and to also expedite normal cortical processing times in listeners. Overall, the findings demonstrate cases by which sensory and perceptual coding approaches jointly continue to expand the enquiry about listeners' personal experience of the communication-rich soundscape.

SENSORY AND PERCEPTUAL CODES IN CORTICAL AUDITORY PROCESSING

by

Francisco Israel Cervantes Constantino

Dissertation proposal submitted to the Faculty of the Graduate School
of the University of Maryland, College Park in partial fulfillment
of the requirements for the degree of Doctor of Philosophy
2017

Advisory Committee:

Professor Jonathan Z. Simon, Chair
Professor Matthew Goupell, Dean's Representative
Professor Behtash Babadi
Professor Daniel A. Butts
Professor Ellen Lau

The projects were funded by the US National Health Institutes (NICDC R01 DC 00843, 014085). I thank support by the Mexican Consejo Nacional de Ciencia y Tecnología through its graduate scholarship program. I would like to thank substantive input, in collecting data for and proof-reading our first project, by Elizabeth Camenga, Katya Dombrowski, Benjamin Walsh, and Marisel Villafañe-Delgado. Also to Elizabeth Nguyen, Natalia Lapinskaya and Anna Namyst for their excellent technical assistance.

Table of contents

List of Figures

Introduction

Physiological origins of MEG signals

Information processing in the nervous system relies on the ability for neuron cell membranes to transfer electric charge in organized manner. Each transfer can be modeled as a point ionic current event, and in turn this generates a magnetic field in its near vicinity. Current transfer constrained by a neural wire-like segment (a 'process') implies an effective displacement along the course of its main axis (Fig. 1). By Maxwell equations, a new magnetic field distribution is there created, with a geometry that can be represented as a series of concentric magnetic field contours whose strength decreases with distance. If all transfer locations and directions were distributed randomly, the resulting local then add and cancel each other at chance, superposing to a resulting global picture of near-zero field at any given time, following a "closed field" (Fig. 1). It follows that for a MEG signal to be measurable, some degree of underlying coherent anatomical organization is necessary so that constructive superposition is favored (e.g. an open field configuration) as an effective magnetic source of greater strength than any of its constituents. Fortunately, in many locations the brain has a sufficient degree of organization for net fields from different locations to add constructively, and thus amplify into a neuromagnetic signal.

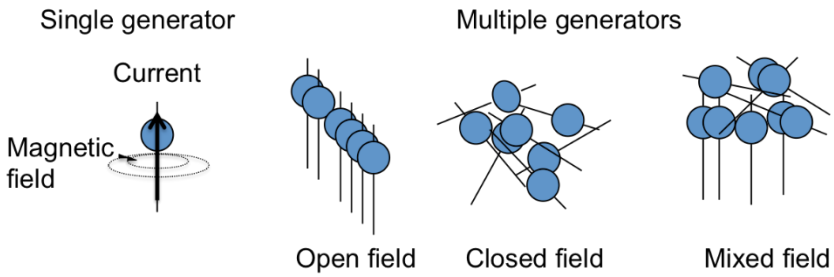


Figure 1. Magnetic field group generators. A single cell generator (left) with a mostly linear architecture may be active and its charge transfer effectively transfer along its axis. This is the source for a magnetic field distribution, and its strength can be depicted as contour lines that decrease with distance from the source direction (thick arrow); its own direction (contour arrows) depends additionally on the source orientation. Given these constraints, the spatial organization (directions and orientations) of multiple current generators (right) may determine how each contribution adds up externally, either adding up as an open field, cancelling out as a closed field, or likely keeping an intermediate form as a mixed field. Activity changes may happen from one another, and an aim of the MEG technique is to measure this global image, in particular *when* do over time. Still, even if a neuromagnetic signal is measurable it does not mean that it is directly interpretable. Fig. 1 illustrates the point that anatomical descriptions of the underlying generators are necessary for an unambiguous interpretations.

At the individual level neurons have a striking variety of shapes and sizes, something that may translate into different current transfer properties. Multipolar neurons with long axon processes and many dendrites are frequently found in the mammalian central nervous system, featuring among them the pyramidal cell (Fig. 2) with a chief excitatory function in layered structures such as cerebral cortex and hippocampus. For our purposes, its main morphological characteristic is its single, long thick apical dendrite that dictates a principal axis of information transfer from dendrites to axon, and of (bidirectional) current flow along its axis. Configurationally, these units are locally arranged in parallel to each other, and normal to the tissue surface layer. Globally, such histological pattern is

maintained along the folded surface of cortex or hippocampus. When jointly active, events at individual units may be amplified generating magnetic fields of sufficient strength to be measurable with current detector systems[1].

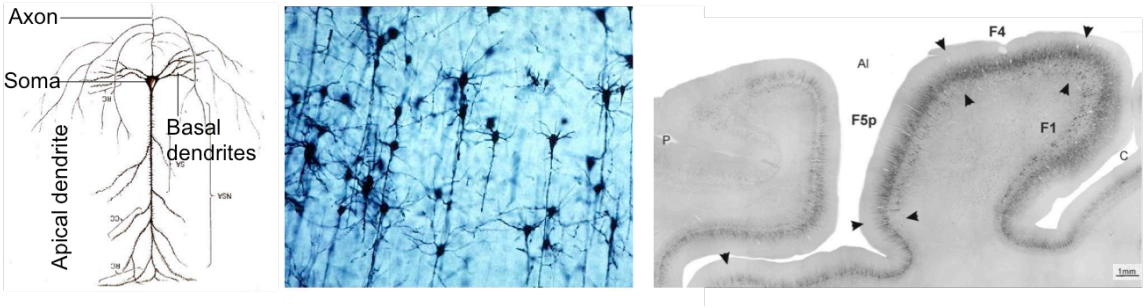


Figure 2. Biophysical structures facilitating neuromagnetic signals. Information transfer in a pyramidal excitatory cell from the hippocampus (left) flows upwards from dendrites to soma to axon. Most dendritic branches appear to end at the main apical dendrite leading to a principal axis of current flow. These cell types have an inverted pyramid-like soma of about 20 μm average size and, histologically, often arrange bidimensionally in parallel as shown by a silver stain of cerebral cortex (middle). Thus if simultaneously active, a magnetic field contour surrounding their surface domain may superpose additively. This normal-oriented-to-surface configuration is a major building block of cortical layered structure as is visible at the millimetre scale in a sample motor cortex section (right). Images modified from: NeuroLex.org, Stritch School of Medicine (Lumen), and [2].

The anatomical basis for neuromagnetic signals arising from these biophysical arrays was directly addressed in an in vitro study by Wu & Okada [3], where a slice of parallel pyramidal-cell tissue that would in principle facilitate an open field distribution was used (Fig. 3). Neuromagnetic signals were successfully measured from a these samples, with

consistent directionality according to the physiological current flow patterns predicted by stimulation at either extreme of the pyramidal cell set (soma or far dendrites). Depending on stimulation distance from the soma (near or far), evoked signals featured bi- or tri-phasic behavior respectively, where the first phase is explainable by intracellular current flow (from stimulation origin to opposite axial end), but the last phase flows in both cases from apex to base, also in consistency with simultaneous electrophysiology data. The difference between first and last phases was then interpreted in terms of cells that have been directly stimulated (~30%) initially, versus late neuromagnetic signals originating from recurrent excitatory connections within the slice [4].

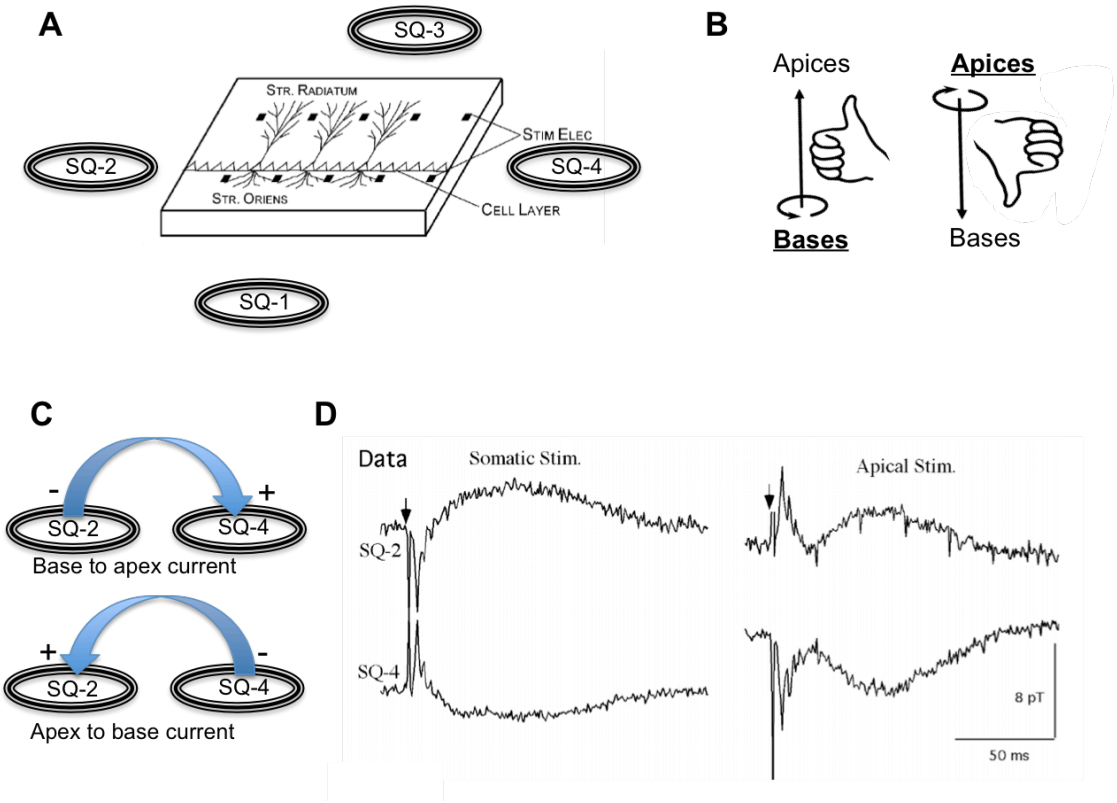


Figure 3. Direct evidence of open field magnetic signals from layer-organized pyramidal cells. (**A**) In the Wu & Okada experiment [3], a slice of hippocampus (CA3) tissue, where pyramidal cells are closely packed in parallel and fixed with equal orientations, is placed in the centre of four detector coil array (SQUIDs) screening

magnetic fields traversing each coil's circular area. Cell soma and a few cell's dendritic processes are represented. The slice was then stimulated along either the apical dendrite or the basal dendrite (soma) regions, defining either of two opposite current flow directions. Modified from [4]. (**B**) The right hand rule for the magnetic field caused by a current wire predicts its direction as it curls around the wire's axis, which are opposite depending on current flow as determined by stimulation location (bold). (**C**) In the coil array, a neuromagnetic signal consistent with the expected open field distribution will be recorded at coil detectors 2 and 4, but not at detectors 1 and 3, as the latter two lie along the current flow axis and no field traverses their circular area. Depending on the current flow direction, any magnetic field reaching coils 2 and 4 may be entering from below (negative) or above (positive) the coils (equivalently, exiting from above or below them). (**D**) Resulting neuromagnetic signals observed only at coils 2 and 4 are near-mirror images of each other, with sign dependent on stimulation site. Inhibitory responses are blocked in this preparation.

While hippocampus and cerebral cortex both have an anatomical organization that favors representations of neural activity as equivalent dipoles, it is not the case that in general all neural activity spanning their entire surfaces are be accessible to noninvasive MEG recording. The Wu & Okada study demonstrates an example of location-dependent signal measurement: virtually zero magnetic fields can be recorded along the current axis, as cytoarchitecture only builds up an open field elsewhere. An implication for human studies is that the convoluted cortical surface may constrain which regions are more or less to MEG sensors (Fig. 4A). As a rule of thumb, areas lying within sulci will be oriented parallel to the scalp surface and thus visible, while those domains located on gyri will be oriented normal to the scalp (but Heschl gyrus a notable exception), thus are magnetically inaccessible to sensors directly above. This problem is diminished by the presence of multiple sensors over the head array, some of which may be further away

from the generator, but in adequate relative orientation so as to measure a signal from the radial source in question. This issue of visibility can be modeled by the "forward problem": that is, to map activity from a hypothesized anatomical source, such as the superior temporal lobe, to a field distribution recorded over the sensor array. This minimally requires projecting the field predicted by the laws of electromagnetism upon the spatial distribution of sensors to generate an image of the resulting magnetic field distributions. In real scenarios, it is the latter which is first available to the experimenter (Fig. 4B), for the "inverse problem" of estimating the likely current distribution anatomical source given the magnetic field distribution, has to be addressed.

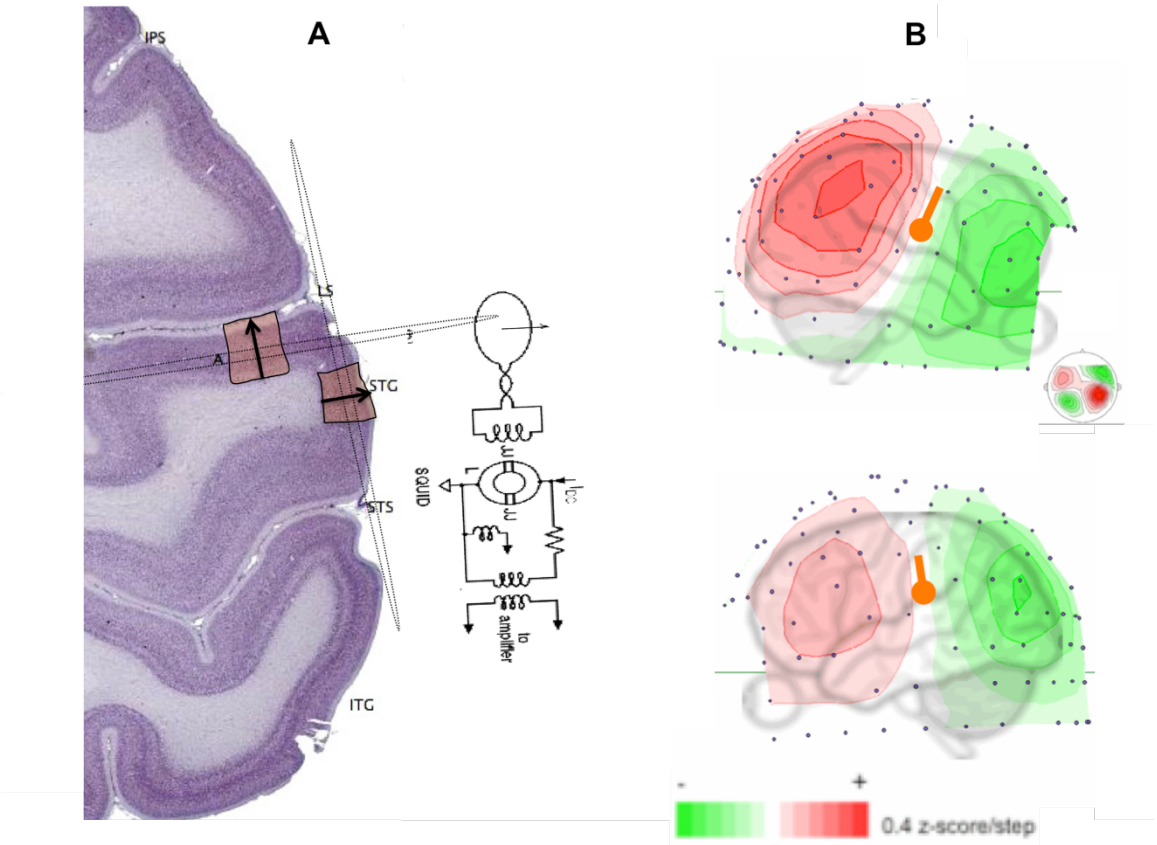


Figure 4. MEG sensitivity and resulting field distributions: forward and inverse approaches. (**A**) A cortical domain organization and location relative to a detector coil

dictates whether neuromagnetic signals may be retrieved from that domain. When apical dendrites across the superior temporal gyrus are active, a collective open magnetic field distribution is generated but little of it will reach outside of the scalp thus rendered invisible to a magnetic field detector at that location. Such neurons are *radially* organized, unlike *tangential* units from domain A, whose open field distribution exits and enters again the scalp, thus able to reach and cross through one detector coil from the MEG sensor set. (**B**) A bi-hemispheric magnetic field distribution, the first activity source map available to the experimenter, arises during auditory stimulation. Red isocontour lines approximate regions of equal magnetic field that cross from scalp into the MEG helmet surface, curling through exterior space and crossing back into the helmet surface (with opposite sign), then back into the scalp. The diagram overlay suggests its origin at the superior aspect of the temporal lobe.

Aside from spatial coherence requirements, open fields also gain strength when generator units are active *with* temporal coherence. Estimates for the minimum amount of excitatory pyramidal units that need to be simultaneously active so as to generate a readable MEG signal, are of the order of between 10^4 to 10^5, resulting from current flow measurement resolution limits of 10 nA·m [5]; the minimum number of units may relate, based on anatomical estimates, to columnar patch ensembles of about 0.6 mm^2 [6] although column densities may vary across different functional cortical areas. Coherence may of course vary over time, and MEG excellent temporal resolution is be used to probe the course of distributed, highly synchronized activation patterns throughout cortical domains. Altogether, the physiological basis of neuromagnetic signals make them amenable for questions of *how and when* do cortical domains covary with experimental timeseries variables.

Neural representations in the auditory system

Finding straightforward encoding relationships or *neural representations* of sensory information has been often done from single unit activity. Information about sensory stimuli may also be encoded in other domains spanning measurement scales other than the single unit, for example both sub-scale membrane fluctuations, and supra-scale aggregate fields such as in in multi-unit ensembles and the M/EEG signal. One consequence of interpreting sensory events within these neural fluctuations is that it extends neural coding strategies from spiking activity to smooth, continuous timeseries.

Encoding and representations: the receptive field

How does the human cortex represent sound? When responses from auditory areas are found to covary with a sound stimulus statistic, the feature in question is said to be *represented* in the neural response. If a set of represented features are found in a way that predicts both behavioral and physiological outcomes – for instance, an encoding predictor arising in both animal models and supporting human psychophysics data (cf. [7]), then it qualifies as one of several possible solutions to the *encoding problem*. The encoding question summarizes as "*how* does a neural domain represent a physical variable *x*?". Part of this question is answered by the extensive tonotopic organization of the auditory system, showing clustering of neighbor neurons by common response selectivity to neighboring spectral tuning regions. Refining the encoding question applied to the

auditory system then to "*when* does a neural domain represent *x*?", the task is then to consider spectral features that dynamically evolve over time. Given a neural domain response, what estimate represents neural selectivity in terms of both spectral tuning and temporal properties such as latency, or periodicity, at the same time?. This description is embodied in the **spectro-temporal receptive field** (STRF), originally obtained by manipulating neuron spike datasets in a procedure known as reverse correlation or triggered correlation [8], [9]. This original method represented single unit selectivity: if each spiking time-series of neural output is reversed in time, then individual spikes conceivably denote time flags or triggers for particular features in the (time-reversed) neural input or stimulus, since it marks the occurrence of a neural event. The collection of such spike-triggered features (equal in number to the amount of spikes in the set) is then the ensemble of stimuli that precedes a neural event, and the ensemble is then summarized by averaging and ordering by frequency, thus representing the correlogram of that neuron's STRF (Fig. 5)[10]. In practice this procedure can be reformulated as a "black-box" operator mapping from sound input to a neural output such as spike activity data, or as later extended, continuous field activity [11], [12]. This latter option allowed examination of implications for neural processing stemming from auditory neural assemblies in the aggregate, such as in MEG.

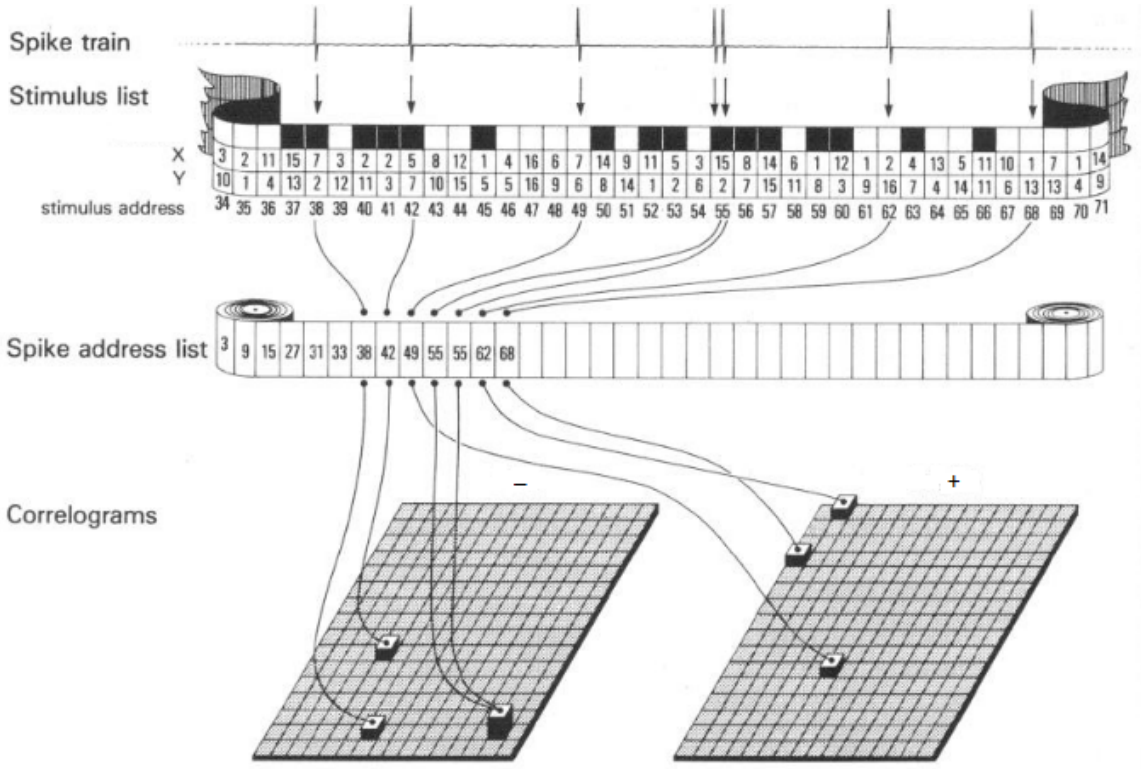


Figure 5. The spike-triggered average method. In this example, a neuron's receptive field is described along two unspecified dimensions X and Y, which parameterize the addresses where both "positive" and "negative" stimuli (black/white) are delivered over the entire experimental session. The positive or negative stimulus delivery at specific addresses in turn associates with generation of one or more spike events. In the image, surface maps of all key addresses are summarized according to the stimulus sign; the spike triggered average is then the difference between maps. Modified from [13].

This systems theory characterization has important limitations, namely that output should not depend on future states, nor or on all past states of the system; that neural domains display nonlinearities as a consequence of simultaneous dependence on several parameters (and/or on several terms of the same parameter); and that responses may depend on high-order factors about the input – arbitrary stimulus categories is a common classic example [14]. A fruitful approach has been is to approximate the system by its

linear representation, with the possibility to add nonlinear terms of increasing order [11] when necessary. When the actual order is unknown (where "order" refers to the structure of present and past dependencies in the system), then the generalized receptive field may require describing an arbitrary number of expansion kernels. In contrast, if prior knowledge of the actual or approximate order has been established, the second-order kernel among the generalized receptive field expansion directly relates to the STRF as a **spectro-temporal *response function***; both second-order kernel and STRF are identical if the system itself is of order two [11], [12], speaking to the severe theoretical restrictions upon of the linear STRF model to capture every aspect of the system behavior. Within limited predictive space however, the potential return of interpretational power may be high given its domain overlapping with a fundamental organization principle of the auditory system

Testing linear models of auditory coding

Only in the unlikely case where the system's response is a mere proportionality relationship between input and output, it can be said to be linear[1], and the actual receptive field is entirely captured by the system's response function. More typically, unknown higher-order nonlinearities in the system will not be captured by the STRF, contributing to a decrease in the predictive power of the model. Therefore, estimation of linear STRF models enable assessment of the extent to which neural output is directly proportional to

[1]This is different from the system's order. In this special case all aspects of a second-order nonlinearity in the system are completely explained by the second-order kernel. In other words, analysis of the system at the second expansion term does not preclude the fact that the system may involve a linear transformation from input to output.

spectral change in the input over time (for this and other interpretations see [15]); if this extent is considerable, then the spectrogram is said to be linearly encoded by the neural domain in question [16]. Among critical issues related to testing STRF model validity is the choice of an adequate stimulus ensemble. For mathematical reasons, random noise stimuli such as Gaussian white noise (GWN) had been instrumental in algorithmic implementations of the original system kernel expansion described above. Failures in generalizing these estimates' predictive ability, namely to implement the model estimate on novel stimuli other than what was used in the original estimation ensemble itself, arise especially beyond peripheral stages [17]. For example, the spike triggered average method is inappropriate for STRF estimation from stimuli beyond white noise, such as natural speech [18] (Fig. 6); rationales behind similar pitfalls are reviewed in [17]. This problem points to the key issue of identifying sound stimulation patterns that may be adequate for the auditory stage in question, as the ascending auditory system shows tuning to increasing levels of spectrotemporal complexity [16]–[18].

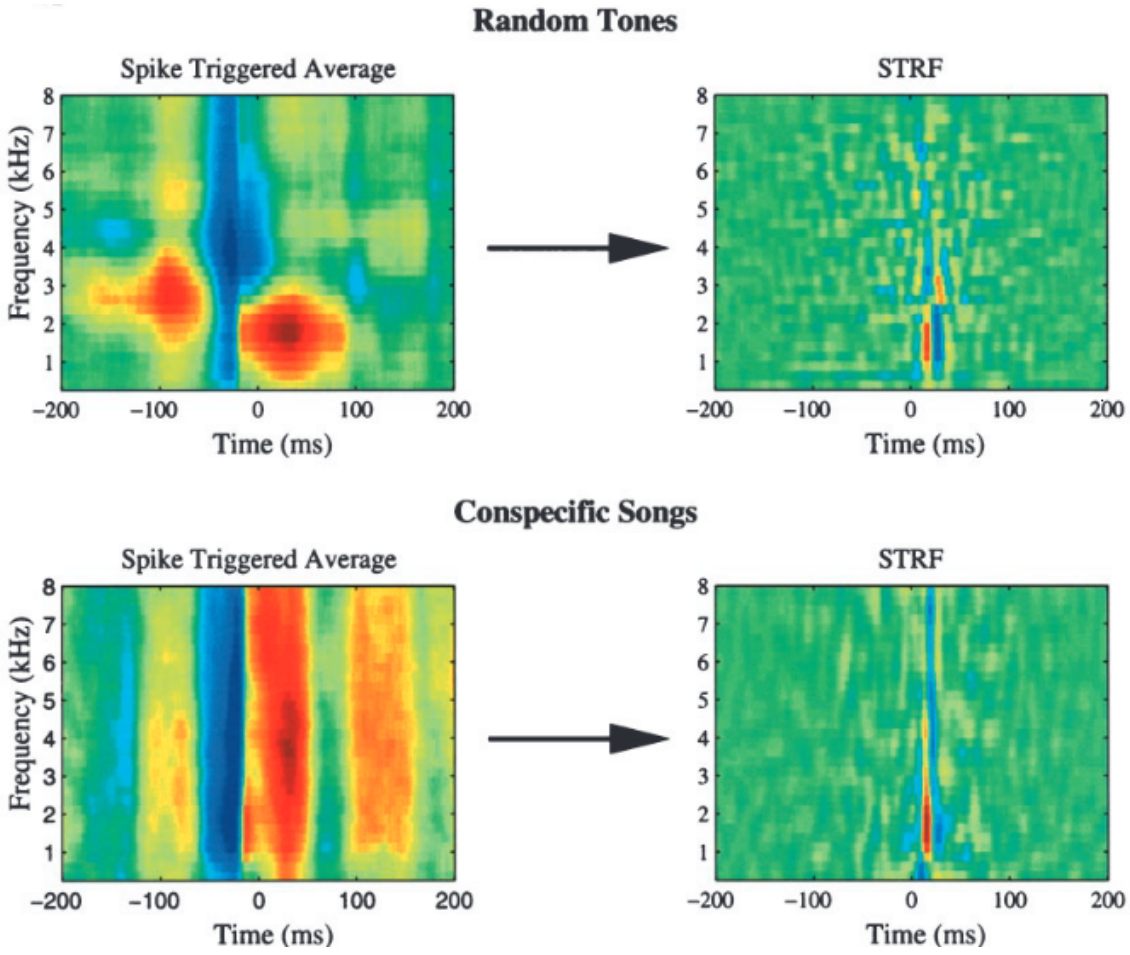


Figure 6. Stimulus ensemble qualitatively affects STRF estimate at higher-order auditory areas. For auditory neurons tuned to complex stimuli, the reverse correlation method yields spurious results under spike triggered averaging, since responses are not driven by uncorrelated stimuli (left column). Strategies accounting for statistical differences between stimulus classes may result in extraction of aspects relevant to the general stimulus-response relationship, which may appear as consistently interpretable STRFs across different stimulus classes (right column) even if from complex-tuned auditory areas, such as illustrated here. Image from [18].

Overall, these observations lead to suggest that the validity of a linear model based on a STRF may be affected not only on the intrinsic nonlinear nature of the system, but also by selection criteria laid by the experimenter, namely the specific neural domain probed

and the stimulus ensemble/class under consideration. In turn, this means that STRF estimates relaying poor predictive power do not necessarily invalidate the encoding relationship between the stimulus class used and the neural site probed; it only suggests that (i) given the encoding model parameters, a linear transform is unlikely to represent deterministically the input-output relationship; (ii) the estimation technique poorly approximates the linear aspect of the system, or (iii) a combination of both.

Estimating linear models of auditory coding

Robust estimation techniques attempt to eliminate the risk of falling under case (ii) above. Among these are estimation optimization models that seek reduction of error between a neural response and the response prediction by the STRF. In one of such approaches, error minimization occurs recursively, via step-wise modifications to the STRF. *Boosting*, an optimization approach in this spirit, belongs to a family of gradient descent techniques applied to auditory data [19]; its name denotes the strategy to run an arbitrary, weak estimation algorithm first, to produce an estimate that will only be required to perform slightly better than chance or than guessing. The same algorithm is then run several times on different instances of the dataset (for example, re-weighted versions of data). Initial low accuracy rates are improved ('boosted') once a single, more accurate estimate is built from the initial poor training outputs in the algorithm by incorporating them into a jointly fitted additive expansion [20]–[22]. Incorporation criteria and the choice of training subsets may vary; the method denoted here as *boosting* corresponds to a forward stage-wise (but not joint) fitting that follows a greedy heuristic, adding the contribution leading to the largest available mean-squared-error reduction at

each given step [19], [23]; indeed such reduction is desirable because it amounts in turn to maximizing the predictive power of the model [24].

Operationally, STRF estimates by *boosting* are initialized as a null matrix of dimensions *T*x*F*, where *T* equals the number of experimental time bins and *F* is the total of frequency bins; optimization follows by exploring fixed increments and decrements per single spectrotemporal bin separately. The exploration yields a total of 2x*F*x*T* possible candidates, among which the best mean-squared-error reduction is selected and accumulated upon to the running STRF in gradient descent (Fig. 7). The procedure is iterated until more modifications introduce undesirable behavior, such as a sustained increase in mean-squared error [19], since the method is not guaranteed to find a global optimum. The final STRF estimate consists then of the history of locally optimal choices added recursively. Formulations of the procedure implementation, along with a description of preventive measures with regards to overfitting (e.g. cross-validation), are available in [19], [25].

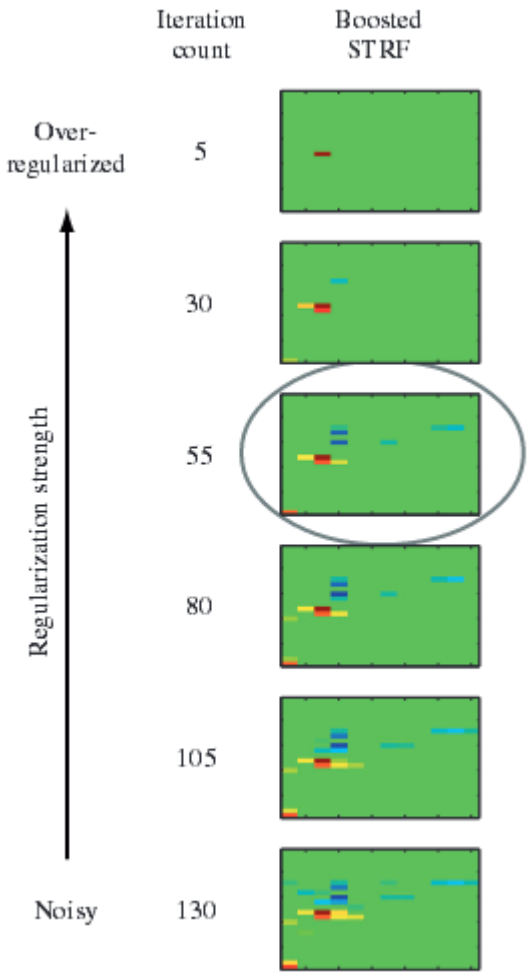


Figure 7. Iterative STRF estimation method via boosting. At each step, only one address in the spectro-temporal map reduces most error between response and prediction, and such bin is kept as a seed for the next iteration. As repetitions eventually lead to noisier estimates, stopping at an intermediate step (circled) prevents further reduction in error due exclusively to statistical properties of the stimulus-response ensemble used for training estimation. Overfitting prevention is done by optimizing with respect to reductions in generalization error involving novel stimulus-response ensembles not used in training. Image from [19].

Chapter II

Functional significance of spectrotemporal response functions obtained using magnetoencephalography

Summary

The spectrotemporal response function (STRF) model of neural encoding quantitatively associates dynamic auditory neural (output) responses to a spectrogram-like representation of a dynamic (input) stimulus. STRFs were experimentally obtained via whole-head human cortical responses to dynamic auditory stimuli using magnetoencephalography (MEG). The stimuli employed consisted of unpredictable pure tones presented at a range of rates. The predictive power of the estimated STRFs was found to be comparable to those obtained from the cortical single and multiunit activity literature. The STRFs were also qualitatively consistent with those obtained from electrophysiological studies in animal models; in particular their local-field-potential-generated spectral distributions and multiunit-activity-generated temporal distributions. Comparison of these MEG STRFs with others obtained using natural speech and music stimuli reveal a general structure consistent with common baseline auditory processing, including evidence for a transition in low-level neural representations of natural speech by 100 ms, when an appropriately chosen stimulus representation was used. It is also demonstrated that MEG-based STRFs contain information similar to that

obtained using classic auditory evoked potential based approaches, but with extended applications to long-duration, non-repeated stimuli.

Introduction

Empirically measured sensory receptive fields and response functions offer analytical characterizations of computations attainable by the auditory system[26]–[28]. Applied linear systems methods such as the spectrotemporal response function (STRF)[11], [14], [29] have similarly led to informative computational characterizations of central auditory neural function with respect to sound encoding and perception[30]. The STRF can be viewed as a representation of the approximate neural response to changing auditory input in time or frequency; any particular functional description will vary according to the location and role of the neurons. Different stimulus classes (e.g. artificially generated sounds vs. natural sounds) may produce related, but dissimilar STRFs from the same neural unit, speaking to fundamental processing differences (and similarities) of auditory encoding[18], [31], [32]. An emerging view in electrophysiology is that the STRF may represent a snapshot of the entire network converging onto that neuron (or group of neurons)[32], incorporating this population's activity in its neural representation of the spectrotemporal features of the stimulus[30]. As seen here, STRFs also have a role in investigations of ensemble auditory coding, using neural recordings obtained from magnetoencephalography (MEG) or electroencephalography (EEG).

STRFs directly characterize the relationship between a sound stimulus and the

accompanying neural response. For neural ensembles, rather than individual neurons, many individual linear components may be jointly pooled, perhaps even superadditively (depending on the underlying neuroanatomy and neurophysiology of the signal source). Also, as in the case of a single-neuron-based STRF, it may be methodologically simpler to use controlled stimuli rather than natural sounds[24], [33], [34]. It remains to be determined how the spectrotemporal features of ensemble-based STRFs correspond to the time-varying evoked-related-potential responses (and other standard MEG/EEG measures) as a function of frequency, and also to what extent the STRF encoding model can provide analogous additional information besides predictive power. Furthermore, the STRF estimate of a stimulus-response relationship may depend on the particular representation chosen for that stimulus; in particular, it remains unknown which specific stimulus representations are optimal for the purpose of matching STRF features to neural function[35], and whether any such choices can address the key question of how to generalize across stimulus classes, from artificial to natural stimuli. Finally, it is important to discuss overlap between these non-invasively obtained STRFs and those available from local field potential (LFP) data or from other invasive recordings.

In order to address these questions, evoked cortical activity recordings from healthy listeners were obtained with MEG during active listening of pseudo-random multi-tone patterns[33], [36] presented at a variety of rates. STRFs were obtained per subject and condition, in order to assess the extent to which the MEG responses were linearly explainable by a sparse representation of the stimulus sound pattern, and whether rate-related changes are consistent with those found using invasive electrophysiological techniques. Peak components in STRFs and temporal response functions (TRFs) were

identified and their latencies compared to those obtained with standard tone-based averaging. Alternative representations of the stimulus, including the auditory spectrogram, were used for reverse correlation in order to constrain the space of stimulus representations given the properties of the MEG cortical signal. Finally, these functionally informative STRFs were compared to those from datasets from studies using natural speech[37] and music processing[38]. This allowed an investigation of ensemble-dependent issues arising from STRF comparisons when using artificial vs. natural stimuli[31].

MEG-based STRFs are shown to functionally explain considerable amounts of response variability while revealing a parsimonious mapping of response features seen in classic averaging methods to those obtained from dynamic stimuli timeseries. Quantitatively, the MEG-based STRFs account for similar levels of predictive power to single and multiunit responses in auditory cortex[24]. Qualitatively, the mappings show reasonable correspondence with those from local field potential activity in animal models[39], [40] and manifest similar stimulus dependencies (e.g., density[33]). We find that similar STRF structure is seen across responses to stimuli as diverse as natural speech and music, demonstrating convergence across stimulus classes. This last result, however, depends on the use of a specific (sparse) representation of acoustic stimuli, the nature of which provides additional knowledge regarding the role of spectrotemporal modulations on predictive frameworks of auditory cortical representations over a wide range of dynamic sound classes.

Results

*MEG cortical responses predictable from the STRF linear model***.** Potential successes of the STRF as a linear model to predict MEG responses from acoustic stimuli are evaluated by comparing the actual vs. predicted responses which, unlike spike-generated STRFs, are continuous waveforms (Fig 1A). Model predictions are obtained by linearly convolving the corresponding STRF with the stimulus representation, using cross-validation (arbitrary separation of training data from testing data) to prevent overfitting, which makes this a conservative estimate due to noise present in the testing data[24], [34]. If instead only the training data is used, i.e., fitting to the same data as is tested, STRF estimates provide a stringent upper limit as to how good any linear prediction can be. STRFs estimated using cross-validation predict the large negative deflections (Fig 1a, red) that follow tone onsets (~100 ms post impulse) well, but unlike those from training (Fig 1a, blue), are less accurate for positive excursions (both data sets summarized in Fig 1b). The ability of the STRF model to predict the encoding relationship between sound patterns and cortical responses can be measured as the fraction of response variability explainable by the linear model, estimated on an individual condition and subject basis, once intrinsic response variability (unrelated to the stimulus) has been removed[24], [34]. MEG STRF predictions were found to range as high as 34% of variance explained across participants and rates, using cross-validated data. When the fraction of variance explainable by the model was compared with normalized noise power (or inverse SNR), the explainable fraction in the theoretical noiseless limit was estimated to be 23.0 ± 2.0% (mean ± st. dev.; CI: 19.0–26.9%) as part of a significant linear regression relationship (F=45.9; p=2.7x10^{-9}; R^2=0.386), with an upper limit of 71% as provided by training-data

only results (Fig 1c).

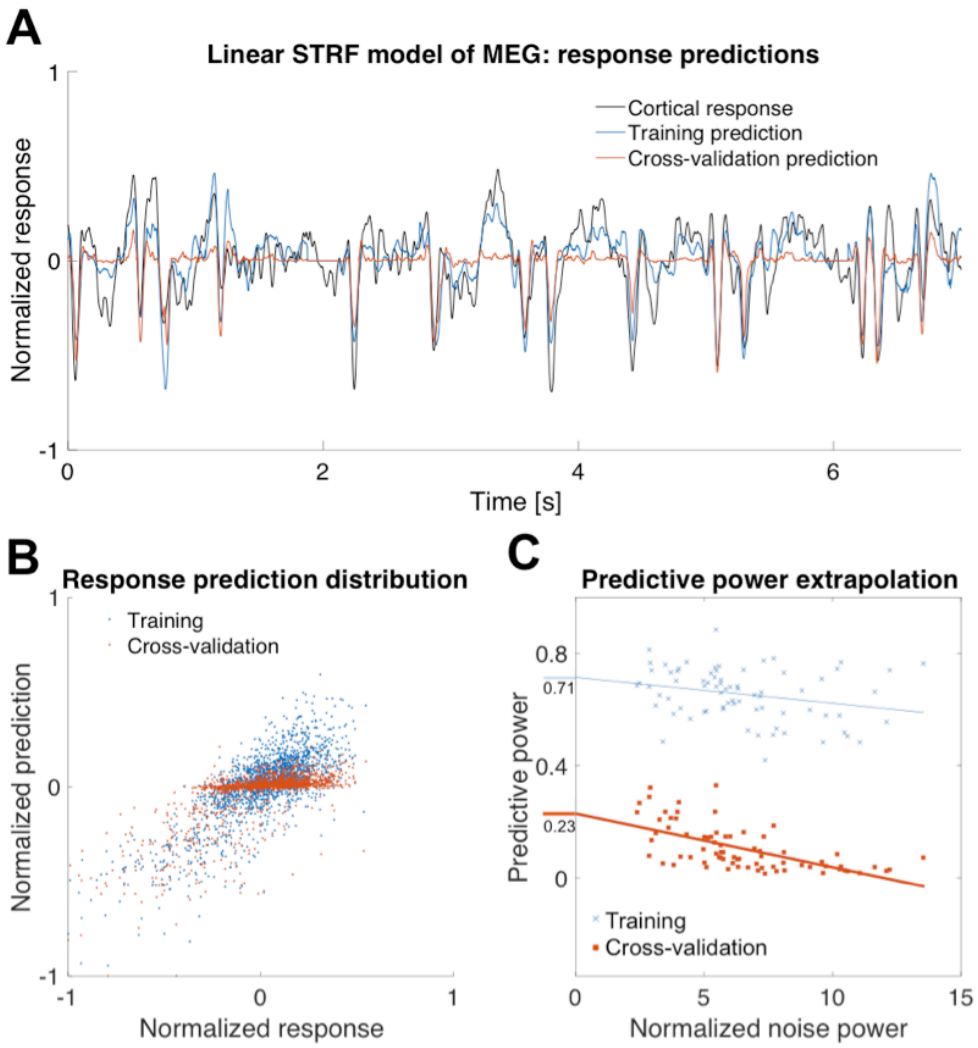


Figure 1. Spectrotemporal encoding models of MEG signals from human auditory cortex. a) A 7 s sample recording of an MEG response to a sparse multitone pattern (2 tones/s), with STRF-based predictions. b) STRFs were optimized by iteratively minimizing prediction error on the entire dataset, referred to as *training* (blue, r=0.74), or alternatively on their ability to generalize (*cross-validate*) over testing datasets (red, r=0.62). c) The predictive power of the STRF models is shown by linear regression of individual STRFs across participants and conditions on their corrected normalized noise power (i.e., inverse signal-to-noise ratio, an indicator of trial by trial reliability, see

Methods). Extrapolation of performance to zero noise power gives a noise-corrected expected performance for both the conservative cross-validation-based estimates and the fundamental-upper-limit training estimates.

Fraction of response explained by STRF features consistent with standard evoked potentials. STRFs based on MEG responses display consistent spectrotemporal structure in the form of positive-negative-positive complex deflections (Fig 2a) coinciding with typical auditory cortical latencies (e.g. those of the P1-N1-P2 complex in averaged EEG responses to isolated tones). In particular, the multitone STRFs demonstrate strong negative responses at ~100 ms post impulse onset ($STRF_{100}$). The specific $STRF_{100}$ latency depends on stimulus frequency, varying ~20 ms over the frequency range 180-700 Hz; at higher frequencies the latency is approximately constant ($STRF_{100}$ latencies for 2 tones/s shown in Fig 2b, black). $STRF_{100}$ latencies were found to follow standard tone-evoked M100 latencies[41]–[46] obtained under various conditions (Fig 2b; also Table 1). The correspondence suggests a quantitative link between the $STRF_{100}$ and M100, and therefore between STRF-based techniques and ordinary auditory evoked cortical potentials. Analyzing the same experimental data using standard evoked response analysis instead (epoching and averaging over responses to all tones in the sparsest multitone pattern) demonstrates strong temporal correspondence at the group level (Appendix A Supplementary Fig 1).

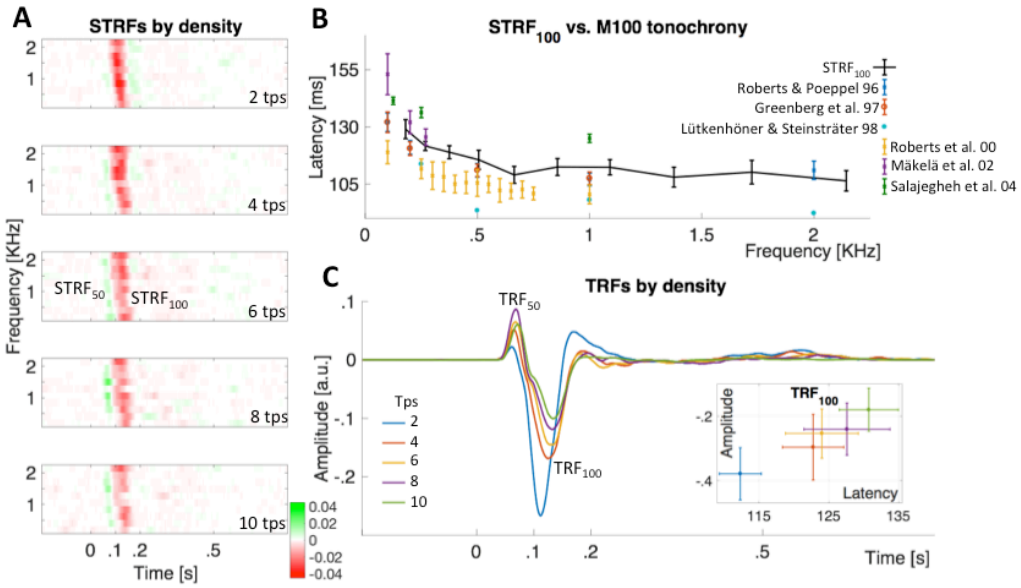


Figure 2. Consistency between response function model predictive model features and evoked potentials. a) Grand average spectrotemporal response functions based on multitone stimuli demonstrate a positive-negative-positive structured sequence between 50 and 200 ms following tone onset; tone cloud density introduces qualitative changes in relative amplitude and delays: at increased rates an early positive component (50-100 ms; $STRF_{50}$) emerges, while the medium latency negative component (100-150 ms; $STRF_{100}$) attenuates, and a late positive component (150-200 ms) present only in the sparsest conditions disappears. b) $STRF_{100}$ components are delayed by over 20 ms as tone carrier frequency decreases from 2 to 0.2 KHz, in a manner consistent with those of evoked potentials in single tone presentations[41], [42], [44]–[46] (Table 1), indicating a correspondence between impulse response functions obtained through reverse correlation and averaged evoked potentials. A common latency decrease across studies and conditions is observed for carrier frequencies in the 180-700 Hz range. c) Temporal response functions, obtained by reverse correlation with the stimulus envelope collapsed across frequencies, show features similar to the P1-N1-P2 complex commonly found in EEG evoked potentials[47]. Higher tone presentation rates result in the emergence of the TRF_{50} and in decreased amplitude and increased latency of the TRF_{100} (inset), as well as the attenuation of a later-latency positive deflection. Error bars are 1 standard error of the mean.

	# of subjects [mean age]	Sound delivery [Sensor location]	Tone duration (ms)	Presentation rate [tones/s]	Peak finding method
Roberts & Poeppel, 1996[41]**; Greenberg et al., 1997**[42]	5 [24-33 y]	Monaural [Contralateral, Left]	400	0.7 - 1.3	Equivalent dipole, Maximum RMS
Lütkenhöner & Steinsträter, 1998[43]	1 [28 y]	Monaural [Contralateral, Left]	520	~0.4	Maximum RMS
Roberts et al., 2000[44]	8 [-]	- [Both hemispheres]	-	-	-
Mäkelä et al., 2002[45]	11 [32 y]	Binaural [Both hemispheres]	200	1	Optimal sensor pair
Salajegheh et al., 2004[46]	11 [45.8 y]	Binaural [Both hemispheres]	400	0.8 - 1.3	Maximum RMS from optimal 12 sensors

Table 1. Comparison of studies reporting M100 absolute latency values in response to pure tones, with participants, recording mode, stimulus details, and M100 peak determination method where available.

To further investigate the correspondence between STRFs and evoked potentials (specifically the effects of tone density), reverse correlation was performed with respect to frequency-collapsed representations of the stimulus, generating the frequency-independent Temporal Response Function (TRF, Fig 2c). The ~100 ms latency negative peak (TRF_{100}) amplitude decreased with increasing tone-density by ~60% across modulation rate range studied, while latency increased 20% (see inset). In contrast, the

~50 ms latency positive deflections (TRF_{50}) had the smallest amplitude for the sparsest multitone condition. Thus sources with ~50 ms latency generate a strong increase in cortical activity with a transition from scattered to continuous pure tones, while sources with ~100 ms latency decrease in strength as they are delayed. Cortical activity in sources with 150 ms latency may also be active, provided the inter-tone interval is long.

STRF most informative for onset-based representations of multitone stimulus. Methodologically, the acoustic representation of the stimulus used to generate the STRF may employ any number of available time-frequency representations of the sound, including the widely-used spectrogram[19], [24], [48], [49]. One reason to consider alternatives to the spectrogram is to compare STRF features with evoked response features, since an evoked response to tones is calculated not with respect to the spectrotemporal duration of the tones but only to their onsets. Thus analyses also included binary and sparse representations of the stimulus: single tones were modeled as trigger-like impulses timed with tone onset and organized by frequency. Indeed, stimulus features that are known to be encoded by auditory cortex include onsets, offsets, and stimulus duration (in the form of sustained responses)[50]–[53]. Since the MEG signal is aggregated across synchronized individual neurons[6], evidence for those same encodings requires investigation. Reverse correlation techniques are well suited for this larger-scale analysis because it explores the outcome of alternative stimulus representations that emphasize such features. The stimulus representations tested here (cf. Fig 3 insets) were (i) the ideal *trigger* representation, (ii) the ideal *edge* representation (both onset and offset triggers), (iii) the ideal stimulus first-order *derivative* (onset and

negatively-signed-offset triggers), which can itself be used to generate the trigger representation if followed by half-wave rectification, (iv) the ideal stimulus *pulse* envelope, which has constant value from onset to offset (and which can itself be used to generate the previous representation if followed by differentiation), (v) the actual acoustic stimulus passed through a filterbank with identical center frequencies as the tone, whose *envelope* is then extracted (see *Methods*), and (vi) a generalized *envelope onset* representation obtained via half-wave rectification of the previously defined filterbank envelope output. Only the last two can be applied to natural (non-discrete) stimuli, and so are especially important in later sections.

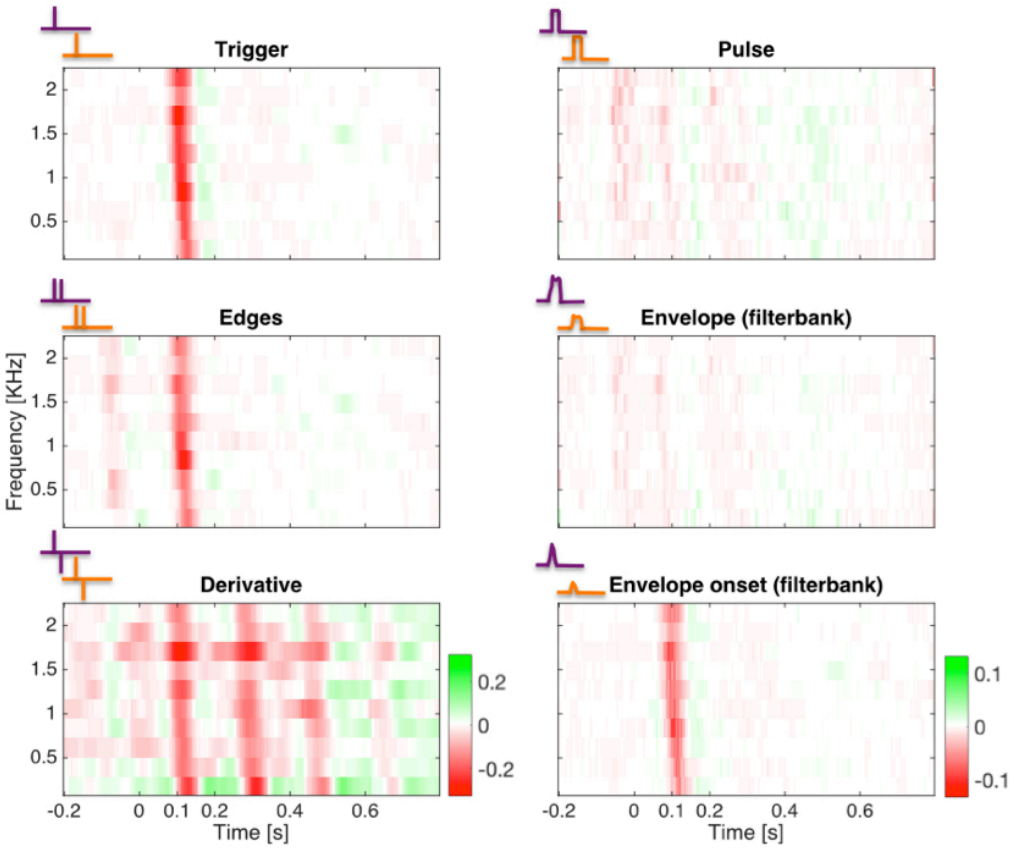


Figure 3. STRFs generated using different stimulus representations achieve different levels of functionality. STRFs generated from multitone patterns are functionally informative (e.g., comparable to evoked potential analysis) when each individual tone is discretely represented by its onset (top left) but not when represented instead by the timing of its temporal edges (middle left), sign (bottom left), or a discrete representation of the entire pulse duration (top right). Related to the spectrogram, the representation based on passing the acoustic signal through a series of filterbanks, then extracting envelopes per band (middle right) yields only barely discernible results. Extracting onset timing information from the filterbank, in contrast, was quite functionally informative (half-wave rectification of the first derivative of the filterbank output; bottom left). Critically, filterbank-based methods do not require *a priori* definitions of temporal edges and can be used for arbitrary stimuli. Color scales as in bottom right inset, except for Derivative STRF.

Grand average STRFs in Fig 3 demonstrate that among such representations, only those expressing tone onset events explicitly yield components comparable to those of evoked potential analysis (first and last STRFs of Fig 3); STRFs from the alternatives introduced ringing and/or pre-causal artifacts. As with the original onset-based trigger representation, reverse correlation with both temporal edges predicted activity from the first edge in accordance with the latency by frequency dependence, but also produced a pre-causal mirror component, in advance of the original and at the tone duration distance. This pattern suggests that tone offset was not explicitly encoded here. This interpretation is supported by analysis of STRFs generated by the derivative representation, which correspondingly flips the sign of the same pre-causal mirror component, but is additionally contaminated via constructive interference by a series of artefactual ringing cycles. The pulse representation, which can be viewed as an idealized envelope, produces

STRFs that are essentially featureless (or at best, whose features are barely discernible above the noise floor). This result is unexpected since typical auditory reverse correlation studies use a duration-based stimulus envelope representation[25], [54] and the temporal envelope is often hypothesized to be the response-driving feature. Similarly, the acoustic envelope representation (using a filterbank model; see *Methods*) also produced featureless STRFs. An attempt to re-create the onset representation (i.e. half-wave rectification of the acoustic envelope representation derivative), did however generate STRFs with features comparable to evoked potential analysis, and enabls the extraction of onset-like information in general from diverse complex natural stimuli. Because of the remarkable agreement between the idealized and the acoustic onset models, interpretations based on evoked potentials may extend to reverse correlation analysis applied to other stimulus classes where definitions of onsets would be *a priori* unknown or not controlled for, such as natural sounds.

Convergent STRF models across artificial and natural stimuli. Because of its potential to reveal hierarchical processing mechanisms, a major goal in auditory reverse correlation has been to examine the encoding relationship for critical natural stimuli including speech and other communication sounds. To this end, datasets from two previously unpublished studies on speech and music processing were submitted to the same analysis methods as the multitone pattern (Fig 4a), with stimuli represented by their envelope onsets. As with the onset-based representation of the multitone patterns, STRFs for speech and music exhibited qualitatively similar structures, with distinctive biphasic components near 50 and 100 ms post rising transient impulses (onsets) along the same investigated spectral region. Inspection of the stimuli under either envelope or envelope onset

representations suggests that the latter procedure effectively increases similarity in the underlying distribution across stimulus classes (Appendix A Supplementary Fig 2a). The frequency dependence of relative peak delays was also maintained for these stimulus classes (Appendix A Supplementary Fig 2b) but with class-dependent timing differences, suggesting a common fundamental mode of spectrotemporal cortical processing up to ~200 ms and after which notable processing differences appear according to the stimulus class. While neural data from all studies were obtained from different subject groups, one subject did participate in those two studies and in a modified pilot version of this experiment (2.4 tones per second presentation rate); these data are presented in Fig 4b, again showing strong qualitative similarity both to group data and class-dependent timing differences. This subject's topographic magnetic field maps associated with the neuromagnetic signals derived in each of the three studies are displayed in Fig 4c; mapping each STRF to overlapping spatial distributions is consistent with source activity at the superior aspect of the temporal lobes.

To better illustrate class-dependent temporal differences across the studies, TRFs were obtained by collapsing STRFs across spectral bins, as shown in Fig 4d. These plots emphasize spectrally consistent changes in temporal processing due to stimulus class, along with relative amplitude differences. As before, early activity appeared least prominent for the spectrotemporally sparsest stimuli; in the case of the single participant tested across all three stimulus classes, a high-temporal resolution analysis of the multitone TRF_{100} shows its dynamics are very close to those of the speech envelope

counterpart (Appendix A Supplementary Fig 3a), with characteristic time constants of ~3 ms (Appendix A Supplementary Fig 3b). The response dynamics for music, however, do not follow similarly, which suggests that features other than overall acoustic onsets may contribute to synchronized auditory responses in these cortical populations.

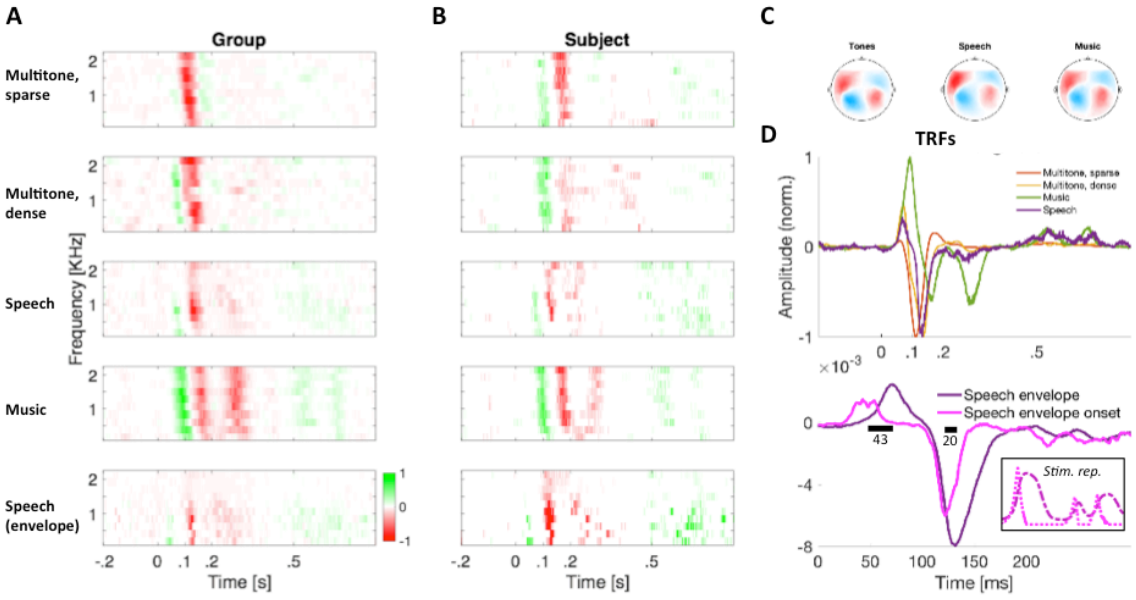


Figure 4. Interpretational power from stimulus representations across STRFs from different stimulus classes. a) Group normalized STRFs from the multitone pattern experiment (*N*=15), and from studies on natural speech (*N*=12) and on music (*N*=15), reveal considerable structural similarity when stimulus onset is extracted as a driving feature of the neuromagnetic response. b) Neuromagnetic STRFs from the same participant across the tones, speech, and music studies, which show substantial consistencies across stimuli when represented by their temporal envelope onsets per frequency band. c) The topographic distribution from same subject as in (b) revealed strong bilateral consistency across classes but with increased left hemisphere-bias during speech processing. d) Top: Timing of major neuromagnetic activity peaks, as shown by TRFs derived from spectral integration of the STRFs in (a), results vary depending on stimulus class and/or context: earliest positive and negative deflections change with increasing acoustic density but also with additional spectrotemporal complexity as found

in natural speech and music. Bottom: Group TRFs comparing both speech envelope and envelope onset related activity. Timing differences are explainable by differential acoustic representation in early (< 0.1 s) but not late activity peaks, suggesting the formation of higher order neural representation of elements in speech acoustics by ~100 ms. Only the first deflection timing difference is explained by slope-to-maximum time differences between stimulus representations (inset, same color coding). Curves smoothed by a 5 ms moving average.

Cortical transformation of natural speech envelope representation. In reverse correlation analyses, exploration of alternative representations of the stimulus may provide complementary insight into the functional operations by the auditory system. Fig 4a and Fig 4b show that for natural speech, STRFs based on the acoustic envelope (row 5) led to functionally informative STRFs, consistent with prior approaches[25], [55]. STRFs based on the envelope *onset* representation (row 3) are similar, which is expected since the envelope onset is correlated with the original envelope. In terms of timing, the corresponding group TRFs (Fig 4d) show a difference of 43 ms between TRF_{50} peak components. This was found to be the same as the characteristic delay between their underlying representations, obtained by cross-correlation of the stimulus representations. Such a close correspondence is evidence that at the level of the neural source of the TRF_{50}, an increasing acoustic envelope operates as a fundamental auditory feature of the stimulus. In contrast, the corresponding comparison of $STRF_{100}$ peaks across the two representations (envelope and envelope onset) gives a much reduced difference of 20 ms (Fig 4d, S5 Fig), not consistent with the acoustic differences between the corresponding representation peaks. Compression in components' relative delays were observed across

spectral bins (Appendix A Supplementary Fig 2b), as well as in individual temporal response functions (Appendix A Supplementary Figs 3a, 4).

Discussion

The present investigation describes STRFs as a series of response function mappings from artificial and natural sounds to auditory neural responses. It has been demonstrated that these STRFs possess similar predictive power as their single-unit cortical counterparts, and, importantly, show strong similarities across stimulus classes when an acoustic envelope onset representation is chosen. Specific choices of spectrotemporal stimulus representations[35] result in STRF models that are not only predictive but whose temporal structure is highly consistent with that from standard evoked potential components.

Comparison to spike-based spectro-temporal receptive fields. The spectrotemporal *receptive field* can be considered a spike-triggered averaged spectrogram, from auditory periphery[8], [14] or central nervous system recordings[27], [56]–[58]. Since reverse correlation is a more general principle than spike-triggered averaging[9] it has been used here to characterize and predict the neural responses of auditory systems where both input and output are continuous time-series[12] via the underlying *response function* of the system. Whether measured by spikes or continuous neural responses, neural systems are non-linear, so predictive linear models of central neural coding are necessarily incomplete descriptions of the underlying coding relationship and are bounded by the predictive power and interpretability they maintain within the limits of the linear

regime[24].

The multitone stimulus employed here is comparable to a dynamic random chord stimulus[24], [59], [60] though it has more temporal degrees of freedom, allowing cross-frequency overlap in a continuing pattern that prevents constant tone presentation rates. It is more similar to dynamic random chords than other artificial stimuli used to estimate STRFs, such as ripple noise and moving ripples[17], [61], which focus on stimulus modulations instead. The predictable fraction of variance in the evoked MEG source timeseries was found to be 19–27%, in close correspondence with 18%[24] and 31%[62] predictive power from primary auditory cortex (A1) single/multiunit responses. Comparisons regarding predictive power (and other STRF properties) should also take into account fundamental differences in the underlying signal (spiking versus dendritic-origin activity) and its scale (neuron or highly local population versus meso-scale cortical patches[6], [63]), the animal model, and state (e.g., performing a task vs. resting vs. anesthetized)[30].

Qualitatively, the STRFs presented here exhibit a general broadband structure with frequency-dependent latencies and amplitude changes depending on stimulus density. Remarkably, similar properties appear in STRFs obtained from LFP in mammalian A1[39], [40], [64], featuring broadband inhibitory-excitatory component sequences and, often, frequency-dependent latencies. Component latencies in mammalian A1 are ~50% shorter than here, which may be explained by reduced equivalent cortico-cortical transmission length delays[65] for the species involved in those studies. With respect to human studies, the component latencies reported here are consistent with multiunit activity[66] and high-gamma activity in

electrocorticography (ECoG)[60] in the functional equivalent of A1, Heschl's gyrus. The STRFs obtained in such datasets principally reflect neural spiking, resulting in mappings with narrow-band features, consistent with their interpretation as units locally sampled along the tonotopic gradient of A1. Indeed, frequency selectivity becomes reduced for local field potential recordings[39], [64] (i.e. ECoG frequencies below high-gamma) as they sample redundant activity across distant recording sites with intra-cortical interactions[67] – which may effectively smooth the spectral selectivity distribution[39]. Unlike local recordings, which due to high-frequency selectivity can require that receptive fields be realigned by best frequency[49] to extract statistical features, MEG STRFs offer distributed access to more global cortical network domains. Plausibly, analog results may be expected from future human auditory LFP STRF studies from invasive procedures, given that these have typically focused on multiunit and high-gamma activity[60].

In addition, these MEG-based tone-generated STRFs show stimulus-dependent differences as seen elsewhere in the receptive field literature (see review by Eggermont[30]), namely, amplitude decreases with density. This is consistent with awake primate results, where three-fold increases in tone presentation rate (9.7 to 31 tones/s) may be accompanied by a magnitude decrease of about a third in the STRF maxima[33]; here, a similar multiplicative change in tone density (2 to 6 tones/s) produced a peak decrease of about half. Suppression of excitatory contributions[36], or emergent inhibitory activity throughout A1 single units[33] have been proposed as mechanisms for response field modulations observable from LFP recordings[64]. In cortical neurons, increased firing rates may accentuate depression rate imbalances

between excitatory synapses and those with increasingly inhibitory activity[33], [40], [49], which is a known factor involved in receptive field modulations in somatosensory[68] and visual areas[69]. For auditory recordings, more inhibition may effectively increment responses' spectral specificity or bandwidth at higher tone densities[33], [64] – the analog of which was not observed in the MEG STRFs (see Westö & May[70] for a cautionary note on interpreting inhibitory contributions to STRFs following dense stimuli). Among factors reducing STRF predictive power is the increase of inhibitory fields in estimates from single unit recordings[33]; in MEG this effect appeared to be mirrored by response function components of opposite sign to the $STRF_{100}$. Further research is thus necessary concerning the coarse-grained level of analysis that is accessible via MEG/EEG respectively, in comparison to that afforded by single/multiunit signals.

Association with auditory evoked potentials. Unlike traditional averaging methods, reverse correlation involves continuous delivery of a dynamic stimulus in order to generate a predictive model (of novel instances of the same sound class). It has been shown here that STRFs and TRFs can be directly compared to standard auditory evoked responses, namely the magnetic M50, M100, late auditory evoked responses, and the P1-N1-P2 complex in EEG[47]:

(i) The earliest positive-polarity component, the $STRF_{50}$, seen at higher multitone densities, is a temporal analog to the M50 response originating from Heschl's gyrus (including core/primary areas)[71], [72]; its amplitude may also be modulated by inter-

stimulus intervals[72] at low presentation rates (<2 tones/s). Known modulators of M50 amplitudes include harmonic versus noise-like bursts[73], [74], prepulse inhibition[75], and automatic processing of redundant information as a form of sensory gating in paired-click stimulus designs[76], [77]. In terms of predictive power, this component did not generalize well over novel instances of the multitone random pattern; this is consistent with an adaptive role contributing to considerable changes in the response profile dependent on the local context on the order of a few seconds or less.

(ii) The subsequent major component, the negative-polarity $STRF_{100}$, exhibited magnitude decrease and delay increase with density, with a sharp transition after the sparsest density level. Suppression of the M100 response from supratemporal cortex has been observed in the transition from low to higher tone presentation rates[72] highlighting the interpretation of increased inhibitory effects that include generalized refractoriness among neurons at denser conditions[78]. This component is also subject to attentional modulation[79]–[81] which may reflect that individual tones in a densely populated scene fail to capture attention individually. Because of this component's involvement in tracking perceptual objects of an auditory scene[55], and of the increasing quality of flow and continuity in these artificial stimuli, sharp transitions in this component may suggest indices of 'crowding' relevant to the figure-ground separation problem[82], [83]. Accounts of spectrally-dependent latency in the evoked M100 components[41]–[46] were consistent at this stage and fall within the sensitivity domain for human voice pitch production and discrimination[45]. These latencies are also consistent with those of pitch-specific onset responses, whether elicited by complex tones or by centrally-generated Huggins pitch percepts[84], [85].

(iii) The second major positive-polarity peak appears only in response to the sparsest stimulus. Auditory event-related potentials at ~200 ms latency have been described in EEG as expectancy indices, exhibiting greater amplitudes for tones whose presentation in time is uncertain[86]. At denser conditions, shorter inter-stimulus-intervals may reduce the tone-evoked analogous EEG P2 component, regardless of presentation within a repetition sequence or as an oddball, suggesting involvement of modulation mechanisms other than habituation[78].

On alternative representations of stimulus state. In addition to their predictive power, STRF profiles are functionally informative in a way similar to trial-averaged evoked responses to isolated stimuli[15], [18], [35]; this was the case for STRFs compared across stimulus classes when the stimulus representations were filterbank-derived onsets. Other abstract representations of this stimulus pattern, including both temporal edges, their directionality, the duration of sustained acoustic energy, or, related to the latter, the spectrogram, did not appear to be similarly functionally informative – even though they contained and extended information from onset representation. Nevertheless, predictive power was similar across representations, suggesting that this metric alone is insufficient to expose which aspects of the stimulus map to the system's response. More complex tone patterns might allow predictive power to become more informative regarding the statistical characterization of a stimulus (c.f.[87]). The lack of evidence for explicit neural encoding of offsets is in accord with neurophysiological evidence suggesting offset-encoding cells to be outnumbered by onset cells, and/or to have minor neural response profiles relative to onset encoders[50], [51], which in the aggregate would result in differential contributions to the neuromagnetic response.

On extension to natural stimuli. Processing of environmental sounds, including conspecific calls, is a critical auditory task. Encoding models incorporating natural sounds with complex spectrotemporal structure provide powerful computational insights into the auditory system that may be inaccessible with synthetic stimuli only[18], [30], [31]. STRFs derived from invasive recordings from A1 perform similarly in terms of predictive power, using random tone chord stimuli, animal calls, environmental sounds, sound effects, and music[24], [48], at the population level. For some subset of these neurons, successful linear encoding of the spectrogram may also occur in the same unit for *both* artificial *and* natural vocalization encodings[32], [49]. The search for predictive models that generalize over novel stimuli not in the training set has proven difficult however[32], [59]. The temporal statistics intrinsic to natural sounds may be critical[32], and some evidence from A1 STRFs demonstrates higher predictive power using conspecific vocalizations that are not dilated or compressed[88]; similarly, comparisons are also favorable for artificial and communication sounds controlled for the span of their temporal and spectral modulations jointly, but allowing differences in their amplitude fluctuations over time[32]. Observed stimulus-class dependencies in STRF spectrotemporal properties appear as small time-shifts of STRF features, plus the emergence of additional late activity for speech and music. Analysis of such differences suffer from confounds arising from statistical non-uniformities among the sampled classes[18], [19], [31], [32] and fully addressing this issue is beyond the scope of this investigation. The question of whether detailed class-dependent temporal coding frameworks may be achieved by means of linear methods remains open.

On speech-derived STRFs. Advances in understanding cognitive processes relevant to speech processing have followed from reverse correlation studies that used the speech acoustic envelope (as represented by low-frequency, 1-15 Hz fluctuations in ECoG[89], [90] and MEG/EEG[25], [55], [91], [92] recordings). We find that the speech envelope $STRF_{100}$ component exhibits similar spectral-dependent latency as the M100 evoked response, thus suggesting a level of speech analysis that still contains independent spectral information. Although in contrast with findings of near-constant M100 latencies for certain synthetic vowels presented in isolation[45], reverse correlation methods over long natural speech presentations are better suited to probe domain-general processing in realistic conditions due to their extended sampling,.

Additionally, the methods may constrain the time course of the change in neural representations of human speech from spectrogram to higher level. The low-frequency speech envelope and its onsets are both operationally related to functionally informative STRFs. The systematic delay between timeseries (peaks in the latter systematically precede those in the former) directly accounted for the relative difference between the resulting pair of $STRF_{50}$ components after each representation. The acoustic mismatch could not explain, however, the reduced relative difference between subsequent $STRF_{100}$ pairs. The interpretation of a compression is consistent with current models of step-wise speech processing, where the formation of speech analysis units or objects is preceded by an earlier spectrogram-like representation of acoustics completed by ~80 ms post speech impulse onset. After this time, response functions did not account for the expected

mismatch, suggesting a neurally-based progression into a modified stage of the neural representation of speech and adding to a body of MEG evidence for a cortical hierarchy of speech object representations (see review by Zhang and colleagues [93]).

Overall, these results demonstrate an important advantage of STRFs over standard epoch-averaging methods commonly when used in MEG applications, e.g., characterizing the phenomenology of disorders in clinical populations[94]: their ability to generalize to critical sounds beyond pure tones, most importantly natural speech. By providing both neural predictions and functional information, it allows noninvasive approaches to understanding developmental[95], learning and associative effects induced by tasks[96]–[98], or behavioral contexts[99], [100] – thus potentially furthering insight into the role of dynamical representations of sound in auditory cognition.

General methods

Participants. 15 subjects (6 women, 23.2 ± 2.9 years of age [mean ± SD]), 1 left-handed[101], participated in the multitone study. 12 subjects (6 women, 24.1 ± 3.0 years of age), all right-handed native English speakers, participated in the speech study. 15 subjects (5 women, 21.0 ± 1.7 years of age), all right-handed, participated in the music study. Each subject received monetary compensation proportional to the study duration (approximately 1.5 hours). Subjects had no history of neurological disorder or metal implants. The experimental protocol was approved by the UMCP Institutional Review Board and before each study session, informed written consent was obtained from the

participant.

Stimuli. Multitone study. Sound stimuli were constructed with the MATLAB® software package (MathWorks, Natick, United States) at a sampling rate of 44.1 KHz, and consisted of 50 s auditory scenes composed of pseudo-randomly presented 180 ms tones, each with frequency f_i taken from a pool of 10 fixed values (range: 180-2144 Hz) in 2 equivalent rectangular bandwidth (ERB) steps[102] specified by $f_i = f_{i-1} + \left(24.7(1 + 4.37 f_{i-1}/1000)\right)$. For each frequency, tone onset times were uniformly distributed with a minimum inter-tone gap of 40 ms. Five tone presentation rates (2, 4, 6, 8, or 10 per second over all channels) were used separately. Tone onset times T_j were independent across frequency bands and selected in 20 ms bins. Individual tones were modulated with 10 ms raised cosine on- and off-ramps. Tone level was calibrated according to frequency based on the 60-phon normal equal-loudness-level contour (ISO 226:2003) in order to adjust for perceived relative loudness differences; relative gains to a 1 KHz reference were determined in 2 dB SPL steps. *Speech and music studies*. For the speech study, a 60 s female voice audiobook excerpt [103] narrated from *The Light Princess* (Macdonald, 1864) was used as part of a related study on reverberant speech processing [37]. For the music study, 55 s samples across 6 different instrumental musical styles reflecting a variety of genres and traditions, were presented: orchestra, *Symphony in F Major, No. 32, Movement I* (Sammartini, c. 1740); swing, *Cascades* (Combelle, c. 1940); blues, *Blues for B&W* (Rogers & Hilden, 2003); sarangi, *Raga Mishra Bhairavi: Alap* (Narayan, 2002); pipa, *Dance of the Yi People* (Huiran, c. 1960); and a euphonium transcription of *Dancing Night Wind* (Benning, 1997).

In all studies, audio signals were normalized and presented through the Presentation®

software package (NeuroBehavioral Systems, Berkeley, United States), using audio equipment equalized to a transfer function approximately flat from 40 to 3000 Hz. Sound stimuli were transmitted to subjects via ear insertion tubes E-A-RTONE® 3A of 50 Ω impedance and E-A-RLINK® disposable foam intra-auricular ends (Etymotic Research, Elk Grove Village, United States) that were inserted in the ear canals.

Experimental design. For the *multitone* study, trials consisted of a main tone cloud pattern scene presented in series with per block, generated anew per each subject. This resulted in trials that contained between 0 and 3 multitone density transitions within the trial, and ranged from 70 to 120 s duration. Each of the five main scenes were repeated 4 times, and only these data epochs were analyzed. After a brief training session, subjects were instructed to attend to the ongoing stimulus with their eyes closed and to report rate transitions via a button press. Optional rests were available every 5 trials, totaling 1.5 hours recording time. Subjects received feedback on the correct number of transitions at the end of each trial. For the *speech* study, trials consisted of various story passages presented in random order at different reverberant noise levels. At the end of a trial, subjects were asked comprehensive questions about the passage, and rated its intelligibility. For the present study purposes, analysis was based exclusively on reverberation-free, no-noise ('clean') trials, repeated 3 times across the experiment. For the *music* study, trials consisted of each of the 6 samples presented individually in random order. At the end of each trial, a 5 s clip taken from the same or a different piece was presented and subjects identified if it was an excerpt of the preceding trial. Each sample repeated 3 times across the experiment.

Neural data recording. Magnetoencephalography (MEG) data were collected with a 160-

channel system (Kanazawa Technology Institute, Kanazawa, Japan)[104] inside a magnetically-shielded room (Yokogawa Electric Corporation, Musashino, Japan) at a sampling rate of 1 KHz. Superconducting quantum interference device (SQUID) sensors (15.5 mm diameter each) were uniformly distributed (~25 mm) inside a Dewar vase containing liquid-He refrigerant, with a concave outer surface fit to the average human head. Sensors are first-order axial gradiometers with 50 mm separation and sensitivity greater than 5 $fT \cdot Hz^{-1/2}$ in the white noise spectral region (> 1 KHz), except for three additional reference magnetometers separated from the neural sensors and arranged orthogonally to each other. A 1 Hz high-pass analog filter, a 200 Hz low-pass analog filter, and a 60 Hz analog notch filter were applied online respectively. Sensor channels with saturating or zero responses over more than 12.5 s recording time were excluded from analysis. Participants laid supine inside the magnetically shielded room and were asked to minimize body movement, particularly from the head.

Neural data processing. *Environmental noise*. To eliminate environmental magnetic noise contributions, time-shifted principal component analysis[105] (TS-PCA) was applied, a process that discards optimally-filtered environmental signals recorded on the reference sensors. Reference sensors were 3 physical magnetometers (see *Neural Data Recording*) plus 2 virtual channels obtained by independent component analysis[106] of the remaining data sensors and selecting the two components with the most unstructured broadband (0-500 Hz) power. *Sensor-specific noise*. Electronic sensor noise was removed via sensor noise suppression (SNS)[107] by substituting each channel signal with its projection onto the orthogonal basis space generated by all other sensors in the system. This method exploits redundant activity across elements of the dense array (where the

number of channels exceeds the number of brain sources of interest) by attenuating components specific to any single channel. *Spatial filtering*. Data-driven spatial filters were derived per participant using responses evoked by repeated trials in each of the respective studies. Response epochs of 45-55 s duration were extracted, band-pass filtered (1-15) Hz with a 2nd order Butterworth filter, and delay corrected (~13 ms). A linear transformation based on this manipulation was obtained per participant[108] to generate spatial filters that correspond to magnetic fields generated by the left and right auditory cortex (Appendix A Supplementary Fig 3). This spatial filter was applied to the raw data and the resulting neural signal, representing the most reproducible component of the evoked data, was selected as a single virtual sensor in analyses henceforth.

Neural data analysis. *Spectrotemporal response function of stimulus representation*. For multitone patterns, pure onset representations only carry information at a time beginning with the onset of a tone. We formulate this representation as

$$O(f,t) = \delta_{t-T_{ij}}\delta_{f-f_i} \quad (1)$$

where every onset has equal weight independent of its tone's frequency band f_i (i = 1,…,10), with onset timesT_{ij} T_{ij} of the *j*-th tone with frequency f_i; δ_n is the discrete unit impulse centered at sample n. The input-output relation between this representation of auditory input and the evoked cortical response $\bar{r}(t)$ is then modeled by a spectrotemporal response function (STRF). For discrete data this linear model is formulated as:

$$r^*_{pred}(t) = \sum_f \sum_\tau STRF(f,t)O(f,t-\tau) + \varepsilon(t) \quad (2)$$

where $\varepsilon(t)$ is the residual contribution to the evoked response not explained by the linear system. Summing only over the frequency term allows evaluating the temporal profile of

the response function model (TRF). Exploration of alternative stimulus representations requires substitution of the $O(f,t)$ term in (2) by the analogous time-frequency representation of the stimulus (e.g. by a spectrogram $S(f,t)$).

For all stimuli, *stimulus envelope filterbank* representations were obtained by passing the original waveform through a filterbank of ten order 1000 FIR filters with passbands at mid-values between f_i neighbors (see *Stimuli*, above) starting at 143 Hz. Filter delays were compensated and the envelope in each band was extracted as above. Sampling rates were reduced to 1 KHz and signals smoothed by a delay-corrected 4th order binomial FIR filter. Half-wave rectification (i.e. setting negative values to zero) of the derivative of the stimulus envelope filterbank output gave *envelope onset* representations of the stimulus signal based on the filterbank. Prior to reverse correlation, both envelope and envelope onset representations were transformed to dB-scale.

Linear STRF model estimation. STRF estimation was performed via *boosting*, a technique where the error estimate $\varepsilon(t)$ (in Eq. 2) is minimized iteratively via sequential modifications to the STRF[19]. The name originates from the ability to improve ('boost') an estimate learning algorithm by establishing aggregate decision rules from across a sequence of many estimation steps, each needing only slightly-better-than-chance accuracy[20]–[22]. This technique can then be implemented as a forward stage-wise fitting that follows a greedy heuristic, by adding the contribution with the largest available mean-squared-error reduction at each given step[19], [23] and in turn maximizing the predictive power of the model[24]. Operationally, STRF estimates by boosting were initialized as a null matrix of dimensions TxF, where T equals the number of experimental time bins and F is the total of frequency bins (=10; for TRF estimates,

F=1); optimization followed through exploring fixed increments and decrements per spectrotemporal bin individually. Among the resulting 2x*F*x*T* possible choices, the outcome with minimum mean-squared-error was selected as the next step in the running STRF estimate. The procedure was iterated, accumulating optimizations, until modifications instead produced a sustained increase in mean-squared error[23], since the method is not guaranteed to find a global optimum. This termination method effectively imposes a sparse structure on the STRF, which allows for extraction of high-temporal resolution features in the STRF even if only low-frequency content was present in the input waveforms (other STRF estimation methods such as normalized reverse correlation[18] and generalized linear models could also used [19], [109], [110]). Other detailed descriptions of the boosting algorithm implementation for timeseries data, including MEG/EEG are available[19], [25].

STRF predictive power bounds. The measured evoked cortical response $r^*(t)$ may include stimulus-independent noise, the presence of which is a consequence of the finite dataset size and leads to STRF model parameters that overfit to the training data. Performance measures that account for stimulus-independent noise are necessarily overestimates and therefore can be considered to act as empirical upper bounds of model performance[24]. In contrast, the risk of overfitting can be minimized using *cross-validation*, where a fraction of the $r^*(t)$ timeseries $\bar{r}(t)$ (e.g. 90%) is reserved for model training, and testing is done on the remaining fraction incorporating only the model's ability to generalize over novel stimulus instances. This would be expected to underperform with respect to an optimal model for the dataset in question and so indicates a lower bound for its

performance [24]. In practice, it is this conservative, cross-validated lower-bound that is used for STRF estimates.

Nonlinear extension. Linear encoding models may fail to characterize firing rate predictions based on effects such as threshold activity, past-history dependencies, dynamic range compression, synaptic transfer, and the non-negative distribution of the neuron response for example. At the single neuron level, predictions can be improved via introduction of static nonlinearities derived by empirical fit[9], or via intermediate nonlinearities in more complex model hierarchies[110]. For coarse-grained continuous neural responses such as local field potentials and the MEG signal here, it appears that such model hierarchies may no longer apply well. When a static nonlinearity was incorporated using a linear-nonlinear (LN) model[111], [112], only a 2% improvement to predictive power resulted (quadratic fit, R^2=0.972, S5 Fig) and so was not pursued.

Estimation of STRF predictive power and noise limit extrapolation. To assess STRF model validity, predictive power was estimated as the fraction of a response signal variance that is stimulus-explained, and corrected for the reduction of noise-related variance achieved by averaging[24]. Namely, for MEG response timeseries $r_1(t)$, … ,$r_N(t)\mathrm{r}_1(\mathrm{t}), \ldots, \mathrm{r}_\mathrm{N}(\mathrm{t})$ where N is the number of repetition trials, total variance is expressed as the average of each trial's individual variance

$$\mathrm{Var}(r) = \frac{1}{N}(\mathrm{Var}(r_1(t)) + \cdots + \mathrm{Var}(r_N(\mathrm{t}))) \quad (3)$$

while evoked variance can be expressed as that of the average response $\mathrm{Var}(\bar{r})$. When N is large, the extent to which total variance is larger than evoked variance indexes

reliability for the response source. Contributions to total variance $\mathrm{Var}(r)$ are then partitioned into those stemming from the evoked *signal*, and the remainder is treated as *noise*:

$$\mathrm{Var}(\mathrm{signal}) = \frac{1}{N-1}\left(N \cdot \mathrm{Var}(\bar{r}) - \mathrm{Var}(r)\right) \quad (4)$$

$$\mathrm{Var}(\mathrm{noise}) = \frac{N}{N-1}\left(\mathrm{Var}(r) - \mathrm{Var}(\bar{r})\right) \quad (5)$$

such that estimates are corrected for cases where *N* is small. Often, STRF model estimates are optimized to produce accurate predictions of the evoked response only; in such cases, use of single-trial variance provides an additional statistic regarding the event-related contribution to available recordings. Once a STRF model has been obtained for a particular condition and subject, its ability to predict the evoked response is assessed as the extent of evoked response variance that is not residual error, that is $\mathrm{Var}(\bar{r}) - \mathrm{Var}\left(\bar{r} - \bar{r}_{pred}\right)$. This expression is the model's predictive power, which after division by the estimated signal power (eq. 4)$\mathrm{Var}(\mathrm{signal})$, represents the fraction of stimulus-evoked variance described by the linear STRF model contingent on a given experimental condition and subject. Analogously, noise power in the same response may be normalized by the estimated signal power, providing the inverse proportion to which the procedure of averaging reduces response variability. When *N* is very large, a normalized noise power of e.g. 10 indicates that averaging reduces variance in the evoked signal to almost a tenth of the original total variance. In the hypothetical case where the procedure of averaging yields no reduction in variability (such as with identical trial response instances), the absence of variability reduction implies an absolute zero noise level. Empirically, each dataset's (condition and subject) predictive power can be indexed by the intrinsic noise power (e.g. Fig 1C). Assuming the responses have been measured from a similar

population, regression analysis may produce an estimate of the STRF model class predictive power, via its extrapolation to the theoretical noise-free limit[24].

Chapter III

Dynamic cortical representation of perceptual filling-in for missing acoustic rhythm

Summary

In the phenomenon of perceptual filling-in, missing sensory information can be reconstructed via interpolation from adjacent contextual cues by what is necessarily an endogenous, not yet well understood, neural process. In this investigation, sound stimuli were chosen to allow observation of fixed cortical oscillations driven by contextual (but missing) sensory input, thus entirely reflecting endogenous neural activity. The stimulus employed was a 5 Hz frequency-modulated tone, with brief masker probes (noise bursts) occasionally added. For half the probes, the rhythmic frequency modulation was moreover removed. Listeners reported whether the tone masked by each probe was perceived as being rhythmic or not. Time-frequency analysis of neural responses obtained by magnetoencephalography (MEG) shows that for maskers without the underlying acoustic rhythm, trials where rhythm was nonetheless perceived show higher evoked sustained rhythmic power than trials for which no rhythm was reported. The results support a model in which perceptual filling-in is aided by differential co-modulations of cortical activity at rates directly relevant to human speech communication. We propose that the presence of rhythmically-modulated neural dynamics predicts the subjective experience of a rhythmically modulated sound in real time, even when the perceptual experience is not supported by corresponding sensory data.

Introduction

The ability to overcome the problem of missing but important sensory information, such as a conversation obscured by heavy background noise, is ethologically valuable. Even when physical information may be lost entirely, restorative phenomena such as the auditory continuity illusion, phonemic restoration, and other forms of perceptual filling-in[113]–[115], allow for the percept of stable hearing in natural environments. These effects have long been hypothesized to rely on the brain's ability to conjecture a reasonable guess as to the nature of the missing fragments[113], [116]. Furthermore, as has been extensively argued, predictive coding is a task well suited for cerebral cortex[117]–[119] but systematic accounts of endogenous cortical mechanisms responsible for these percepts remain unspecified.

Rhythmically-modulated sounds generate steady predictable events for which disruptions and resumptions may indicate the grouping strength of dynamic perceptual streams[120], [121]. If replacement of these sounds by noise may, under some circumstances, preserve the perceived rhythm in apparent continuity, how are such streams instantiated at the neural level? Rhythmic sounds drive auditory steady-state responses (aSSR) in auditory cortex and can be recorded non-invasively via magnetoencephalography (MEG)[122]–[124], with responses to rhythmic rates <10 Hz being especially prominent[125]–[128]. To the extent to which the neural responses track the stimulus rhythm, they can be considered sparse neural representations of the modulation rate. This experimental framework was employed to investigate the cortical effects of briefly masking and removing an ongoing low-frequency rhythmic pattern. We hypothesize that for cases where perceptual restoration of the removed rhythm occurs, the neural signature of the

removal is attenuated—akin to stabilization of a cortical representation, in line with perceptual grouping under dynamic continuity. This predicts that during perceptual filling-in, the dynamical evolution of a listener's cortical response retains oscillation in synchrony with the expected but acoustically missing rhythm.

Listeners' perception of a continuous 5 Hz rhythmic pattern during masking was probed in a two-alternative forced choice task, where the acoustic pattern may or may not have been removed with equal probability. Simultaneously obtained MEG responses were then partitioned according to both physical and perceptual conditions, using wavelet analysis to localize oscillatory responses in time and frequency. The finding of rhythmic aSSR-like responses in cases where perceptual filling-in occurs is consistent with underlying mechanisms requiring a sustained neural representation of the restored feature[114]. Importantly, it demonstrates dynamical restoration processes occurring at scales commensurate with informal speech articulation rates[129], as well as within MEG frequency bands that reflect cortical phase-locking to the slow temporal envelope of natural stimuli[25], [127].

Results

Sustained neural rhythm follows acoustic rhythm in noise. Subjects listened to four blocks (~14 min each) of a 5 Hz frequency modulated (FM) rhythmic stimulus, repeatedly masked by noise probes at pseudo-random times (see *Methods*). Half of the probes replaced the underlying rhythmic FM tone with a constant frequency tone, and half instead simply masked the underlying rhythmic stimulus, here called non-rhythmic

and rhythmic probes, respectively (Fig. 1A insets). Between noise masker segments, MEG responses to steady rhythmic intervals show strong aSSR, even on a per-trial basis. Noise masker segments generate strong transient onset-like responses, after which any residual phase-locked response may disappear, on average, for rhythm-absent probes but not rhythmically-driven probes (Fig. 1A). To determine whether across subjects this change results from a decrease in aSSR power, or increased temporal jitter that would reduce averaged aSSR, inter-trial phase coherence (ITPC) and power analyses were performed on single-trial and evoked data respectively (e.g. Fig. 1B). Results of inter-trial phase coherence (ITPC) analysis reveal that, within the 0.55 – 1.22 s post probe onset interval, the ITPC difference is significant across ($N = 35$) listeners ($p < 0.001$; non-parametric permutation test). Testing for evoked rhythmic power for across listeners similarly reveals a significant difference ($p< 0.001$) within the 0.56 – 1.23 s post probe onset interval. Thus the dual phase and power analyses show that both decreased aSSR power and increased intertrial jitter contribute to the decrease of the neural 5 Hz component in rhythmically absent versus driven probes.

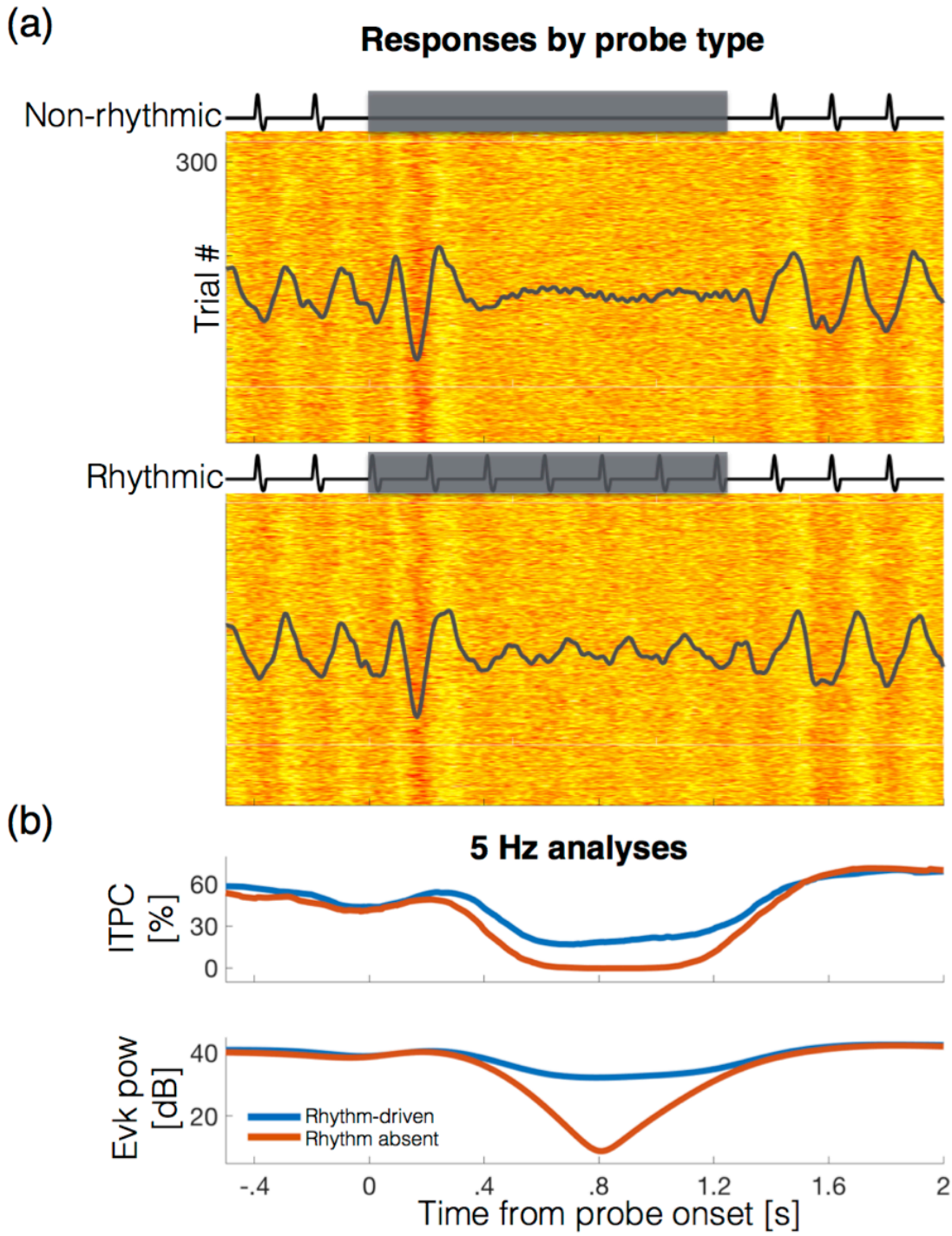


Figure 1. Neural representations of (un)modulated masked sound from a representative subject. (**a**) MEG responses before, during, and after a noise probe are shown (single MEG component obtained via spatial filtering; see *Methods* and Appendix B Supplementary Fig 1). The basic stimulus consists of a 5 Hz pulsatile (short duty-

cycle) FM tone, centered at f_0 = 1024 Hz, to which 1.24 s noise probes were applied. Insets: illustration of a non-rhythmic probe where pulses are replaced by the constant tone (top); and a rhythmic probe, where the FM continues under the noise (bottom). Before and after the probes, phase locking to the main rhythmic stimulus is apparent even on a per-trial basis. Overlaid on each response raster, evoked activity (averaged separately for each probe type) reveals a measurable aSSR during rhythmically-driven probes (top) but not during rhythm-absent probes (bottom). (**b**) Top: Phase analysis at 5 Hz shows estimated phase-locking over time as measured by ITPC. During masking ITPC values drop to near floor in rhythm absent probes (orange) but only to half of baseline levels in rhythm-driven probes (blue). Bottom: Analysis of spectral power (also at the 5 Hz rhythm rate) also shows considerable difference between probe types for this subject.

Sustained neural rhythm follows listeners' perceived rhythm in noise. In order to determine how neural representations of rhythm co-varied with perception, after each trial the probe was classified by the subject as perceived as rhythmic or as non-rhythmic. This resulted in a 2-by-2 partition of analyzed trials: (1) non-rhythmic probes perceived rhythmic ('filling-in'); (2) non-rhythmic probes perceived non-rhythmic (rhythm 'absent'); (3) rhythmic probes perceived rhythmic (rhythm 'present'); and (4) rhythmic probes perceived non-rhythmic (rhythm 'missed'). Fig. 2 shows the grand average evoked 5 Hz response power before, during, and after noise probes, for each combined condition of stimulus and percept. Transient (and broadband) masker-onset responses were evident during the initial 0.3 s post masker onset (cf. Appendix B Supplementary Fig 2) (brief pre-causal dips accompanying these transients are due to convolution residuals from the continuous wavelet transform).

For non-rhythmic probes (Fig. 2A), phase coherence dropped to almost 0% for

both perceptual conditions (filling-in and absent, right panel). Rhythmic spectral power also dropped from the initial baseline for both perceptual states, but the decrease was on average 7.9 dB worse when subjects reported the rhythm absent than present (filling-in). Decreases were restored to baseline values by 0.8 to 1.2 s post probe offset (equivalent to between 4 and 6 rhythmic pulse cycles. Thus, within non-rhythmic probes, a sustained and significant percept-specific difference was observed in rhythmic evoked power (0.56 to 1.19 s, $p < 0.001$), but this was not the case for phase locking ($p > 0.18$).

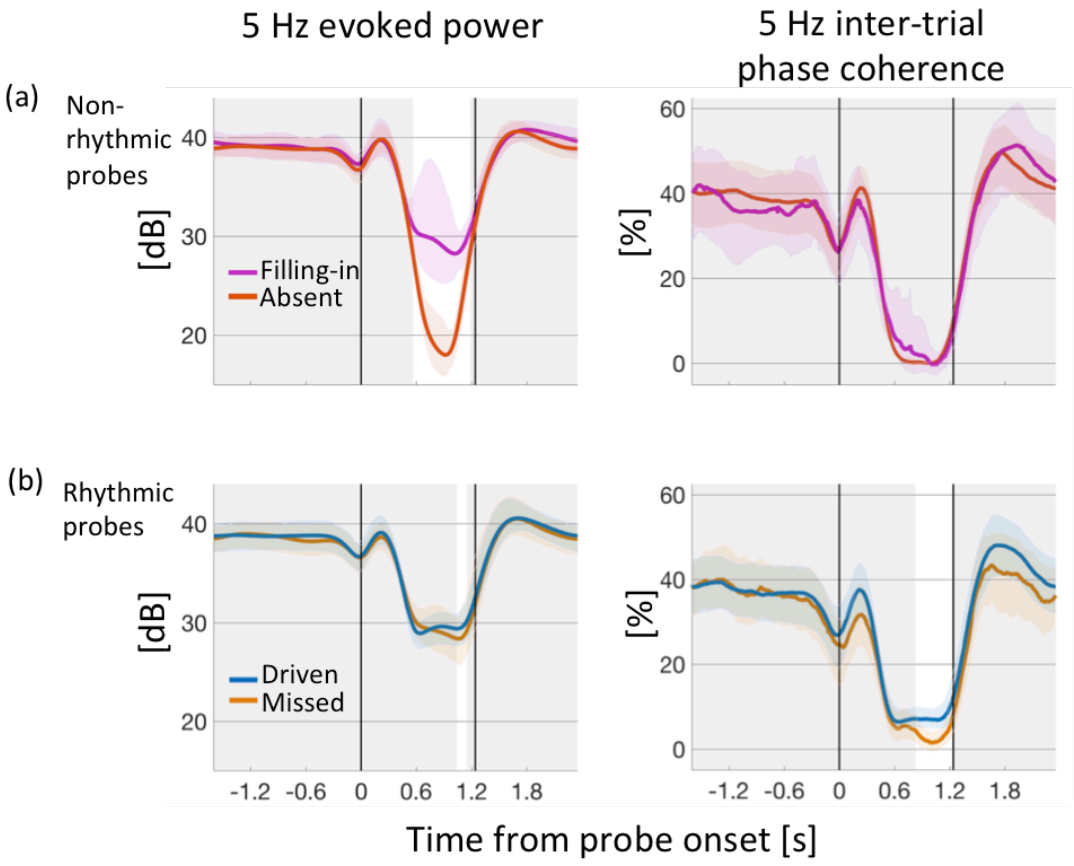


Figure 2. Percept-specific endogenous representations of patterned sound. Grand averages ($N = 35$) of rhythmic evoked power and intertrial phase coherence partitioned

by probe type and reported percept. Noise probe starts at the first vertical line at $t = 0$ s and continues until the next vertical line at $t = 1.24$ s. (**a**) Non-rhythmic probes: (Left) After an initial transient, rhythmic evoked power was reduced regardless of percept, but differentially by 7.9 dB depending on percept as present (magenta), or absent (orange). (Right) No significant difference was observed for ITPC, where there was a reduction to near floor during the probe. (**b**) Rhythmic probes: (Left) During masking, rhythmic evoked power drops by 9.5 dB in average, holding relatively steady for the duration of the probe. (Right) Similarly, inter-trial phase coherence drops by about 81% for the duration of the probe. For probes in which the rhythm was missed (brown), however, both evoked power and ITPC showed an additional reduction (only near the end of the probe) compared to rhythmically-driven probes (blue). Solid lines: mean across subjects and trials; Color bands: bootstrap 95% confidence of the mean over subjects; Grey bands: time intervals with no significant difference by percept.

For rhythmic probes (Fig. 2B), the masker was associated with an average relative decrease of 9.5 dB evoked power regardless of perceptual condition (driven and missed), and with a relative decrease of ~75% in trial-to-trial phase locking. When subjects missed the rhythm, evoked power and inter-trial phase coherence both further decreased, with percept-specific decreases sustained over a longer period for ITPC (0.84 – 1.25 s, $p < 0.001$; right panel) than evoked power (1.04 – 1.15 s, $p = 0.008$; left panel).

Rhythmic neural power as discrimination statistic in a rhythm detection task. With the observation that differential neural processing of masked rhythm depends on listeners' percept, it was next investigated whether the observed divergence might have properties of an internal variable underlying discrimination. Based on the previous result, we hypothesized that the 5 Hz target neural processing power in the final ~600 ms of the

probe interval might act as such variable. For each subject, a metric was created from the rhythmic evoked power differences contrast, integrated over the 0.56 – 1.24 s interval of interest post probe onset. To illustrate the use of this latent variable as a discrimination statistic, a bootstrap resampling of trials (with replacement) was used to produce distributions of evoked power sustained over the critical window (two representative subjects shown in Fig. 3A). A neural discriminability metric was then computed from their relative separation (see *Methods*). To assess the potential of this sustained evoked power to operate as a variable relevant to perceptual discrimination, the neural metric was compared with psychometric *d'* scores that index behavioral sensitivity of listeners to the detection task[130] (Fig. 3B, blue), with the result that the two are significantly correlated ($\rho = 0.728, p = 1.04\text{x}10^{-6}$).

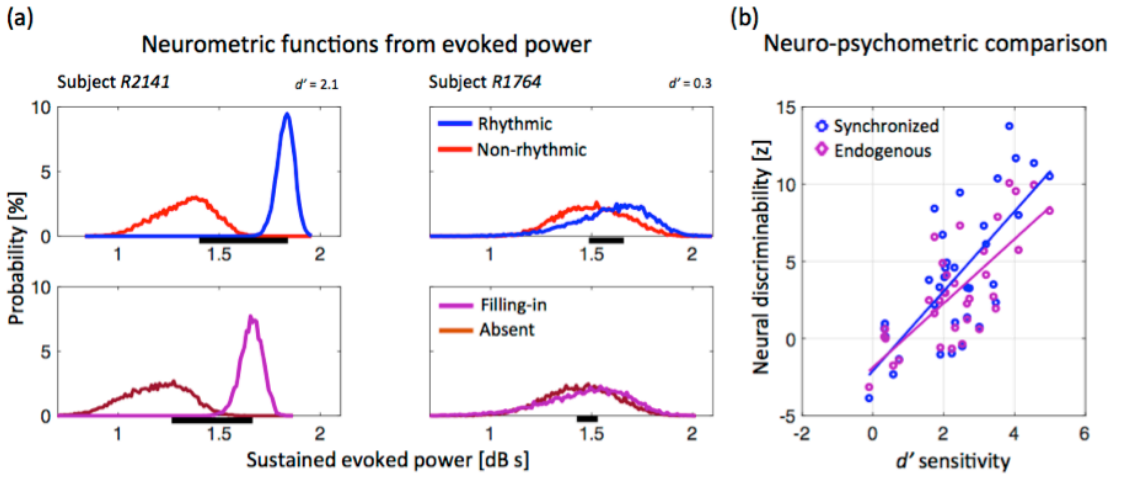


Figure 3. Rhythmic target power acts as a discriminant neural statistic for perceived rhythm. (**a**) Top: In two representative subjects, behavior covaries with empirically-derived neural discriminability distributions. Probability distributions of a given level of sustained (time-integrated) evoked power depend on the acoustic presence (blue) or absence (red) of stimulus rhythmic FM; a neural discriminability score (proportional to

horizontal black bar length) can be obtained from them. In the first subject (left panel), the small overlap between the distributions gives high neural discriminability; for the second subject (right panel), both distributions overlap substantially, giving poor discriminability. Bottom: Next, empirically-derived neural distributions were obtained *only from non-rhythmic probes* (i.e., the red curves in the top panels), now conditioned instead by percept. A similar pattern in the distributions is observed. Distributions obtained via bootstrap. (**b**) Over subjects, the psychometric *d'* sensitivity index (abscissa) correlates with the neurometric discriminability index based on acoustic contrast (rhythmic versus non-rhythmic probe, blue; $\rho = 0.73$, $p = 1.0\text{x}10^{-6}$). Critically, behavioral sensitivity to 'filling-in' also correlates with rhythmic evoked power differences despite the absence of stimulus rhythm via the related neurometric discriminability index based on perceptual contrast (filling-in versus reported absent, magenta; $\rho = 0.69$, $p = 6.1\text{x}10^{-6}$).

A related latent discrimination statistic, directly relevant to the phenomenon of filling-in, is computed with contributions only from endogenous (non-sensory) factors, by analyzing the responses to non-rhythmic probes exclusively (Fig. 3A, bottom). In these purely percept-specific (constant acoustics) distributions, neural power discriminability was defined analogously as the difference in rhythmic evoked power between filling-in and rhythm-absent trials, integrated over the time at which significant differences were observed at the group level in the previous section (0.56 to 1.19 s post probe onset, as in Fig. 2B). Just as for the acoustic contrasts, this discriminability index also correlates strongly with the psychometric sensitivity indices across listeners (Fig. 3B, magenta) (ρ=0.745, p=4.23x10^{-7}). Thus, consistent with the properties of a latent discrimination statistic, sustained evoked power may account for both stimulus- and percept-specific differential processing, where the latter reflects only endogenous neural processes.

Spectrum of power increase in target-related neural rhythm dynamics with filling-in.

Given the possibility that increased power at the 5 Hz rhythmic frequency would be accompanied by increased spectral power at other frequencies, it is important to consider whether change arises as a power gain specific to the target frequency or as a modulatory effect over a larger spectral region that includes the target frequency band. By extending the wavelet analysis over a broader frequency range (1-25 Hz), the spectral extent of restoration was probed to address whether changes are target-specific, or instead accompanied by other activity that may be behaviorally relevant.

Evoked power analyses across probe conditions and subjects reveal that the evoked response contains two frequency ranges, one centered on the target 5 Hz, and the other centered on the 10 Hz first harmonic (Fig. 4A). To analyze time-frequency power contrast between conditions, corresponding spectrograms (baseline corrected per frequency band) were subtracted. In particular, the 'driven' minus 'absent' map results in a contrast whose differences arise from synchronization to physical differences in the sound, while 'filling-in' minus 'absent' maps differences due entirely to endogenous activity (Fig. 4B, left panels). For the first case, the defined 'synchronized' contrast (Fig. 4B, top left) group average data shows a spectrotemporal region, ~600 ms post probe offset until the end of the probe, of significant differential neural processing ($p = 3.3\text{x}10^{-4}$), rooted in physical stimuli differences. The region is limited to the spectral neighborhood of the target (half maximum 4.1-6.7 Hz; maximum 3.8-7.5 Hz), which may be expected as smearing from Fourier/Heisenberg uncertainty. For the 'endogenous' contrast (Fig. 4B, bottom left), a similar profile was found (half maximum 4.1-6.6 Hz; maximum 3.8-6.8 Hz; $p = 6.7\text{x}10^{-4}$), with additional enhancement around the target first

harmonic (0.4 to 1.1 s post probe onset; half maximum 9.7-11 Hz; maximum 8.9 to 11.9 Hz; $p = 0.01$). In a related analysis of a third partition contrast, 'rhythm-driven' minus 'missed', no spectrotemporal cluster of significance was found (p=0.29).

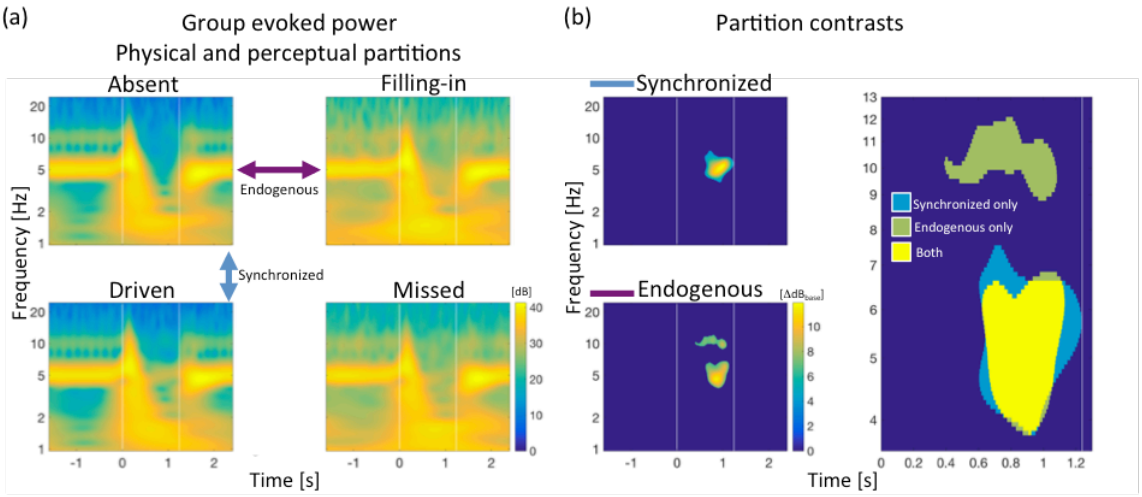


Figure 4. Stimulus- and percept-specific spectrotemporal modulations of cortical activity during restored rhythm. (**a**) Wavelet power correlograms, in a 1-25 Hz frequency range, reveal qualitative differences in steady neural responses post probe onset, across participants (N = 32). Color arrows indicate spectrogram pairs submitted to difference contrasts as follows. (**b**) Differences between spectrograms reveal differential processing under alternative percepts, whether based on different physical sounds (top left), or on endogenous restorative processes (bottom left), in both cases specific to the target 5 Hz frequency band. The latter case of filling-in generates enhanced sustained power in the first harmonic band (~10 Hz) as well. Synchronization maps are shown masked by regions of group-level significance, as determined by permutations within contrast pairs, performed independently across subjects ('driven', $p = 3.3\text{x}10^{-4}$; 'filling-in'-near 5 Hz $p = 6.7\text{x}10^{-4}$, filling-in'near10 Hz $p = 0.01$). The lower-rate rhythmic enhancements (~5 Hz) coincide spectrotemporally even though the sensory bases for each are different (right). White vertical lines indicate noise probe temporal edges.

Upon examination of whether the additional spectral information conveyed by these

maps improved neural predictions regarding listeners' behavior, we found that neural discriminability indices based upon the 'synchronized' region in this section showed no improvement over the target frequency specific index obtained previously for 5 Hz only measures ($\rho = 0.53$; $p = 0.001$). The 'endogenous' regions, jointly, showed no improvement in predictive power of listener's performance ($\rho = 0.72$; $p = 1.4\text{x}10^{-6}$) over that of the target-based index alone. Separating these regions into 5 Hz and 10 Hz domains revealed that the lower (target rhythm) region was more predictive (5 Hz only: $\rho = 0.73, p = 8.2\text{x}10^{-7}$; 10 Hz only: $\rho = 0.44, p = 0.01$). These results suggest that differential narrowband 5 Hz power is most critical to explain listeners' detection performance shown previously, and that for filling-in trials, some improvement also arises from integrating over the broadened filter to include neighbor target frequencies present in the average timeseries of endogenous neural activity.

Discussion

The subjective experience of attending effectively to complex sound scenes in noisy environments can be substantially assisted by perceptual restoration. This effect is investigated using MEG to record the neural dynamics of a steady temporal pattern while repaired perceptually. Measures of differential cortical processing contributed to the identification of a discrimination statistic predicting a subject's behavioral performance sensitivity. The data are consistent with the view that perceptual restoration is attributable to endogenous neural processes, emerging from learnable temporal patterns present in the tracked auditory object, at modulation rates that dominate natural communication speech

sounds.

Perceptual restoration, the effect of hearing the continuation of a sound regardless of an interrupting masker, includes descriptions of “auditory induction”, “temporal induction”, “perceptual synthesis”, or “contextual catenation” of dynamic sounds in classic studies[131], [132]. It implies an ability to discount disruptive but extraneous interruptions to relevant acoustic signals, so much so that even noise-filled *gaps* are more likely to be discounted as such[132]. Where multiple interpretations of a relevant acoustic signal are possible (e.g. phonemes), perceptual restoration has been probed in identification tasks; for more constrained decision spaces, it may be probed based on sound delivery quality assessments, such as gap localization of the excised token signal (e.g. Warren’s paradigm[132]), and by discrimination of noise-added vs. noise-replaced token gap alternatives (e.g. Samuel’s paradigm[133]). Our method subscribes to the latter approach, also referred to as ‘filling-in’, which emphasizes the signal detection strategy followed in cases where a listener classification is inconsistent with the token absence in a gap[115], [134]–[140]. As has been noted[141], from the listener’s utilitarian perspective, this effect of induction in a challenging environment is not aimed at the production of decision errors (or illusions) but to assist against masking. Restoration refers to the perception of a token projected by a context (such as a speaker’s intention), with apparent intactness[141]. Critical to this is a strong masker, along with contextual evidence favoring a specific acoustic token with high probability. This combination allows inference that the lack of auditory evidence of the token could be ascribed to energetic masking[113], [116], [142].

A simple and compelling example of perceptual restoration is that of a pure tone followed by a brief noise-filled gap where the tone has been excised: this leads to a strong illusory percept of continuity of the tone[143]. The percept appears to rely on two related effects, the more obvious being conveying the original signal as uninterrupted, but also, critically, accompanied by an attenuation of discontinuity boundaries[144]. Neural correlates of both effects have been observed in single units in macaque primary auditory cortex (A1), where up to 35% of sampled single units respond to a gap with noise as though the tone were continuously present[145], [146]. In some cases there is also failure of a transient response at the end of the gap[145]. For human listeners, there is evidence that such compensatory principles may extend to disruptions to dynamically modulated sound, including amplitude-modulated (AM) sound, single vowels, and consonants within words[120], [121], [135], [139], [147], [148], the latter of which fall under the concept of phonemic restoration[113], [115], [133]. Depending on stimulus, neural correlates have been localized to different areas, including Heschl's gyrus for missing AM noise[147], the posterior aspect of superior temporal gyrus for disrupted vowels[139], and wider brain networks including the superior temporal lobe in the case of missed phonemes[135], [148]. In addition, mixed evidence points to a basis for restoration in terms of endogenous modulations to boundary encoding: on the one hand, the search for differential onset responses to noise when under restoration, indexing alternative encoding, has yielded negative results so far[138], [139]; on the other, induced narrow-band (3-4 Hz) desynchronizations that are restoration-specific, and occur after gap onset, have been suggested by results from EEG[139], [149].

In this study the differential temporal boundary encoding under restoration was not specifically addressed[149], but instead the emphasis was on the neural representation of the missing rhythm itself, via measures of evoked rhythmic MEG responses. While restoration of continuous tones has been observed for segments as long as 1.4 s[150] behaviorally, to our knowledge this is the first investigation where cortical aSSRs are directly implicated in perceptual restoration, sustained in real time representing a temporal code. That neural phase information was not reliable, despite an apparent continuity of the rhythm, is consistent with behavioral analyses suggesting that listeners may not track phase information under illusory FM continuity[120], [121]. An cortical EEG study by Vinnik and colleagues[138] showed no change to neural spectral power sustained along noise gaps embedded in a 40 Hz AM context stimulus during restoration; on the other hand, it has been shown that changes to neural spectral power in brainstem responses may occur during restored pitch of a missing 800 Hz carrier tone[140]. It is possible that while gamma-rate acoustic modulations can be represented cortically with a temporal code[125], [151], they are also at rates that involve pitch quality – a representation of which implies substantially distinct cortical coding modes[152] assisting restoration.

In other sensory modalities, some restorative phenomena may fall in the category of perceptual experience that does not represent the absence of a physical stimulus, but rather, an alternative interpretation based on additional contextual information, e.g., the case of illusory induction of perceived kinesthetic trajectories[153], [154], and of spatial contours in certain visual displays[155]–[157]. Context-sensitivity in general is considered a requisite for cortical predictive coding[158], which in the case of hearing

may depend on known priors regarding the sound temporal dynamics. A compelling example arises from missing, but highly expected, click-like sounds that generate auditory onset-like responses locked to the nominal time of delivery of the missed sound[159]. Additionally, long duration, rhythmic metric structures may produce endogenous neural locking to a subharmonic frequency of the actual acoustic beat when it has the potential to be perceived as the underlying rhythm, whether listeners are instructed to do so[160], or passively listen in the absence of instruction[161]. Correspondingly, the data here show that with perceptual restoration of masked rhythm, endogenous representational differences may emerge as early as 0.6 s post masking, at the target rhythm. There is also activity at the first harmonic, 10 Hz, but there one cannot entirely rule out yet alternative explanations involving enhanced alpha activity[138], since with increased alertness at some trials over others, a systematic differential in spontaneous alpha activity might be responsible[162]. For filling-in and rhythm-missed trials related to inattention, reduced vigilance might be expected to effectively increase alpha activity. We did not, however find this; instead, filling-in trials displayed a narrow-band 10 Hz power increase strongly concurrent with the target duration, therefore consistent with being a harmonic of the endogenous 5 Hz rhythm. Alpha-band related effects due to non-uniform attentional states should be investigated in future studies using rhythms whose first harmonics are not in the alpha band. Our data does not reject the possibility of spontaneous and temporally patterned cortical activity profiles influencing sensory processing, as in ongoing slow-wave activity that may interact with evoked signals as a temporally coordinated modulation of excitability across distributed cortical fields[163], [164] .

Focus on analysis of endogenous activity may address circumstances under which the brain repairs certain temporal features of highly stereotyped sound. This is part of the general problem of determining what relationship does a neurally-instantiated representation of a missed pattern has with a template representation mapping to actual acoustic experience. Solutions may offer key insight into biologically-inspired applications dealing with incomplete information. In particular, the modulation studied here corresponds to the temporal scale of syllabic production in human speech[165] and the slow temporal envelope of natural stimuli[166], thus raising the question of whether similar restorative phenomena exist during sequences of inner or imagined speech, as well as during auditory hallucinations.

General methods

Participants. 35 subjects (12 women, 25.7 ± 4.4 years of age) with no history of neurological disorder or metal implants participated in the study, and received monetary compensation proportional to the study duration (~ 2 hours). The experimental protocol was approved by the UMCP Institutional Review Board, and all experiments were performed in accordance with its relevant guidelines and regulations. Informed written consent was obtained from all participants before study sessions.

Stimuli. Four template sound stimuli were constructed with MATLAB® (MathWorks, Natick, United States), each consisting of ~15 minutes of a 1024 Hz tone frequency-modulated (FM) at 5 Hz with modulation range (log-sinusoidal) 512–2048 Hz and a 20% duty cycle[128]. 420 rhythmic probes were created by adding 1.24 s of noise to the basic

stimulus, at pseudo-random times. Noise was generated de novo per probe, and spectrally matched to the FM but with random phase. A fixed signal-to-noise ratio value was chosen from the -4 to 4 dB range, per participant. 420 non-rhythmic type trials were additionally created in the same manner, except that the underlying FM was replaced with constant carrier frequency. Inter-probe time intervals were 1.6 s plus a discrete Poisson-distributed random delay ($\lambda = 1.2$ s); the exact onset time was rounded to a multiple of the stimulus period (0.2 s), so that all probe onset times kept constant phase with the main rhythm. Sound stimuli were delivered through Presentation® (NeuroBehavioral Systems, Berkeley, United States), equalized to be approximately flat from 40–3000 Hz, at a sound pressure level ~ 70 dB. Sounds were transmitted via E-A-RTONE® 3A tubes (impedance 50 Ω) and E-A-RLINK® disposable foam intra-auricular ends (Etymotic Research, Elk Grove Village, United States) inserted in the ear canals.

Experimental design. After a brief practice session, subjects were instructed to push one of a pair of buttons based on whether they detected a 5 Hz rhythm. In order of importance, participants were instructed to: (i) wait until probe ended before pressing the button, weighting accuracy over reaction time; (ii) respond only to the probe immediately presented; (iii) modify their choice by pressing the other button only if certain and still before the next trial. Trials that did not meet the requirements, and corrected trials, were excluded (median 6.8% and 1.3% of trials respectively). To avoid transient cortical dynamics associated with motor response execution[167], trials beginning less than 250 ms from the previous response were also excluded (median 6.3% of trials). To more

evenly distribute the proportion of correct answers across participants, the masker signal-to-noise ratio (SNR) was fixed in advance, from on of 0, ±1, ±2 or ±4 dB. Silent films were presented concurrently, which subjects were instructed to watch.

Data recording. MEG data were collected with a 160-channel system (Kanazawa Technology Institute, Kanazawa, Japan) inside a magnetically-shielded room (Yokogawa Electric Corporation, Musashino, Japan). 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. Every session had four experimental blocks. In the case of seven participants, the experiment had to be suspended early due to time constraints (mean 89% completion in these participants, minimum 75%); for one participant only 2 blocks out of 4 were recorded due to transfer failure. Two participants requested pauses during a block, which was terminated and later repeated in whole.

Data processing. A 1-30 Hz band-pass third order elliptic filter with at most 1 dB ripple and 20 dB stopband attenuation was applied and noise sources were removed as follows. *Environment noise*. Time-shifted principal component analysis[105] (TS-PCA) was applied to remove environmental noise, using the three reference magnetometers (N_{lags} =

43). *Sensor-specific noise*. Sensor-generated sources unrelated to brain activity were subtracted using sensor noise suppression (SNS)[107]. *Spatial filtering*. A per-participant data-driven model was used to synthesize spatial filters from the responses to the unmasked rhythmic sound stimulus via denoising spatial separation (DSS)[108]. The responses were structured as a matrix of dimensions T x N x K; where T is the number of samples (=1400), N is the number of usable recording segments (average=514.3), and K the number of active sensors (average=156.8). This spatial filter selects for the most reproducible aSSR component over trials, generating a single virtual sensor used in the remaining analysis.

Data analysis. Trials were classified a posteriori, according to subjects' reports, into one of four groups: rhythmic-trial perceived such ('driven') or as not as non-rhythmic ('missed'); non-rhythmic trial perceived as such ('absent'), or as rhythmic ('filling-in'). Time-frequency analysis used a Morlet wavelet transform with 0.2 s scale, permitting estimation of spectral evoked power at the bandwidth of experimental interest (5 Hz). For evoked power and ITPC contrasts, statistical clusters were found during which there were significant differences across experiment conditions according to non-parametric permutation tests[168]. A measure of 'neural discriminability'

$$\Delta P_i^{A,B} \equiv \int_{T_0}^{T_1} \left(P_i^A - P_i^B\right) dt \qquad (1)$$

is defined as the area between two evoked power curves P obtained each at conditions A and B for the i-th subject, and computed over a fixed time interval (T_0 = 0.58 s and T_1 =1.2 s post noise onset on average), as defined by statistical clusters of significance found at the group level for the given contrast AB. Measures for shifts in ITPC were computed in similar way. Perceptual sensitivity of a subject in detection is given by d-prime

analysis[130], $d'_i = z(H_i) - z(F_i)$ where for each subject i, H_i the fraction of rhythmic probes labeled rhythmic, and F_i the fraction of non-rhythmic probes labeled rhythmic, undergo a z-transformation.[169]

To investigate whether the observed pattern of percept-specific differences was due to unintended acoustical or statistical properties in the stimulus constructs, stimulus probes were analyzed a posteriori. No significant differences were found in stimulus temporal modulations when partitioned by percept, within rhythmic (p=0.85) nor non-rhythmic (p=0.84) probes (paired-sample t-tests, Appendix B Supplementary Fig 3).

Subjects' reported percepts corresponded to the physical acoustics (presence or absence of rhythm) approximately 5 times as often as not, resulting in data pools with differing signal-to-noise ratio improvement from averaging. Therefore inter-trial phase coherence measures included bias correction[170] as small sample sizes are especially prone to bias. The unbiased estimator is based on the squared ITPC (also defined as squared 'modified resultant length'[170]), which may be negative after estimated bias subtraction. To investigate the possibility of related biases in the rhythmic evoked power measures, post hoc two-sided non-parametric permutation tests were performed by collecting, for each subject, all trials from the two conditions to be compared, and instantiating resampled of partitions of fixed size (original sample sizes per subject); the group-level test statistic obtained in the actual partition was then contrasted against those obtained at group level across the distribution of resampled instances. Using the 5 Hz evoked power difference between conditions in the same intervals of significance, it was found that responses to non-rhythmic probes show significantly greater power when reported perceived as rhythmic versus non-rhythmic (0.56 to 1.19 s ; p=0.007); a similar

result held for responses to rhythmic probes, which also show significantly greater power when reported perceived as rhythmic versus non-rhythmic (1.04 to 1.15 s ; p=0.034). Potential systematic differences resulting from the per-subject signal-to-noise (SNR) ratio were also investigated, but no evidence was found of differences, neurally (ρ=0.10, p=0.57) or behaviorally (ρ=0.33, p=0.054). One participant was excluded from the analysis due to zero reported perceptual differences from the acoustics.

Data availability. Relevant data are available in a public repository accessible at http://hdl.handle.net/1903/19593.

Chapter VI

Prior knowledge influences cortical latency and fidelity of the neural representation of missing speech

Summary

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 speech-related dynamic 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.

Introduction

The ability to interpret speech elements across interruptions masking a conversation is a hallmark of human communication [171]. In many cases, possessing contextual knowledge poses clear informational advantages for a listener, so as to successfully disengage the masker and restore the intended template signal [135], [139], [172], [148]. Information can typically be obtained from multimodal sources and/or low-level auditory and higher-order linguistic analyses, although it remains unclear how and which factors are most effective in assisting speech restoration under natural conditions. For instance, it is possible to identify cortical network activity profiles consistent with phonemic restoration, the effect where missing phonemes in a signal may be heard [115], [133], in binary semantic decision tasks [148]; still, the description of factors that bias into either of two alternatives in the direction of perception remains unclear. To this end, there is evidence that restorative processes may be influenced by contributions from audiovisual integration cues [173], lexical priming [174], and within the auditory domain, predictive template matching [159] or even intentional expectations about temporal patterns in sound [160], [161].

It is clear that in order to lead to informational gain, potential contributors must be readily accessible before and during missing auditory input. Presumably, the mechanism would involve (i) generation of a provisional template about forthcoming speech, (ii) that the template is 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* [175]–[177]. In some cases, the informational value added by such putative mechanism in ameliorating the neural representation of speech may also involve speeding up cortical processing during integration [178].

Here we test how natural speech tokens spanning over several words may be represented cortically in the midst of masking noise, under varying levels of informational gain added by prior knowledge about the missed element. The low-frequency envelope of speech indexes slow acoustic energy changes over time and is known to entrain and phase-lock neural activity at the auditory cortex, as measured by magnetoencephalography (MEG) and electroencephalography (EEG) [25], [91], [179], [180]. Due to its characteristic timescale, the envelope is also related to prosodic attributes such as syllabic lengths and loudness, which themselves may include intonation, rhythm and stress cues. We hypothesize that by presenting the same verse units several times, it is possible to manipulate listeners' ability to develop detailed predictions about forthcoming elements in long speech sentences, plausibly forming a template about them, that may serve for a form of point-wise matching at a later time when spontaneous maskers disrupt the same parts of the narrated story. This implies the possibilities that (a) the template about the envelope may be decoded from cortical signals in response to noise, and (b) because the template must be have been present in advance, that the mechanism could be facilitated at subsequent times by speeding up processing of the incumbent envelope token, at least indirectly. We apply neural coding methods to neural responses in order to reconstruct the original verse template envelope [181], an approach that has been successfully applied in auditory electrophysiology [16], [18], EEG/MEG [25], [55], [91] , electrocorticography [90], [148], [180], and fMRI

[182]; and also to provide estimates of the forward stimulus-response mapping [25], [91] under normal speech conditions. From such decoding performance we assess the extent to which prior knowledge about speech may enhance endogenous representations that assist restoration of intended speech signals. In the case of forward models, we address the cortical latencies involved in natural speech encoding under the same conditions. The latter is a relevant question at least because (i) reduced processing times have been observed in visual contexts that facilitate integration of detailed predictions with auditory representations of incoming speech [172], [178]; and (ii) within timescales of the order of seconds or minutes, task-related adaptive changes can occur to the shape of stimulus-response mappings, which would in turn suggest mediation by cortical plasticity [98], [183] as a biophysical basis for restorative mechanisms given the present task demands.

We provide evidence that the speech temporal envelope may be better reconstructed if listeners have obtained sufficient knowledge about a particular speech sequence, and this effect extends to cases in which they are presented with noise instead. The data also show that cortical latencies in natural clean speech processing can be reduced by the order of milliseconds under similar conditions. Overall, the results suggest that formation of online templates about low-level features of frequently experienced speech may facilitate more efficient neural representations, by means of faster encoding and improved access to endogenous modulations time-locked to expected but missing speech, thus assisting its restoration.

Results

Reconstruction of missing speech timeseries from noise with context. Fixed-duration spectrally-matched static noise bursts were used to mask word sets within a narrated story. Each noise probe was designed to have the same spectral composition over time as the replaced speech segment (Fig. 1A) but 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 entrain auditory cortical activity as 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 using unmasked speech and reverse correlation to establish an optimal mapping from cortical activity to the original speech envelope. To test whether the acoustic presence of a target is a strictly necessary condition for such reconstruction, the listeners were exposed to extensive repetitions of some of the speech, and less extensive repetitions (or none at all) to the rest. Sentences that were maximally repeated over the hour-long session (Fig. 1B, left) resulted in greatest relative performance in reconstruction of the envelope of the missing speech: approximately 25% of the performance for the actual speech presented free of masking. Lesser amounts of repetition resulted in further reductions to relative performance down to baseline floor level for masked speech with which the listener had little or no prior experience. Because relative performance measures include data entries from clean speech reconstruction as references, it is important to verify whether reconstructions from noise alone independently reveal similar effects. Absolute effect sizes of repetition in reconstruction of the missing speech envelope were thus confirmed to display a similar pattern as with

relative performance (Fig. 1B, right). To determine whether decoding success of the linear model of the envelope did significantly change across conditions, the Mauchly test of sphericity was run to evaluate whether corrections would be necessary for a posterior repeated measures model. Results for independent reconstructions using exclusively noise-derived independent scores showed that this condition was not violated in the absolute effect of envelope reconstruction in noise epochs ($\chi^2(5)=5.409$; $p=0.368$). The subsequent four-level repeated measures ANOVA with subject and verse as predictors resulted in a significant main effect of repetition ($F=3.332$; $p=0.023$), with no interaction from subject ($F=0.411$; $p=0.726$), no interaction from verse ($F=0.622$; $p=0.603$), and no three-way interaction between repetition, verse and subject ($F=1.229$; $p=0.304$). Post hoc comparisons using the *t*-test with Bonferroni correction indicated that the average effect size at High repetition rates was significantly different than that at the Control condition ($t(34)=4.319$;$p<1.3\text{x}10^{-4}$), and Low repetition rates ($t(34)=3.918$;$p<4.1\text{x}10^{-4}$).

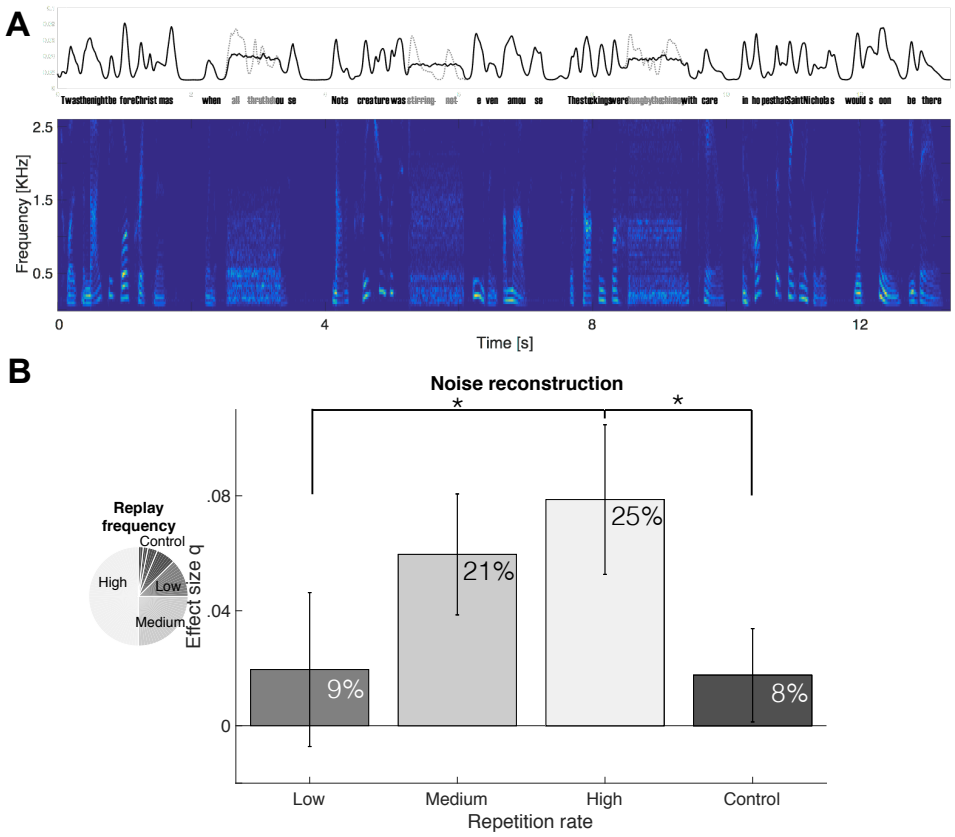


Figure 1. Cortical reconstruction of acoustically missing word-level speech envelope from noise by repeated replays of narrated story. (**A**) A set of speech materials from a poem were repeatedly presented to 35 listeners, but every 4-5 s the signal was replaced with spectrally-matched noise (three instances shown in spectrogram, bottom). This manipulation leads to loss of critical temporal modulation related to the missed words, as shown by the slow envelope (top). (**B**) For repeatedly presented identical material, over 30, 15, 7.5 minutes or less out of an hour-long MEG recording session (left), the missing dynamic speech envelope was reconstructed from responses to static the noise maskers, with performance up to 25% of that obtained under clean conditions (insets, right). This effect on relative performance was not due to the fractional contribution from clean speech reconstructions, as analog results were reproducing by using the absolute reconstruction effects only (right), by assessment of independent performance by noise-trained decoders only, suggesting influence of prior experience in low-level sensory encoding of the temporal envelope over connected words. Error bars indicate confidence intervals for the means (Bonferroni-corrected α-level).

Expedited auditory cortical processing of frequent natural speech replays. temporal response function (TRF) is a functionally informative statistic which can be used to predict the neural response to a given stimulus. When applied to natural sound conditions, it reveals information similar to that of evoked responses to pure tones (identifiable peaks with different polarities at specific latencies, corresponding to distinct neural sources and processing stages) but directly derived from the neural processing of the speech [25], [55]. We examined the effect of extensive prior experience on the TRF temporal structure in general, and a specific peak, the TRF_{100}, occurring 100-150 ms post envelope change in particular(Fig. 2A). A significant latency shift of 5.3 ± 2.2 ms was observed for $TRF_{100\text{-}High}$ versus $TRF_{100\text{-}Control}$ peaks ($t(33)=2.387$; $p=0.023$) indicating that occurrence of this processing cortical may become expedited for listeners, compared on a within-subject basis (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.

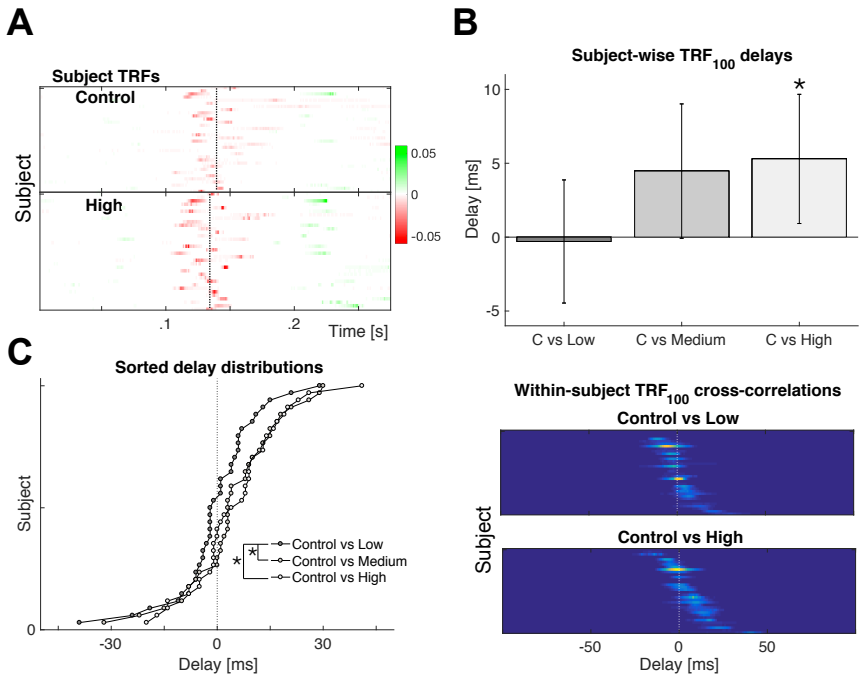


Figure 2. Frequent repetitions of natural speech speed-up their cortical processing. (**A**) Temporal response functions across participants reveal a common cortical processing step about 100-150 ms after unitary variations in the speech envelope, referred to as TRF_{100}. (**B**) Depending on context, the same processing component step may occur at different times at the millisecond stage in high-resolution recordings, with processing of frequently-repeated speech occurring about 5 ms earlier than with 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**) These shifts are obtained by cross-correlating TRF_{100} signals obtained per condition in each subject.

Discussion

The perceptual phenomenon of sensory restoration relies on inference of the missing sections from a sensory signal. The results demonstrate that auditory cortical activity

possesses critical envelope information to reconstruct missing fragments of speech replaced by noise, but only when previously and repeatedly exposed to the missing speech. Results suggest that access and maintenance of a detailed representation of the stimulus, under a template format compatible with the acoustic envelope, is enabled by prior experience, which may also additionally speed up cortical processing time, and together, point to the generation of a time-locked neural activity pattern consistent with the expected but absent sensory input. These findings complement those from designs based on perceptual reports at the phonemic level (e.g. <200 ms), suggesting that acoustic delivery is not a necessary condition for spectrogram reconstructability when interpretation of a phoneme is actively ongoing through a noise [148], as long as the immediate acoustic context is present. These results imply that neural activity matching processes must rely on endogenous activity, possibly as top-down restorative modulations of auditory cortex populations [144], [145]. Our data is consistent with the notion that this activity can be influenced by prior learning and storage of speech information, at the level of its explicit temporal structure. Under this interpretation, enhanced listeners' expectations about forthcoming speech tokens may predispose them to restorative encoding, but when contextual information is poor or insufficient, neural dynamics default to failure in predicting of the missing stimulus. Spontaneous neural background activity known to influence perceptual processing in general, includes the ability to entrain to a complex, natural signals such as speech [184], to optimize behavioral performance of detection tasks [185], or even robustness of an illusory experience [149].

On the plausibility of auditory memory involvement

Besides neural coding, adaptive capabilities of auditory cortical areas include analysis and storage of sound features relevance [186]. This process requires that memory traces be hold, in a format that is considered to develop from a low-level sensory code, held in register by up to 15 s, to categorical terms that are more efficient for storage over the long term [187]. While in sensory format, storage has been argued to assist in the ability to restore missing fragments of a sound source, e.g. as a internal replay of the fragment [188]; also, other potential perceptual effects of memory-based reactivations over auditory object representations, including attention, are an area of current research [189]–[191].

The auditory effect studied here can be considered to belong to the multimodal class of *attractive temporal context effects* [192], a group of facilitatory mechanisms including perceptual hysteresis [193], [194] and perceptual stabilization [195] effects in the vision literature. These are considered critical for improving invariance in the face of external demands imposed by discontinuously fluctuating, broadly cluttered environments. Conceptually, this group 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 alternative perceptual manifolds for which neural activity has not yet been adapted [192], [194]. These may include semantic satiation effects, namely the subjective experience of increasingly meaningless words after fast and prolonged repeats [196], [197].

On access and format of stored auditory representations

Over repeated sound stimuli, attractive contextual effects may rely on forms of implicit auditory memory as they are considered to intervene regularly in sensory and perceptual encoding [198]. A clear example is the improved detection after sequential presentations of arbitrary noise structures, and its time-locked sensory potential covariates [199], [200]. Foreknowledge of acoustic features may adapt listeners to a likely communication source, as demonstrated by perceptual facilitation when advance notice about the identity of a forthcoming instrument play is given [201], and by preferential activation in auditory association areas specific to speaker familiarity [202]. The notion that higher 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 lobe areas, and over time activity can be mediated by structures in inferior frontal cortex instead [203]. Evidence from these studies is consistent with the hypothesis of variable memory trace formats, where high temporal resolutions may be available for readout at sensory buffers, and coarser elsewhere at stores encoding for categorical higher-order input features(cf. [204], [205]).

On the subjective conditions of listening in noise

With regards to the cognitive state of listeners during masking, it is relevant to address whether the findings are consistent with conditions that normally lead to auditory imagery processes, which are (distinctly) analog to perceptual restoration phenomena. In masked circumstances, sensory imagery is postulated to involve ‘schemata’ or prior abstractions actively formed with perceptual input, better resolved with increased

familiarity and which remain online while an expected stimulus fails to be presented [206]. For these purposes, auditory imagery is defined as the persistence of an auditory experience without prompting by direct sensory input [207]; the methodological implication is its existence is judged either directly by subjective reports, or indirectly by tasks and measures hypothesized to involve imagery with reasonable probability [206]. This latter approach comprises the study of conditions or stimuli that may automatically evoke auditory imagery, including substantial prior experience.

Our findings suggest that among these indirect measures it is plausible to include perceptual coding principles of the missing envelope of natural. Behaviorally, it is consistent with findings that the prevalence of auditory imagery episodes depends on the level of familiarity with original sound pieces [208], in natural sound classes (e.g. speech versus music) [209]. 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 [210]. While there is some agreement that imagery and rehearsal, a related process, of natural complex sounds may be subserved by auditory association cortex areas (see reviewed in [206]), evidence on similar activation at primary areas is mixed (cf. [211]). There is dual evidence on the format of representations sustained during active rehearsing, under both auditory-specific (sometimes termed ‘echoic memory’) and modality-general codes; these two types have been shown to occur each over distinct locations on superior temporal cortical areas, over distinct timescales as transient (<5 s) versus sustained phases respectively [209], [212]. Therefore, our data are 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 [213]. This is also supported by findings of retrieval processes in vision and hearing that involve reactivation of sensory regions active during perception [214], something in addition found with auditory verbal imagery [215], [216] thus pointing to the notion that both involve overlapping processes [206].

On the low-level envelope representation during masking

Our suggestion that a key structural property of natural sound encoding lies with the acoustic envelope representation, is compatible with preservation of a temporal coding scheme in auditory imagery based on prior experience as a necessary context. However, while formation of auditory 'images' may entail activity consistent with that elicited by original sound input [217], preservation of properties such as temporal acuity of original stimuli may be deteriorated under imagery depending on factors such as context and experience [218]. This finding was consistent with our relative effect sizes in the ability to reconstruct of missing speech, which disappeared for relatively novel stimuli. In this sense, frequent "refreshing" echoes the *auditory memory reactivation hypothesis* which states that storage of individual sound features occurs embedded in the context of neighboring patterns and sequences that can be representable by the auditory system as regularities. Reactivation is then the automatic process whereby variable sound input is matched to constancies extracted previously, and proximity between prior 'rule' and current 'update' tokens increases memory likelihood [205]. This description, originated from oddball sequence studies, has direct analogy in the present design because across verses, the ability to form regularities differs by uneven repetition rates, and therefore verse sequences may differentially reinstate prior constancies attained throughout the

session. In this interpretation, envelope features over speech preceding a masker serve as referents enabling translation of verse regularities, to be learned and represented over the course of the experiment, into specific values under the same feature format [205], as is the acoustic envelope. This does not preclude that additional stimulus features may likely be extracted and synthesized along the timing information represented in the envelope [219], including higher-order linguistic elements. The idea suggests that restorative processes may be inclusive of learning or storage strategies of linguistically-informative template in formats alternative formats to the envelope (e.g. [172]). Although outside the scope of this study, it is likely that mechanistic accounts of the restoration effect may invoke multiple levels of language analysis. The envelope correlates chiefly indicates timing of specifics phonemic utterances in natural speech [91], [220], with evidence for restoration pointing to adaptive use of their temporal cues in assisting real-time, natural listening conditions.

On adaptive changes to envelope encoding

The accompanying effect that cortical processing timelines were changed under the same circumstances that promoted restoration suggests that active, task-related endogenous changes may be present in order to optimize low-level envelope processing with relevant experience. Plausibly, 'speeding up' is related to increased excitability among populations normally active at the later stages of the immediately preceding processing step (presumably as shown by the TRF_{50} component). This may have a modulatory effect on the early low-level analysis stages (or be a consequence of facilitation already occurring there) – something that could help improve prediction over representational formats held by auditory areas. Determining conditions under which the acoustic

temporal envelope is relevant to initiating this endogenous process, may in the future result in the technical ability to provide real-time noninvasive indices of the subjective states by which a person maintains in register a template auditory pattern. Overall, the results manifest the brain's ability to form a model of a speech scene, independently of feed-forward, bottom-up sensory information, but driven by expectations and learned experience in general [221]. It will be interesting to address whether this may assist in strategies for potential stimulation principles when seek to circumvent some derived auditory peripheral damage, as in prosthetic devices.

General methods

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. Due to excessive artifact caused by misfit within the MEG helmet, data from an additional subject was not included. Each subject received monetary compensation proportional to the study duration (approximately 1.5 hours). Conduction of the experiment protocol was approved by the UMCP Institutional Review Board, and all experiments were performed in accordance with its relevant guidelines and regulations. Before each study session, informed written consent was obtained from participants.

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) from an online database (http://archive.org/details/AVisitFromSt.Nicholas-ByClementClarkeMoore-

NarratedByGrantRaymond. All fourteen stanza (from now on, stimuli) in the poem were separated (see Appendix for materials) and silence gaps within each were reduced so as to be matched in duration (range: 13.1 – 13.6 s), and then presented into 4 blocks. For the first block, 64 stimuli were used by repeating several times those in the poem's first half. A 'High' frequency stimulus was chosen by selected a stanza and repeating it for half of cases (32/64), and this was similarly done for another 'Medium' and 'Low' frequency stimuli, in a quarter and eighth of cases, respectively. The remainder of the block was filled with 'Control' stimuli, namely the four remaining stanza 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 in the same stimuli used as in 1 and 2 respectively, but with randomized order again. The procedure was recreated de novo for each subject resulting in a total of 35 different stimulus sets of about 1 hour each in total duration. Importantly, the choice of stimuli at a given repetition level was titrated across participants, resulting in 7 groups of 5 listeners each that underwent the same 'High', 'Medium', 'Low', and 'Control' stimuli selection.

For each stimuli, 2–4 spectrally-matched (SM) 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 syllable onsets times, as per the envelope rising slope maxima. An expected 768 noise probe samples were presented per experiment, and each was individually constructed by randomization of phase values across the specific frequency-domain phase information contained in the underlying speech stimulus occurring at the same time as the masker noise, yielding a noise with

equal spectral amplitude characteristics[222]. The original speech content occurring during the same time was removed and substituted by the respective SM noise, at a power signal level matching that in the clean original. Subjects listened to the speech sounds while watching a silent film to keep the participant engaged. To maintain their attention upon the auditory stimulus, after each probe, they were report via a button press whether they understood what the speaker meant to during noise.

Data recording. We record responses to selected speech sequences by using MEG, a non-invasive neuroimaging technique optimally retrieving neural activity from human cortex regions such as the auditory cortex on the temporal lobes. Such recordings may reflect direct entrainment to speech low frequency modulations, namely its acoustic energy envelope, with remarkable temporal resolution [25].

Data processing. *Pre-processing and sensor rejection*. The time series of *K* raw recordings $s_k(t)$ from the MEG sensor array (sampling frequency 1 KHz) will be submitted to a fast implementation of independent component analysis [223], from which two independent components will be used as surrogate reference channels for environment noise reduction purposes. Independent components will be selected if they contain the largest proportion of broadband (0-500 Hz) power; this selection will be done by finding the independent component yielding the most power at each spectral bin (of fixed linear size, determined by dataset), and then computing the histogram of independent components that most frequently outnumbered all others in power across the spectrum. Because spectral bins are linearly spaced, and given the 1/f power spectrum of typical MEG fluctuations, this

approach weighs favorably unusual components that consistently show extreme power at higher frequencies.

Environmental noise sources arising from unwanted electrical signals not related to brain activity of interest will be reduced by time-shifted principal component analysis (TS-PCA). This technique discards environmental sources that have dissimilar convolutive properties when they mix at reference sensors in the EEG system, in contrast with the convolutive properties of sources that mix at the data sensors in the array[105]. Provided that reference sensors record noise and no primary sources of interest, such mismatch is exploited as a basis for rejection: projections of recordings from the brain sensor array which do match in their convolutive properties with those from reference sensors recordings are removed via PCA. We set TS-PCA parameters to N=200 taps (equivalent to the range ±100 ms at the original sampling frequency), and regressor principal components whose variance amount to less than 10^{-6} times that of the first component will discarded as negligible for numerical purposes. This combination of parameters is commensurate with a reduction to less than 3% of residual noise for a simulation with three reference sensors in a MEG system [105]. Signal delays introduced by the TS-PCA procedure will be corrected.

Sensor noise. Sensor-specific sources of unwanted electrical signals unrelated to brain activity of interest are reduced by sensor noise suppression (SNS). We substitute each channel recording by its projection formed by the orthogonal basis span of all other channels[107]. This method exploits redundancy from a dense array -where the number of sensors exceeds the number of brain sources- by

rejection of sensor-specific components whose presence cannot be explained by the redundancy manifold laid by data from in other channels – potentially including sensor-specific noise from those channels themselves. Collectively, this separation does not necessarily eliminate all sensor-specific noise, since at each substitution it can be imported from other sensors, yet this may promote instances where such redistribution will add these components in incoherent manner and thus become attenuated [108]

Data analysis. To assess low-frequency cortical entrainment, recordings will be band-pass filtered between 1 and 8 Hz with an order-2 Butterworth design, correcting for the group delay created by the filtering procedure. A data-driven spatial filter will be derived, following trial-by-trial repeatability as the basis for a source-separation model[108]. Spatial filter coefficients from the most reproducible signal component (i.e. having greatest evoked-to-total power ratio) in individual subject data, as obtained by denoising source separation, will be applied to sensor data as a weighted sum forming the resulting virtual MEG output channel.

$$s_1^{DSS} = \sum_{k=1}^{K} a_{k,1} s_k(t)$$

This approach effectively improves signal quality, and the data-driven virtual sensor distributions will be estimated based on recordings to clean speech epochs only. This will ensure that neural activity recorded during noise probes has been projected to the span determined by the neuromagnetic source that represents the most reproducible processing modes of the original speech template stimulus.

Stimulus reconstruction. The ability to reconstruct speech from MEG epochs will be assessed. Aside from the component described in equation M.1, sources from the next three top reproducible components will be obtained and submitted to a trained linear decoder estimation procedure (Figure 4), mapping from these sources back to original template stimulus. These components are considered reproducible signals in contrast with the bottom rank bottom components, which may serve as a reference devoid of stimulus-related activity. In either case reconstruction produces a timeseries whose similarity with the original envelope was assessed via Pearson's *r* correlation coefficient. These scores were contrasted, after reproducible (r_e) and, separately, reference (r_f) response signals, for (1) reconstructions of clean speech from neural activity following clean speech; (2) reconstructions of clean speech from neural activity following noise. The referencing procedure was introduced to obtain a necessary baseline in decoding performance given that timeseries' lengths varied across conditions as a result of the different repetition rates and verses involved, something that we observed may produce positive biases in *r* for shorter sequences, irrespectively of underlying relationship to the stimulus to be decoded. To compute absolute reconstruction effect sizes, each of the Pearson's *r* pairs (reproducible versus reference activity) were transformed to Cohen's *q*[224] indices 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 were computed by the fraction q_2/q_1 of absolute effect sizes given the stimulus presentation conditions above.

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

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

where $\varepsilon(t)$ is the residual waveform, which is the contribution to the evoked response not explained by the linear system. As stimulus representation, the envelope was extracted by extraction of the instantaneous amplitude of each channel's analytic representation via the Hilbert transform[225], their sampling rates were reduced to 1 KHz and transformed to dB-scale.

Statistical analyses. For reconstructions, repeated measures ANOVA were run across all four levels: 'Control', and 'Low', 'Medium', and 'High' repetitions, in order to detect any differences between their related means overall. For temporal response functions, in each participant, activity related to the TRF100 component was obtained from the 100-200 ms window and cross-correlations were performed on 'Control' versus all other repetition conditions. The resulting peak delays were submitted to a non-parametric one-tailed two-sample Kolmogorov-Smirnov test for differences in the underlying delay populations.

Chapter V: Conclusions

A cornerstone of human auditory cognition is the dynamic interplay between the sensory and perceptual bases of sound encoding, as performed by the auditory cortex – a key structure dedicated to analysis and inferences related to hearing.

The study of each basis presents advantages and challenges. With regards to sensory encoding, access to a wealth of physiological characterizations about the stimulus-to-response mapping will continue to provide invaluable means to establishing a biological basis for computation. However these means almost always involve non-human animal models, implying a relatively limited range of cognitive tasks available, especially those related to speech.

On the perceptual side, incentives exist in attaining a comprehensive understanding of subjective states and of processes relating a human listener more efficiently to her environment. This because it is crucial information for pressing mental health issues and communication disorders. Yet, assessments of "inner experience" can be problematic, as it involves activity difficult to tag in time and can be prone to intractable amounts of variability between subjects.

This series of projects addressed these issues in part. First, the studies present a framework where analysis and representation of *basic* versus *complex* sound encoding models are, at the cortical level, substantially closer than previously assumed. Findings of regions where models overlap suggest a means to extract sensory, domain-general processing stages. Such models may also mirror the structure of encoding models derived across animal models in the electrophysiology literature. Therefore, an avenue for further biologically-informed hypotheses to enter neuroimaging research lies ahead in exploiting their joint accessibility to models of spectrotemporal coding.

Second, access to subjective information can be addressed by approaching perceptual coding models, which posit more several alternatives in representing the same stimulus. Across this manifold, the search for a neural representation closest to the receiver's experience can be reduced for instance by setting up a sensory context that suggests what to expect and when. This tactic is compatible with the study of perceptual restoration phenomena, a set of strategies to fill-in missed information. In these conditions, temporally-patterned rhythmic sounds were found to be represented by auditory cortex also in to cases when they are only perceived to be so but indeed physically absent. This implies a means of access to endogenous, dynamic perceptual representations following the subjective experience of sound.

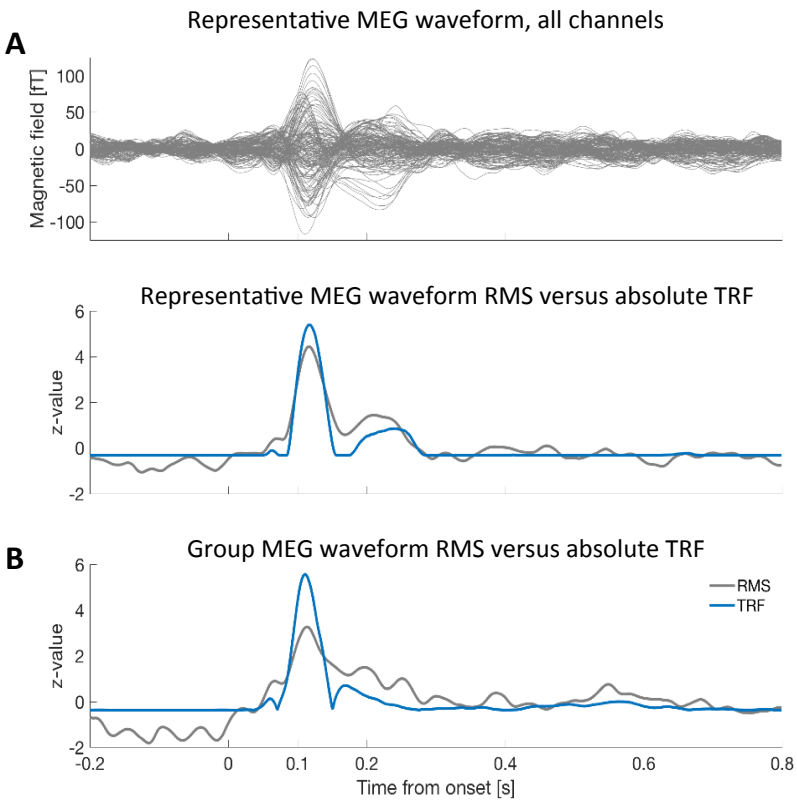
Last, in realistic conditions natural speech sounds are sometimes clear and predictable, and by times ambiguous or uncertain. Dual sensory and perceptual coding mechanisms may aid to sustain stable hearing in the face of disruptions unrelated to our acoustic interactions; one such way is to speed up auditory cortical processing when conditions allow to infer from prior knowledge what occurs next, and to possibly invest that gain at later instances where uncertainty demands further explorations through the tree of perceptual possibilities. There is now evidence that adaptive modifications to the sensory coding model, in the form of cortical processing time facilitation, may stem from comprehensive knowledge about the speech sequence being listened to. The finding is accompanied, from a perceptual coding perspective, by the increased likelihood for restoration of prolonged missed speech sounds encoded in cortical activity. Overall, the results demonstrate how sensory and/or perceptual coding approaches may further expand enquiry windows about a listener's personal experience of the communication-rich soundscape.

Appendix A

Supplemental information for "Functional significance of spectro-temporal response functions obtained using magnetoencephalography"

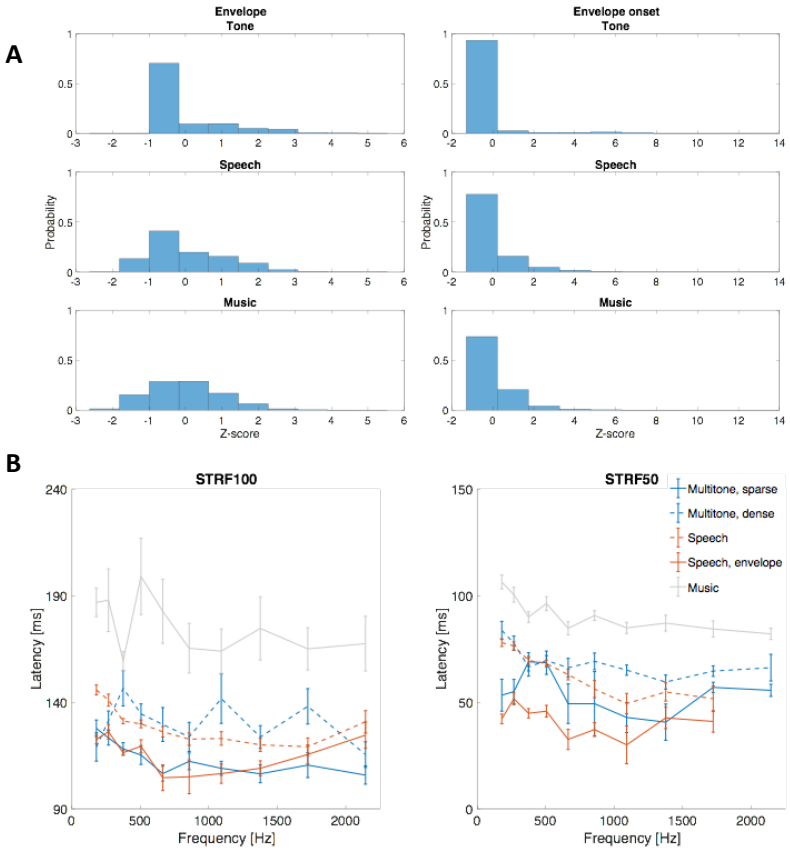
Relevant data are available in a public repository accessible at

http://hdl.handle.net/1903/19601



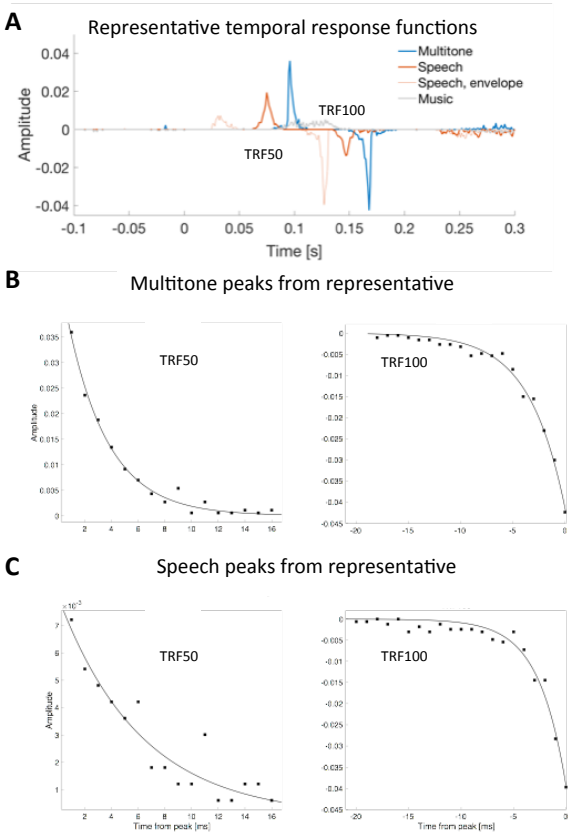
Supplementary Fig 1. Example of equivalence between standard evoked potentials and temporal response function components. (**A**) Complete MEG sensor dataset from a representative subject (R1946) following tone delivery shows a typical 'butterfly plot' waveform pattern, when the data are processed as standard evoked potential (bandpass filter 1-15 Hz, averaging, and baseline correction, top). The standardized root mean square (RMS) of the sensor array reveals close similarity to the absolute value of the sparser temporal response function obtained for this subject via reverse correlation. (**B**)

Consistency of the TRF with the classic RMS is also apparent at the group level (*N*=15), albeit at improved contrast by means of both spatial filtering and cross-validated predictive model techniques across participants.



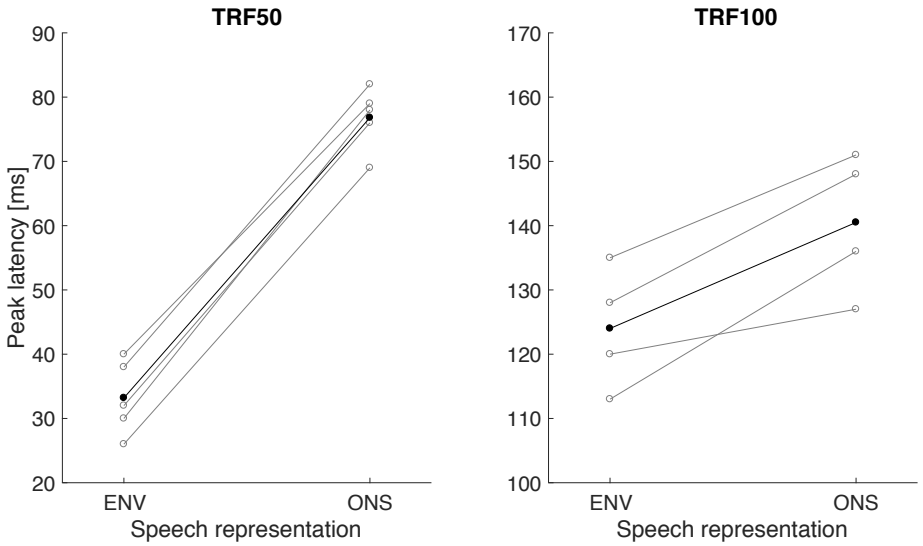
Supplementary Fig 2. Stimulus and transfer functions differences across stimulus classes. (A) *Differences in distributions of envelope and envelope onset representations*. When represented by their envelope, stimuli have distributions that show variable spread over different classes, as a consequence of their different statistical structure. The envelope-onset procedure reduces its extent by referencing rises in acoustic energy to its immediate preceding context, increasing the similarity among stimuli classes. **(B)** *Differences in transfer function delay distributions*. Average group latencies of the $STRF_{50}$ and $STRF_{100}$ components across the random multitone pattern (*N*=15), speech (*N*=12), and music processing (*N*=15) datasets shown. Across the three studies, lower frequencies entail longer delays. Dense multitone

patterns entail longer delays (sparse: 2 tones per second, dense: 10 tones per second). For speech, envelope (thick red) and envelope onset (thin red) representations exhibit a relative delay that is greatest at the early $STRF_{50}$ but reduces for $STRF_{100}$. Error bars indicate the standard error of the mean. Music shows the longest delays, although a subject group confound cannot be ruled out therefore absolute value comparisons across studies are for reference only.

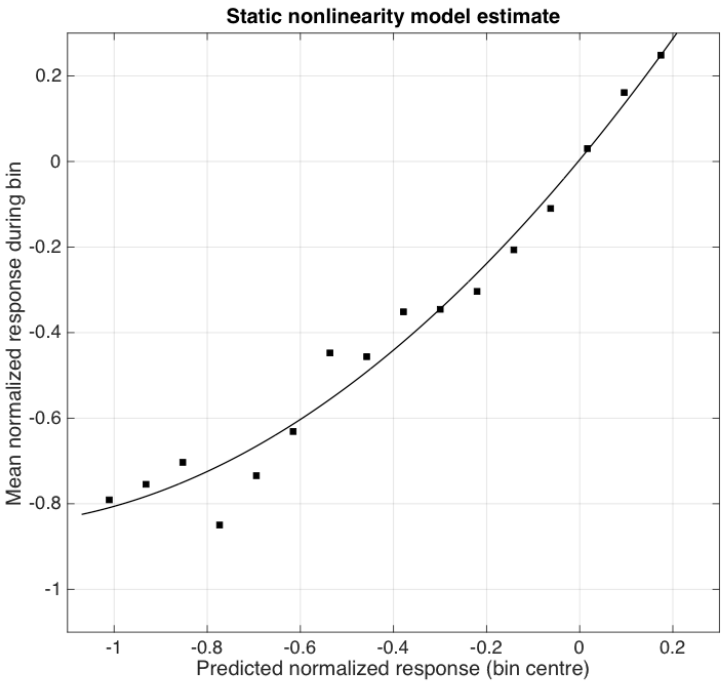


Supplementary Fig 3. Models of subject temporal response function principal peaks. (**A**) TRFs from the same subject display the timing of principal activity related to different stimulus features and classes, some of which are consistent with exponential growth/decay models. Processing timescales differ according to both stimulus feature and class being modeled. (**B**) Within the first 200 ms following both tone and speech stimulus onsets, early and late peak deflections of the neuromagnetic signal may be described as transient exponential decay/growth curves (cf. Fig 3C) for the one subject in which both stimuli were tested. *Top:* Following tone onsets, both deflections were fit by exponential models, achieving time constant estimates at high goodness of fit values (TRF_{50}: τ=3.1

ms with 2.8-3.4 ms CI-95%; R^2=0.989; TRF_{100}: τ=3.5 ms with 3.2-3.8 ms CI-95%; R^2=0.988). *Bottom:* Analog cortical activity described by the TRF and signal envelope from natural speech reveals expanded and similar temporal processing windows at early and late latencies respectively (TRF_{50}: τ=6.2 ms; 5.0-8.2 ms CI-95%; R^2=0.892; TRF_{100}: τ=2.6 ms; 2.3-3.1 ms CI-95%; R^2=0.967).



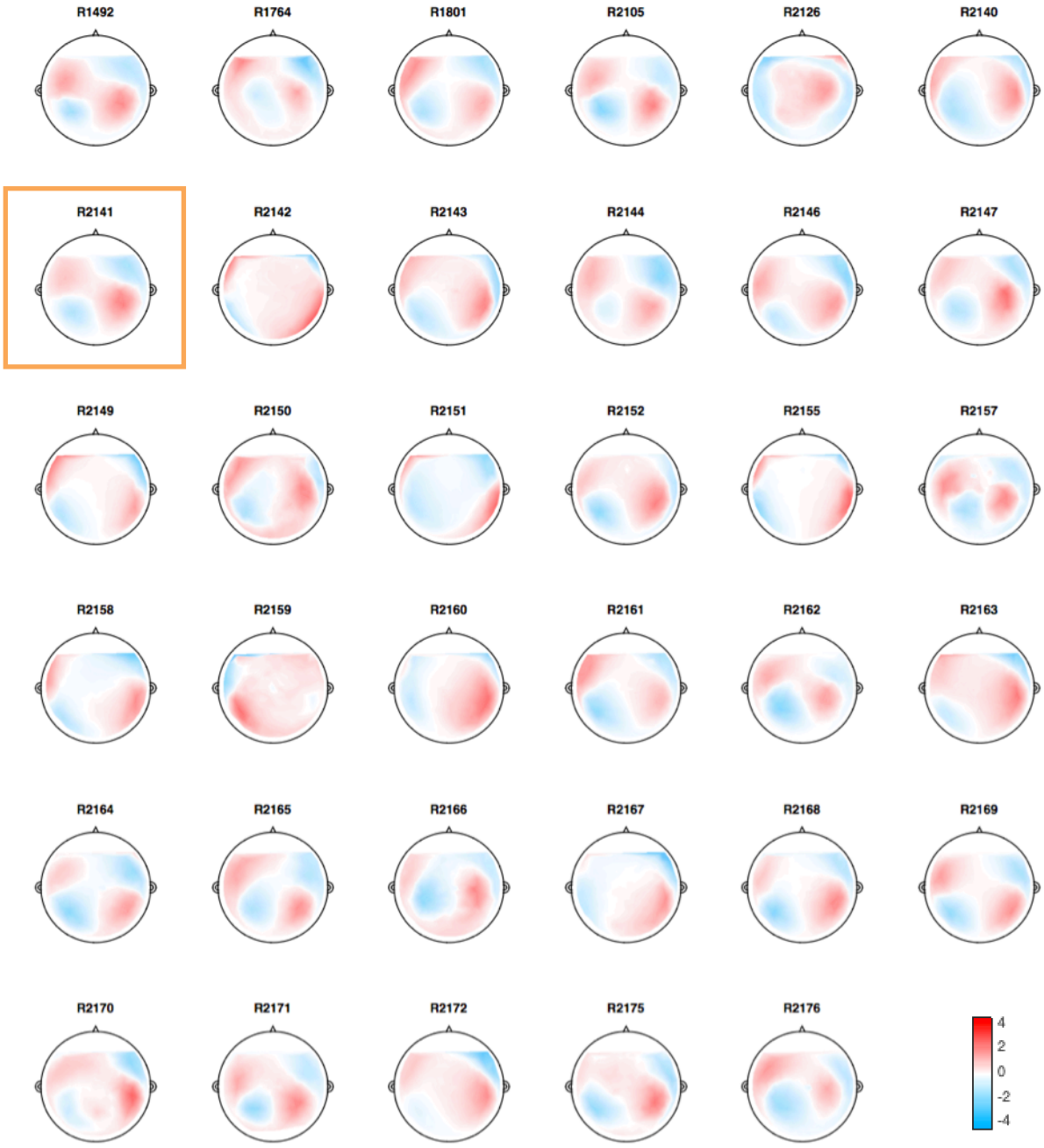
Supplementary Fig 4. Representation format transformed from early to mid latency speech processing at individual level. In single subjects, TRF components were timed differently for the speech envelope and envelope onset representations, as indicated by grey lines. The difference between early stage components (left) was about 43 ms, which aligns with the average delay (black) between maxima in the representations. Individual response functions showed a considerably reduced delay at mid latency stage (after 100 ms, right), thus suggesting a transformation to the acoustic envelope-based representation by this time.



Supplementary Fig 5. Addition of static nonlinearity to multitone response properties. As intrinsic nonlinear response features may be potentially precluded by linear analysis, a static nonlinearity was estimated given the MEG response and response prediction by the linear STRF model approximation (data as in Fig 1B). The STRF prediction timeseries was binned according to its magnitude range (values normalized), and an average computed across all values in the MEG response timeseries that map to each bin. Although the graphical procedure produces an accelerating function of the linear contribution of improved goodness of fit over the linear prediction (R^2=0.972 quadratic; R^2=0.940 linear), cascading the linear prediction with this function marginally improves power explained by a <1.5% margin, suggesting that the linear portion accounts sufficiently for the original model's predictive power.

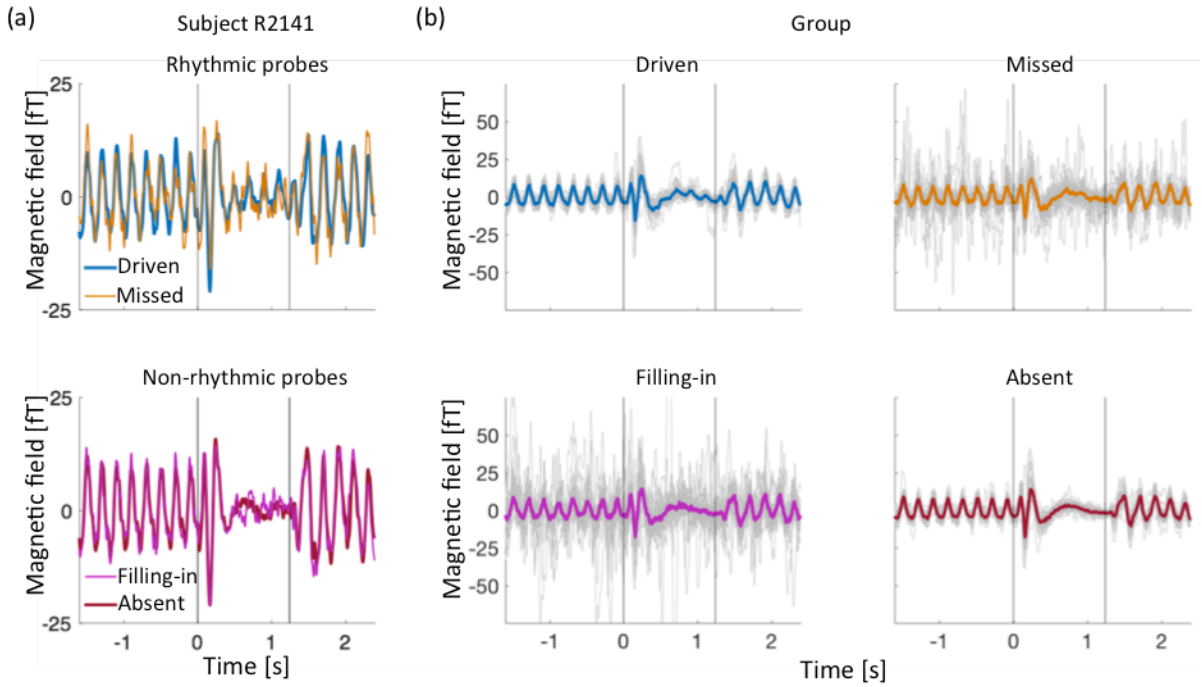
Appendix B

Supplemental information for "Dynamic cortical representation of perceptual filling-in of missing acoustic rhythm"

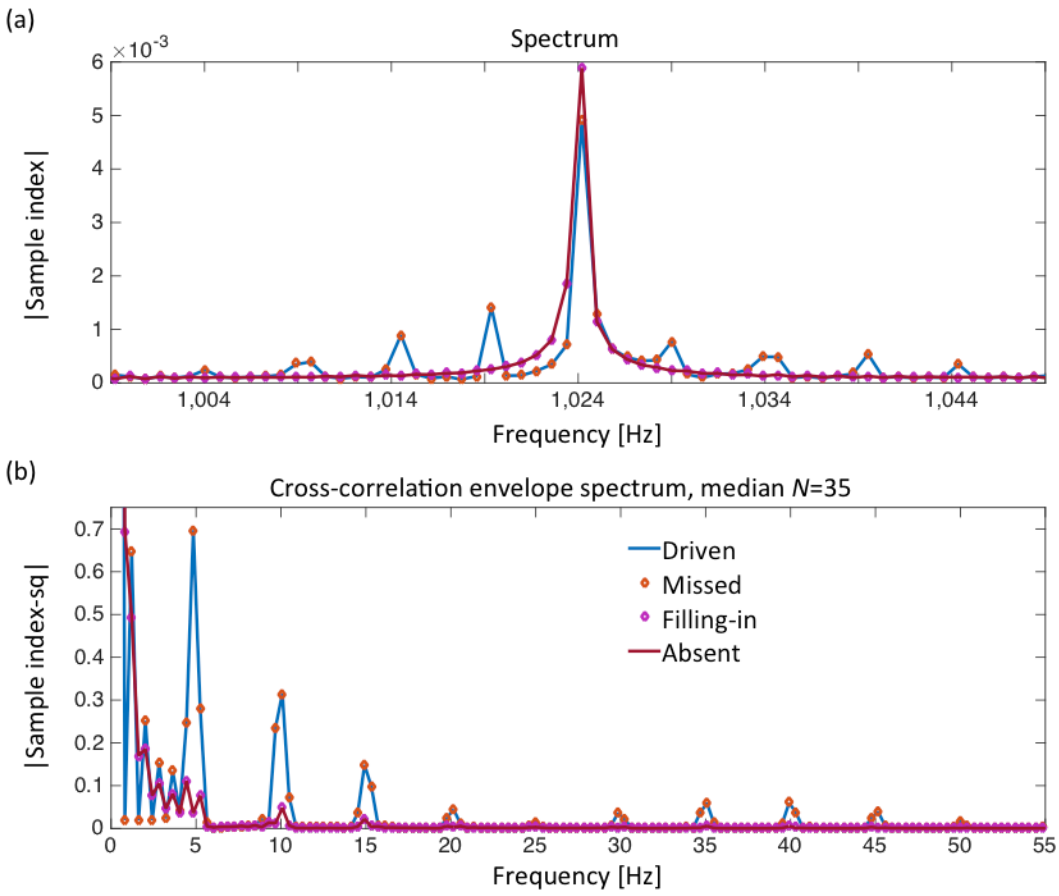


Supplementary Fig 1. Spatial filters associated with auditory steady-state responses. Spatial filters were obtained from participant datasets using responses from the unmasked

acoustic pulse train only; the magnetic field distribution corresponding to that filter is displayed for each subject (*N*=35). The procedure constructs a fixed virtual MEG sensor on the basis of the most reproducible component of each participant's aSSR; neural dynamics during noise probes are investigated using this virtual sensor. The large majority of the field distributions are consistent with MEG evoked potentials originating from bilateral auditory cortex. The distribution from the representative subject in Fig. 1 is highlighted. Units are in *z*-scores.



Supplementary Fig 2. Neural representations of a rhythmic pattern embedded in noise. (**a**) Representative stimulus-locked neural activity as measured by virtual MEG sensor. After a transient noise-onset response, the acoustic presence of a rhythmic pattern (top) may elicit an auditory steady-state response (aSSR) weaker in magnitude relative to baseline levels; (bottom) the acoustic absence of the target rhythmic pattern entails a similar noise-specific onset response, but an apparent lack of the aSSR depending on perception. (**b**) Median data across subjects with color convention as in as in (A); grey indicates individual subjects.



Supplementary Fig 3. No systematic acoustic influence of ambiguous perception of stimuli. (**a**) Spectral analysis of all rhythmic [respectively, non-rhythmic] noise probes in the experiment, shows that as per stimuli design, spectral content in probes appears virtually identical regardless of a listener's posterior report on their rhythmic [non-rhythmic] content. Spectra predominantly feature the 1024 Hz tone carrier, and FM interactions where expected (color code in (B)). (**b**) To assess for unforeseen random temporal modulations appearing systematically in the probe distributions, each probe trial was cross-correlated with a stimulus segment consisting of the basic pulse train without noise. 5 Hz modulations in the cross-correlation envelope are observed only when the pulse train was present were expected (i.e. rhythmic probes), because signal similarity peaks at periodic lags. Between identical-acoustics probe partitions, tests for pairwise differences with mean different than zero were rejected (paired-sample *t*-tests; rhythmic versus missed, p=0.85; filling-in versus absent, p=0.84).

Appendix C

Supplemental information for “Prior knowledge influences cortical latency and fidelity of the neural representation of missing speech”

’Twas the night before Christmas, when all through the house
not a creature was stirring, not even a mouse.
The stockings were hung by the chimney with care,
in hopes that St. Nicholas soon would be there.

The children were nestled all snug in their beds,
while visions of sugar plums danced in their heads.
And Mama in her 'kerchief, and I in my cap,
had just settled our brains for a long winter's nap.

When out on the lawn there arose such a clatter,
I sprang from my bed to see what was the matter.
Away to the window I flew like a flash,
tore open the shutter, and threw up the sash.

The moon on the breast of the new-fallen snow
gave the lustre of midday to objects below,
when, what to my wondering eyes should appear,
but a miniature sleigh and eight tiny reindeer.

With a little old driver, so lively and quick,
I knew in a moment it must be St. Nick.
More rapid than eagles, his coursers they came,
and he whistled and shouted and called them by name.

“Now Dasher! Now Dancer! Now, Prancer and Vixen!
On, Comet! On, Cupid! On, Donner and Blitzen!
To the top of the porch! To the top of the wall!
Now dash away! Dash away! Dash away all!"

As dry leaves that before the wild hurricane fly,
when they meet with an obstacle, mount to the sky
so up to the house-top the coursers they flew,
with the sleigh full of toys, and St. Nicholas too.

And then, in a twinkling, I heard on the roof
the prancing and pawing of each little hoof.
As I drew in my head and was turning around,
down the chimney St. Nicholas came with a bound.

He was dressed all in fur, from his head to his foot,
and his clothes were all tarnished with ashes and soot.
A bundle of toys he had flung on his back,
and he looked like a peddler just opening his pack.

His eyes--how they twinkled! His dimples, how merry!
His cheeks were like roses, his nose like a cherry!
His droll little mouth was drawn up like a bow,
and the beard on his chin was as white as the snow.

The stump of a pipe he held tight in his teeth,
and the smoke it encircled his head like a wreath.
He had a broad face and a little round belly,
that shook when he laughed, like a bowl full of jelly.

He was chubby and plump, a right jolly old elf,
and I laughed when I saw him, in spite of myself.
A wink of his eye and a twist of his head
soon gave me to know I had nothing to dread.

He spoke not a word, but went straight to his work,
and filled all the stockings, then turned with a jerk.
And laying his finger aside of his nose,
and giving a nod, up the chimney he rose.

He sprang to his sleigh, to his team gave a whistle,
And away they all flew like the down of a thistle.
But I heard him exclaim, 'ere he drove out of sight,
"Happy Christmas to all, and to all a good night!"

References

[1] F. Lopes da Silva, "EEG: Origin and Measurement," in *EEG - fMRI*, C. Mulert and L. Lemieux, Eds. Springer Berlin Heidelberg, 2009, pp. 19–38.

[2] G. Rizzolatti, L. Cattaneo, M. Fabbri-Destro, and S. Rozzi, "Cortical Mechanisms Underlying the Organization of Goal-Directed Actions and Mirror Neuron-Based Action Understanding," *Physiol. Rev.*, vol. 94, no. 2, pp. 655–706, Apr. 2014.

[3] J. Wu and Y. C. Okada, "Physiological bases of the synchronized population spikes and slow wave of the magnetic field generated by a guinea-pig longitudinal CA3 slice preparation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 107, no. 5, pp. 361–373, Nov. 1998.

[4] S. Murakami, A. Hirose, and Y. C. Okada, "Contribution of Ionic Currents to Magnetoencephalography (MEG) and Electroencephalography (EEG) Signals Generated by Guinea-Pig CA3 Slices," *J. Physiol.*, vol. 553, no. 3, pp. 975–985, diciembre 2003.

[5] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Rev. Mod. Phys.*, vol. 65, no. 2, p. 413, 1993.

[6] F. H. Lopes da Silva, "Electrophysiological Basis of MEG Signals," in *MEG: An Introduction to Methods*, P. Hansen, M. Kringelbach, and R. Salmelin, Eds. Oxford University Press, 2010, pp. 1–23.

[7] G. D. Kidd, C. R. Mason, V. M. Richards, F. J. Gallun, and N. I. Durlach, "Informational Masking," in *Auditory Perception of Sound Sources*, W. A. Yost, A. N. Popper, and R. R. Fay, Eds. Springer US, 2008, pp. 143–189.

[8] E. de Boer and P. Kuyper, "Triggered Correlation," *IEEE Trans. Biomed. Eng.*, vol. BME-15, no. 3, pp. 169–179, Jul. 1968.

[9] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, 2005.

[10] A. M. H. J. Aertsen and P. I. M. Johannesma, "Spectro-temporal receptive fields of auditory neurons in the grassfrog," *Biol. Cybern.*, vol. 38, no. 4, pp. 223–234, Nov. 1980.

[11] A. M. H. J. Aertsen and P. I. M. Johannesma, "The Spectro-Temporal Receptive Field," *Biol. Cybern.*, vol. 42, no. 2, pp. 133–143, 1981.

[12] J. J. Eggermont, "Wiener and Volterra analyses applied to the auditory system," *Hear. Res.*, vol. 66, no. 2, pp. 177–201, abril 1993.

[13] J. P. Jones and L. A. Palmer, "The two-dimensional spatial structure of simple receptive fields in cat striate cortex," *J. Neurophysiol.*, vol. 58, no. 6, pp. 1187–1211, Dec. 1987.

[14] J. J. Eggermont, P. I. M. Johannesma, and A. M. H. J. Aertsen, "Reverse-correlation methods in auditory research," *Q. Rev. Biophys.*, vol. 16, no. 03, pp. 341–414, 1983.

[15] J. Z. Simon, D. A. Depireux, D. J. Klein, J. B. Fritz, and S. A. Shamma, "Temporal Symmetry in Primary Auditory Cortex: Implications for Cortical Connectivity," *Neural Comput.*, vol. 19, no. 3, pp. 583–638, Feb. 2007.

[16] D. J. Klein, J. Z. Simon, D. A. Depireux, and S. A. Shamma, “Stimulus-invariant processing and spectrotemporal reverse correlation in primary auditory cortex,” *J. Comput. Neurosci.*, vol. 20, no. 2, pp. 111–136, Feb. 2006.
[17] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, “Robust Spectrotemporal Reverse Correlation for the Auditory System: Optimizing Stimulus Design,” *J. Comput. Neurosci.*, vol. 9, no. 1, pp. 85–111.
[18] F. E. Theunissen, K. Sen, and A. J. Doupe, “Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds,” *J. Neurosci.*, vol. 20, no. 6, pp. 2315–2331, Mar. 2000.
[19] S. V. David, N. Mesgarani, and S. A. Shamma, “Estimating sparse spectro-temporal receptive fields with natural stimuli,” *Netw. Comput. Neural Syst.*, vol. 18, no. 3, pp. 191–212, enero 2007.
[20] R. E. Schapire, “The strength of weak learnability,” *Mach. Learn.*, vol. 5, no. 2, pp. 197–227.
[21] Y. Freund, “Boosting a Weak Learning Algorithm by Majority,” *Inf. Comput.*, vol. 121, no. 2, pp. 256–285, Sep. 1995.
[22] J. Friedman, T. Hastie, and R. Tibshirani, “Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors),” *Ann. Stat.*, vol. 28, no. 2, pp. 337–407, Apr. 2000.
[23] T. Zhang and B. Yu, “Boosting with Early Stopping: Convergence and Consistency,” *Ann. Stat.*, vol. 33, no. 4, pp. 1538–1579, 2005.
[24] M. Sahani and J. F. Linden, “How Linear are Auditory Cortical Responses?,” in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2003, pp. 125–132.
[25] N. Ding and J. Z. Simon, “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening,” *J. Neurophysiol.*, vol. 107, no. 1, pp. 78–89, Jan. 2012.
[26] S. M. N. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen, “Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds,” *Nat. Neurosci.*, vol. 8, no. 10, pp. 1371–1379, Oct. 2005.
[27] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, “Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex,” *J. Neurophysiol.*, vol. 85, no. 3, pp. 1220–1234, Mar. 2001.
[28] S. Kumar and W. Penny, “Estimating neural response functions from fMRI,” *Front. Neuroinformatics*, vol. 8, p. 48, 2014.
[29] A. M. H. J. Aertsen and P. I. M. Johannesma, “Spectro-temporal receptive fields of auditory neurons in the grassfrog,” *Biol. Cybern.*, vol. 38, no. 4, pp. 223–234, 1980.
[30] J. J. Eggermont, “Context dependence of spectro-temporal receptive fields with implications for neural coding,” *Hear. Res.*, vol. 271, no. 1–2, pp. 123–132, Jan. 2011.
[31] F. E. Theunissen and J. E. Elie, “Neural processing of natural sounds,” *Nat. Rev. Neurosci.*, vol. 15, no. 6, pp. 355–366, Jun. 2014.
[32] J. Laudanski, J.-M. Edeline, and C. Huetz, “Differences between spectro-temporal receptive fields derived from artificial and natural stimuli in the auditory cortex,” *PloS One*, vol. 7, no. 11, p. e50539, 2012.

[33] D. T. Blake and M. M. Merzenich, "Changes of AI Receptive Fields With Sound Density," *J. Neurophysiol.*, vol. 88, no. 6, pp. 3409–3420, Dec. 2002.
[34] A. F. Meyer, R. S. Williamson, J. F. Linden, and M. Sahani, "Models of Neuronal Stimulus-Response Functions: Elaboration, Estimation, and Evaluation," *Front. Syst. Neurosci.*, vol. 10, Jan. 2017.
[35] P. Gill, J. Zhang, S. M. N. Woolley, T. Fremouw, and F. E. Theunissen, "Sound representation methods for spectro-temporal receptive field estimation," *J. Comput. Neurosci.*, vol. 21, no. 1, pp. 5–20, Aug. 2006.
[36] P. A. Valentine and J. J. Eggermont, "Stimulus dependence of spectro-temporal receptive fields in cat primary auditory cortex," *Hear. Res.*, vol. 196, no. 1–2, pp. 119–133, Oct. 2004.
[37] M. Villafañe-Delgado, "The Cortical Representations of Speech in Reverberant Conditions," Thesis, University of Maryland, College Park, 2013.
[38] E. Camenga *et al.*, "Cortical Representations of Music in Human Listeners," presented at the Midwinter Meeting of the Association for Research in Otolaryngology, Baltimore, 2013.
[39] Q. Gaucher, J.-M. Edeline, and B. Gourévitch, "How different are the local field potentials and spiking activities? Insights from multi-electrodes arrays," *J. Physiol.-Paris*, vol. 106, no. 3–4, pp. 93–103, May 2012.
[40] B. Gourévitch, A. Noreña, G. Shaw, and J. J. Eggermont, "Spectrotemporal receptive fields in anesthetized cat primary auditory cortex are context dependent," *Cereb. Cortex N. Y. N 1991*, vol. 19, no. 6, pp. 1448–1461, Jun. 2009.
[41] T. P. Roberts and D. Poeppel, "Latency of auditory evoked M100 as a function of tone frequency," *Neuroreport*, vol. 7, no. 6, pp. 1138–1140, Apr. 1996.
[42] S. Greenberg, D. Poeppel, and T. Roberts, *A Space-Time Theory of Pitch and Timbre Based on Cortical Expansion of the Cochlear Traveling Wave Delay*. 1997.
[43] B. Lütkenhöner and O. Steinsträter, "High-precision neuromagnetic study of the functional organization of the human auditory cortex," *Audiol. Neurootol.*, vol. 3, no. 2–3, pp. 191–213, Jun. 1998.
[44] T. P. Roberts, P. Ferrari, S. M. Stufflebeam, and D. Poeppel, "Latency of the auditory evoked neuromagnetic field components: stimulus dependence and insights toward perception," *J. Clin. Neurophysiol. Off. Publ. Am. Electroencephalogr. Soc.*, vol. 17, no. 2, pp. 114–129, Mar. 2000.
[45] A. M. Mäkelä, P. Alku, V. Mäkinen, J. Valtonen, P. May, and H. Tiitinen, "Human Cortical Dynamics Determined by Speech Fundamental Frequency," *NeuroImage*, vol. 17, no. 3, pp. 1300–1305, Nov. 2002.
[46] A. Salajegheh *et al.*, "Systematic latency variation of the auditory evoked M100: from average to single-trial data," *NeuroImage*, vol. 23, no. 1, pp. 288–295, Sep. 2004.
[47] T. W. Picton, *Human auditory evoked potentials*. San Diego: Plural Pub, 2010.
[48] C. K. Machens, M. S. Wehr, and A. M. Zador, "Linearity of cortical receptive fields measured with natural sounds," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 24, no. 5, pp. 1089–1100, Feb. 2004.
[49] S. V. David, N. Mesgarani, J. B. Fritz, and S. A. Shamma, "Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in

primary auditory cortex by natural stimuli," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 29, no. 11, pp. 3374–3386, Mar. 2009.
[50] G. H. Recanzone, "Response profiles of auditory cortical neurons to tones and noise in behaving macaque monkeys," *Hear. Res.*, vol. 150, no. 1–2, pp. 104–118, Dec. 2000.
[51] D. P. Phillips, S. E. Hall, and S. E. Boehnke, "Central auditory onset responses, and temporal asymmetries in auditory perception," *Hear. Res.*, vol. 167, no. 1–2, pp. 192–205, May 2002.
[52] X. Wang, T. Lu, R. K. Snider, and L. Liang, "Sustained firing in auditory cortex evoked by preferred stimuli," *Nature*, vol. 435, no. 7040, pp. 341–346, May 2005.
[53] L. Qin, S. Chimoto, M. Sakai, J. Wang, and Y. Sato, "Comparison between offset and onset responses of primary auditory cortex ON-OFF neurons in awake cats," *J. Neurophysiol.*, vol. 97, no. 5, pp. 3421–3431, May 2007.
[54] A. J. Power, R. B. Reilly, and E. C. Lalor, "Comparing linear and quadratic models of the human auditory system using EEG," *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.*, vol. 2011, pp. 4171–4174, 2011.
[55] N. Ding and J. Z. Simon, "Emergence of neural encoding of auditory objects while listening to competing speakers," *Proc. Natl. Acad. Sci.*, vol. 109, no. 29, pp. 11854–11859, Jul. 2012.
[56] L. M. Miller, M. A. Escabí, and C. E. Schreiner, "Feature selectivity and interneuronal cooperation in the thalamocortical system," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 21, no. 20, pp. 8136–8144, Oct. 2001.
[57] R. C. deCharms, D. T. Blake, and M. M. Merzenich, "Optimizing sound features for cortical neurons," *Science*, vol. 280, no. 5368, pp. 1439–1443, May 1998.
[58] M. A. Escabi and C. E. Schreiner, "Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 22, no. 10, pp. 4114–4131, May 2002.
[59] Y. Bitterman, R. Mukamel, R. Malach, I. Fried, and I. Nelken, "Ultra-fine frequency tuning revealed in single neurons of human auditory cortex," *Nature*, vol. 451, no. 7175, pp. 197–201, Jan. 2008.
[60] R. L. Jenison, R. A. Reale, A. L. Armstrong, H. Oya, H. Kawasaki, and M. A. Howard, "Sparse Spectro-Temporal Receptive Fields Based on Multi-Unit and High-Gamma Responses in Human Auditory Cortex," *PloS One*, vol. 10, no. 9, p. e0137915, 2015.
[61] M. A. Escabí and H. L. Read, "Representation of spectrotemporal sound information in the ascending auditory pathway," *Biol. Cybern.*, vol. 89, no. 5, pp. 350–362, Nov. 2003.
[62] R. S. Williamson, M. B. Ahrens, J. F. Linden, and M. Sahani, "Input-Specific Gain Modulation by Local Sensory Context Shapes Cortical and Thalamic Responses to Complex Sounds," *Neuron*, vol. 91, no. 2, pp. 467–481, Jul. 2016.
[63] F. L. da Silva, "EEG: Origin and Measurement," in *EEG - fMRI*, C. Mulert and L. Lemieux, Eds. Springer Berlin Heidelberg, 2009, pp. 19–38.
[64] A. J. Noreña, B. Gourévitch, M. Pienkowski, G. Shaw, and J. J. Eggermont, "Increasing spectrotemporal sound density reveals an octave-based

organization in cat primary auditory cortex," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 28, no. 36, pp. 8885–8896, Sep. 2008.
[65] S. S.-H. Wang *et al.*, "Functional trade-offs in white matter axonal scaling," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 28, no. 15, pp. 4047–4056, Apr. 2008.
[66] S. Da Costa, W. van der Zwaag, J. P. Marques, R. S. J. Frackowiak, S. Clarke, and M. Saenz, "Human Primary Auditory Cortex Follows the Shape of Heschl's Gyrus," *J. Neurosci.*, vol. 31, no. 40, pp. 14067–14075, Oct. 2011.
[67] S. Kaur, R. Lazar, and R. Metherate, "Intracortical pathways determine breadth of subthreshold frequency receptive fields in primary auditory cortex," *J. Neurophysiol.*, vol. 91, no. 6, pp. 2551–2567, Jun. 2004.
[68] M. Galarreta and S. Hestrin, "Frequency-dependent synaptic depression and the balance of excitation and inhibition in the neocortex," *Nat. Neurosci.*, vol. 1, no. 7, pp. 587–594, Nov. 1998.
[69] J. A. Varela, S. Song, G. G. Turrigiano, and S. B. Nelson, "Differential depression at excitatory and inhibitory synapses in visual cortex," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 19, no. 11, pp. 4293–4304, Jun. 1999.
[70] J. Westö and P. J. C. May, "Capturing contextual effects in spectro-temporal receptive fields," *Hear. Res.*, vol. 339, pp. 195–210, Sep. 2016.
[71] C. Liégeois-Chauvel, A. Musolino, J. M. Badier, P. Marquis, and P. Chauvel, "Evoked potentials recorded from the auditory cortex in man: evaluation and topography of the middle latency components," *Electroencephalogr. Clin. Neurophysiol. Potentials Sect.*, vol. 92, no. 3, pp. 204–214, May 1994.
[72] T. Onitsuka, H. Ninomiya, E. Sato, T. Yamamoto, and N. Tashiro, "The effect of interstimulus intervals and between-block rests on the auditory evoked potential and magnetic field: is the auditory P50 in humans an overlapping potential?," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 111, no. 2, pp. 237–245, Feb. 2000.
[73] I. Hertrich, K. Mathiak, W. Lutzenberger, and H. Ackermann, "Differential impact of periodic and aperiodic speech-like acoustic signals on magnetic M50/M100 fields," *Neuroreport*, vol. 11, no. 18, pp. 4017–4020, Dec. 2000.
[74] M. Chait, J. Z. Simon, and D. Poeppel, "Auditory M50 and M100 responses to broadband noise: functional implications," *Neuroreport*, vol. 15, no. 16, pp. 2455–2458, Nov. 2004.
[75] D. L. Braff, M. A. Geyer, and N. R. Swerdlow, "Human studies of prepulse inhibition of startle: normal subjects, patient groups, and pharmacological studies," *Psychopharmacology (Berl.)*, vol. 156, no. 2–3, pp. 234–258, Jul. 2001.
[76] T. Grunwald *et al.*, "Neuronal substrates of sensory gating within the human brain," *Biol. Psychiatry*, vol. 53, no. 6, pp. 511–519, Mar. 2003.
[77] O. Korzyukov *et al.*, "Generators of the intracranial P50 response in auditory sensory gating," *NeuroImage*, vol. 35, no. 2, pp. 814–826, Apr. 2007.
[78] D. R. Pereira *et al.*, "Effects of inter-stimulus interval (ISI) duration on the N1 and P2 components of the auditory event-related potential," *Int. J. Psychophysiol.*, vol. 94, no. 3, pp. 311–318, 2014.
[79] S. A. Hillyard, R. F. Hink, V. L. Schwent, and T. W. Picton, "Electrical Signs of Selective Attention in the Human Brain," *Science*, vol. 182, no. 4108, pp. 177–180, Oct. 1973.

[80] V. L. Schwent, S. A. Hillyard, and R. Galambos, “Selective attention and the auditory vertex potential. I. Effects of stimulus delivery rate,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 40, no. 6, pp. 604–614, Jun. 1976.
[81] N. Fujiwara, T. Nagamine, M. Imai, T. Tanaka, and H. Shibasaki, “Role of the primary auditory cortex in auditory selective attention studied by whole-head neuromagnetometer,” *Cogn. Brain Res.*, vol. 7, no. 2, pp. 99–109, Oct. 1998.
[82] N. I. Durlach, C. R. Mason, B. G. Shinn-Cunningham, T. L. Arbogast, H. S. Colburn, and G. K. Jr, “Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity,” *J. Acoust. Soc. Am.*, vol. 114, no. 1, pp. 368–379, Jul. 2003.
[83] R. A. Lutfi, L. Gilbertson, I. Heo, A.-C. Chang, and J. Stamas, “The information-divergence hypothesis of informational masking,” *J. Acoust. Soc. Am.*, vol. 134, no. 3, pp. 2160–2170, Sep. 2013.
[84] K. Krumbholz, R. D. Patterson, A. Seither-Preisler, C. Lammertmann, and B. Lütkenhöner, “Neuromagnetic Evidence for a Pitch Processing Center in Heschl’s Gyrus,” *Cereb. Cortex*, vol. 13, no. 7, pp. 765–772, Jul. 2003.
[85] M. Chait, D. Poeppel, and J. Z. Simon, “Neural Response Correlates of Detection of Monaurally and Binaurally Created Pitches in Humans,” *Cereb. Cortex*, vol. 16, no. 6, pp. 835–848, Jun. 2006.
[86] P. F. Sowman, A. Kuusik, and B. W. Johnson, “Self-initiation and temporal cueing of monaural tones reduce the auditory N1 and P2,” *Exp. Brain Res.*, vol. 222, no. 1–2, pp. 149–157, Aug. 2012.
[87] T. Sharpee, N. C. Rust, and W. Bialek, “Analyzing neural responses to natural signals: maximally informative dimensions,” *Neural Comput.*, vol. 16, no. 2, pp. 223–250, Feb. 2004.
[88] I. M. Carruthers, R. G. Natan, and M. N. Geffen, “Encoding of ultrasonic vocalizations in the auditory cortex,” *J. Neurophysiol.*, vol. 109, no. 7, pp. 1912–1927, Apr. 2013.
[89] E. M. Zion Golumbic *et al.*, “Mechanisms underlying selective neuronal tracking of attended speech at a ‘cocktail party,’” *Neuron*, vol. 77, no. 5, pp. 980–991, Mar. 2013.
[90] N. Mesgarani and E. F. Chang, “Selective cortical representation of attended speaker in multi-talker speech perception,” *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.
[91] G. M. Di Liberto, J. A. O’Sullivan, and E. C. Lalor, “Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing,” *Curr. Biol.*, vol. 25, no. 19, pp. 2457–2465, Oct. 2015.
[92] A. Presacco, J. Z. Simon, and S. Anderson, “Effect of informational content of noise on speech representation in the aging midbrain and cortex,” *J. Neurophysiol.*, p. jn.00373.2016, Sep. 2016.
[93] J. Z. Simon, “The encoding of auditory objects in auditory cortex: Insights from magnetoencephalography,” *Int. J. Psychophysiol.*, vol. 95, no. 2, pp. 184–190, Feb. 2015.
[94] T. P. L. Roberts *et al.*, “MEG Detection of Delayed Auditory Evoked Responses in Autism Spectrum Disorders: Towards an Imaging Biomarker for Autism,” *Autism Res. Off. J. Int. Soc. Autism Res.*, vol. 3, no. 1, pp. 8–18, Feb. 2010.

[95] L. I. Zhang, S. Bao, and M. M. Merzenich, "Persistent and specific influences of early acoustic environments on primary auditory cortex," *Nat. Neurosci.*, vol. 4, no. 11, pp. 1123–1130, Nov. 2001.
[96] B. Delgutte, "Two-tone rate suppression in auditory-nerve fibers: dependence on suppressor frequency and level," *Hear. Res.*, vol. 49, no. 1–3, pp. 225–246, Nov. 1990.
[97] A. J. Noreña, B. Gourévitch, N. Aizawa, and J. J. Eggermont, "Spectrally enhanced acoustic environment disrupts frequency representation in cat auditory cortex," *Nat. Neurosci.*, vol. 9, no. 7, pp. 932–939, 2006.
[98] J. Fritz, S. Shamma, M. Elhilali, and D. Klein, "Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex," *Nat. Neurosci.*, vol. 6, no. 11, pp. 1216–1223, Nov. 2003.
[99] N. M. Weinberger, "Dynamic Regulation of Receptive Fields and Maps in the Adult Sensory Cortex," *Annu. Rev. Neurosci.*, vol. 18, no. 1, pp. 129–158, 1995.
[100] L. Ma, C. Micheyl, P. Yin, A. J. Oxenham, and S. A. Shamma, "Behavioral measures of auditory streaming in ferrets (Mustela putorius)," *J. Comp. Psychol.*, vol. 124, no. 3, pp. 317–330, 2010.
[101] R. C. Oldfield, "The assessment and analysis of handedness: the Edinburgh inventory," *Neuropsychologia*, vol. 9, no. 1, pp. 97–113, Mar. 1971.
[102] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1–2, pp. 103–138, agosto 1990.
[103] "LibriVox." [Online]. Available: https://librivox.org/the-light-princess-by-george-macdonald. [Accessed: 23-Nov-2016].
[104] H. Kado *et al.*, "Magnetoencephalogram systems developed at KIT," *IEEE Trans. Appl. Supercond.*, vol. 9, no. 2, pp. 4057–4062, Jun. 1999.
[105] A. de Cheveigné and J. Z. Simon, "Denoising based on time-shift PCA," *J. Neurosci. Methods*, vol. 165, no. 2, pp. 297–305, Sep. 2007.
[106] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Netw.*, vol. 13, no. 4–5, pp. 411–430, Jun. 2000.
[107] A. de Cheveigné and J. Z. Simon, "Sensor noise suppression," *J. Neurosci. Methods*, vol. 168, no. 1, pp. 195–202, Feb. 2008.
[108] A. de Cheveigné and J. Z. Simon, "Denoising based on spatial filtering," *J. Neurosci. Methods*, vol. 171, no. 2, pp. 331–339, Jun. 2008.
[109] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, and S. M. N. Woolley, "A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds," *PloS One*, vol. 6, no. 1, p. e16104, Jan. 2011.
[110] N. Schinkel-Bielefeld, S. V. David, S. A. Shamma, and D. A. Butts, "Inferring the role of inhibition in auditory processing of complex natural stimuli," *J. Neurophysiol.*, vol. 107, no. 12, pp. 3296–3307, Jun. 2012.
[111] E. J. Chichilnisky, "A simple white noise analysis of neuronal light responses," *Netw. Comput. Neural Syst.*, vol. 12, no. 2, pp. 199–213, Jan. 2001.
[112] T. O. Sharpee, K. D. Miller, and M. P. Stryker, "On the importance of static nonlinearity in estimating spatiotemporal neural filters with natural stimuli," *J. Neurophysiol.*, vol. 99, no. 5, pp. 2496–2509, May 2008.

[113] R. M. Warren, C. J. Obusek, and J. M. Ackroff, "Auditory Induction: Perceptual Synthesis of Absent Sounds," *Science*, vol. 176, no. 4039, pp. 1149–1151, Jun. 1972.
[114] A. S. Bregman, *Auditory scene analysis: the perceptual organization of sound*, 2. paperback ed., Repr. Cambridge, Mass.: MIT Press, 2006.
[115] A. Samuel, "Phoneme Restoration," *Lang. Cogn. Process.*, vol. 11, no. 6, pp. 647–654, diciembre 1996.
[116] J. A. Bashford and R. M. Warren, "Multiple phonemic restorations follow the rules for auditory induction," *Percept. Psychophys.*, vol. 42, no. 2, pp. 114–121, Aug. 1987.
[117] K. Friston, "A theory of cortical responses," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 360, no. 1456, pp. 815–836, Apr. 2005.
[118] A. Clark, "Whatever next? Predictive brains, situated agents, and the future of cognitive science," *Behav. Brain Sci.*, vol. 36, no. 3, pp. 181–204, Jun. 2013.
[119] R. P. Rao and D. H. Ballard, "Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects," *Nat. Neurosci.*, vol. 2, no. 1, pp. 79–87, Jan. 1999.
[120] J. Lyzenga, R. P. Carlyon, and B. C. J. Moore, "Dynamic aspects of the continuity illusion: Perception of level and of the depth, rate, and phase of modulation," *Hear. Res.*, vol. 210, no. 1–2, pp. 30–41, Dec. 2005.
[121] R. P. Carlyon, C. Micheyl, J. M. Deeks, and B. C. J. Moore, "Auditory processing of real and illusory changes in frequency modulation (FM) phase," *J. Acoust. Soc. Am.*, vol. 116, no. 6, p. 3629, 2004.
[122] B. Ross, C. Borgmann, R. Draganova, L. E. Roberts, and C. Pantev, "A high-precision magnetoencephalographic study of human auditory steady-state responses to amplitude-modulated tones," *J. Acoust. Soc. Am.*, vol. 108, no. 2, pp. 679–691, Aug. 2000.
[123] R. Schoonhoven, C. J. R. Boden, J. P. A. Verbunt, and J. C. de Munck, "A whole head MEG study of the amplitude-modulation-following response: phase coherence, group delay and dipole source analysis," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 114, no. 11, pp. 2096–2106, Nov. 2003.
[124] R. Draganova, B. Ross, A. Wollbrink, and C. Pantev, "Cortical Steady-State Responses to Central and Peripheral Auditory Beats," *Cereb. Cortex*, vol. 18, no. 5, pp. 1193–1200, May 2008.
[125] Y. Wang, N. Ding, N. Ahmar, J. Xiang, D. Poeppel, and J. Z. Simon, "Sensitivity to temporal modulation rate and spectral bandwidth in the human auditory system: MEG evidence," *J. Neurophysiol.*, vol. 107, no. 8, pp. 2033–2041, Apr. 2012.
[126] H. Luo, Y. Wang, D. Poeppel, and J. Z. Simon, "Concurrent encoding of frequency and amplitude modulation in human auditory cortex: MEG evidence," *J. Neurophysiol.*, vol. 96, no. 5, pp. 2712–2723, Nov. 2006.
[127] A. L. Giraud *et al.*, "Representation of the temporal envelope of sounds in the human brain," *J. Neurophysiol.*, vol. 84, no. 3, pp. 1588–1598, Sep. 2000.
[128] R. E. Millman, G. Prendergast, P. T. Kitterick, W. P. Woods, and G. G. R. Green, "Spatiotemporal reconstruction of the auditory steady-state response to

frequency modulation using magnetoencephalography," *NeuroImage*, vol. 49, no. 1, pp. 745–758, Jan. 2010.
[129] E. Jacewicz, R. A. Fox, C. O&#39, Neill, and J. Salmons, "Articulation rate across dialect, age, and gender," *Lang. Var. Change*, vol. 21, no. 2, pp. 233–256, Jul. 2009.
[130] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*, Repr. ed. Los Altos Hills, Calif: Peninsula Publ, 2000.
[131] R. M. Warren, "Perceptual restoration of obliterated sounds," *Psychol. Bull.*, vol. 96, no. 2, pp. 371–383, Sep. 1984.
[132] R. M. Warren, "Perceptual Restoration of Missing Speech Sounds," *Science*, vol. 167, no. 3917, pp. 392–393, 1970.
[133] A. G. Samuel, "Phonemic restoration: Insights from a new methodology.," *J. Exp. Psychol. Gen.*, vol. 110, no. 4, pp. 474–494, 1981.
[134] K. R. Kluender and R. L. Jenison, "Effects of glide slope, noise intensity, and noise duration on the extrapolation of FM glides through noise," *Percept. Psychophys.*, vol. 51, no. 3, pp. 231–238, May 1992.
[135] A. J. Shahin, C. W. Bishop, and L. M. Miller, "Neural mechanisms for illusory filling-in of degraded speech," *NeuroImage*, vol. 44, no. 3, pp. 1133–1143, Feb. 2009.
[136] A. J. Shahin, J. R. Kerlin, J. Bhat, and L. M. Miller, "Neural restoration of degraded audiovisual speech," *NeuroImage*, vol. 60, no. 1, pp. 530–538, Mar. 2012.
[137] L. Riecke, C. Micheyl, M. Vanbussel, C. S. Schreiner, D. Mendelsohn, and E. Formisano, "Recalibration of the auditory continuity illusion: sensory and decisional effects," *Hear. Res.*, vol. 277, no. 1–2, pp. 152–162, Jul. 2011.
[138] E. Vinnik, P. M. Itskov, and E. Balaban, "β- And γ-band EEG power predicts illusory auditory continuity perception," *J. Neurophysiol.*, vol. 108, no. 10, pp. 2717–2724, Nov. 2012.
[139] L. Riecke, M. Vanbussel, L. Hausfeld, D. Başkent, E. Formisano, and F. Esposito, "Hearing an Illusory Vowel in Noise: Suppression of Auditory Cortical Activity," *J. Neurosci.*, vol. 32, no. 23, pp. 8024–8034, Jun. 2012.
[140] G. M. Bidelman and C. Patro, "Auditory perceptual restoration and illusory continuity correlates in the human brainstem," *Brain Res.*, vol. 1646, pp. 84–90, Sep. 2016.
[141] B. H. Repp, "Perceptual restoration of a 'missing' speech sound: auditory induction or illusion?," *Percept. Psychophys.*, vol. 51, no. 1, pp. 14–32, Jan. 1992.
[142] J. Verschuure and M. P. Brocaar, "Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise," *Percept. Psychophys.*, vol. 33, no. 3, pp. 232–240, Mar. 1983.
[143] L. Elfner and J. L. Homick, "Some Factors Affecting the Perception of Continuity in Alternately Sounded Tone and Noise Signals," *J. Acoust. Soc. Am.*, vol. 40, no. 1, pp. 27–31, Jul. 1966.
[144] C. I. Petkov and M. L. Sutter, "Evolutionary conservation and neuronal mechanisms of auditory perceptual restoration," *Hear. Res.*, vol. 271, no. 1–2, pp. 54–65, enero 2011.

[145] C. I. Petkov, K. N. O'Connor, and M. L. Sutter, "Encoding of Illusory Continuity in Primary Auditory Cortex," *Neuron*, vol. 54, no. 1, pp. 153–165, Apr. 2007.
[146] C. I. Petkov, K. N. O'Connor, and M. L. Sutter, "Illusory Sound Perception in Macaque Monkeys," *J. Neurosci.*, vol. 23, no. 27, pp. 9155–9161, Oct. 2003.
[147] L. Riecke, A. J. van Opstal, R. Goebel, and E. Formisano, "Hearing Illusory Sounds in Noise: Sensory-Perceptual Transformations in Primary Auditory Cortex," *J. Neurosci.*, vol. 27, no. 46, pp. 12684–12689, Nov. 2007.
[148] M. K. Leonard, M. O. Baud, M. J. Sjerps, and E. F. Chang, "Perceptual restoration of masked speech in human cortex," *Nat. Commun.*, vol. 7, Dec. 2016.
[149] L. Riecke, F. Esposito, M. Bonte, and E. Formisano, "Hearing Illusory Sounds in Noise: The Timing of Sensory-Perceptual Transformations in Auditory Cortex," *Neuron*, vol. 64, no. 4, pp. 550–561, Nov. 2009.
[150] L. Riecke, A. J. van Opstal, and E. Formisano, "The auditory continuity illusion: A parametric investigation and filter model," *Percept. Psychophys.*, vol. 70, no. 1, pp. 1–12, 2008.
[151] B. Roß, T. W. Picton, and C. Pantev, "Temporal integration in the human auditory cortex as represented by the development of the steady-state magnetic field," *Hear. Res.*, vol. 165, no. 1–2, pp. 68–84, Mar. 2002.
[152] K. V. Nourski and J. F. Brugge, "Representation of temporal sound features in the human auditory cortex," *Rev. Neurosci.*, vol. 22, no. 2, pp. 187–203, Apr. 2011.
[153] C. Thyrion and J.-P. Roll, "Perceptual Integration of Illusory and Imagined Kinesthetic Images," *J. Neurosci.*, vol. 29, no. 26, pp. 8483–8492, Jul. 2009.
[154] L. Casini, P. Romaiguère, A. Ducorps, D. Schwartz, J.-L. Anton, and J.-P. Roll, "Cortical correlates of illusory hand movement perception in humans: A MEG study," *Brain Res.*, vol. 1121, no. 1, pp. 200–206, Nov. 2006.
[155] T. S. Lee and M. Nguyen, "Dynamics of subjective contour formation in the early visual cortex," *Proc. Natl. Acad. Sci.*, vol. 98, no. 4, pp. 1907–1911, Feb. 2001.
[156] M. M. Murray, G. R. Wylie, B. A. Higgins, D. C. Javitt, C. E. Schroeder, and J. J. Foxe, "The spatiotemporal dynamics of illusory contour processing: combined high-density electrical mapping, source analysis, and functional magnetic resonance imaging," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 22, no. 12, pp. 5055–5073, Jun. 2002.
[157] L. Montaser-Kouhsari, M. S. Landy, D. J. Heeger, and J. Larsson, "Orientation-selective adaptation to illusory contours in human visual cortex," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 27, no. 9, pp. 2186–2195, Feb. 2007.
[158] K. Friston, "Learning and inference in the brain," *Neural Netw.*, vol. 16, no. 9, pp. 1325–1352, Nov. 2003.
[159] I. SanMiguel, A. Widmann, A. Bendixen, N. Trujillo-Barreto, and E. Schröger, "Hearing Silences: Human Auditory Processing Relies on Preactivation of Sound-Specific Brain Activity Patterns," *J. Neurosci.*, vol. 33, no. 20, pp. 8633–8639, May 2013.

[160] S. Nozaradan, I. Peretz, M. Missal, and A. Mouraux, "Tagging the Neuronal Entrainment to Beat and Meter," *J. Neurosci.*, vol. 31, no. 28, pp. 10234–10240, Jul. 2011.
[161] I. Tal *et al.*, "Neural Entrainment to the Beat: The 'Missing-Pulse' Phenomenon," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 37, no. 26, pp. 6331–6341, Jun. 2017.
[162] B. S. Oken, M. C. Salinsky, and S. M. Elsas, "Vigilance, alertness, or sustained attention: physiological basis and measurement," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 117, no. 9, pp. 1885–1901, Sep. 2006.
[163] B. J. Farley and A. J. Noreña, "Spatiotemporal Coordination of Slow-Wave Ongoing Activity across Auditory Cortical Areas," *J. Neurosci.*, vol. 33, no. 8, pp. 3299–3310, Feb. 2013.
[164] B. S. W. Ng, T. Schroeder, and C. Kayser, "A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 32, no. 35, pp. 12268–12276, Aug. 2012.
[165] H. Luo and D. Poeppel, "Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, Jun. 2007.
[166] C. Chandrasekaran, H. K. Turesson, C. H. Brown, and A. A. Ghazanfar, "The Influence of Natural Scene Dynamics on Auditory Cortical Activity," *J. Neurosci.*, vol. 30, no. 42, pp. 13919–13931, Oct. 2010.
[167] K. Yamanaka and Y. Yamamoto, "Lateralised EEG power and phase dynamics related to motor response execution," *Clin. Neurophysiol.*, vol. 121, no. 10, pp. 1711–1718, Oct. 2010.
[168] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *J. Neurosci. Methods*, vol. 164, no. 1, pp. 177–190, agosto 2007.
[169] N. A. Macmillan and C. D. Creelman, *Detection theory: a user's guide*. Mahwah, N.J.: Lawrence Erlbaum Associates, 2005.
[170] R. Kutil, "Biased and unbiased estimation of the circular mean resultant length and its variance," *Statistics*, vol. 46, no. 4, pp. 549–561, Aug. 2012.
[171] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.
[172] V. van Wassenhove and C. E. Schroeder, "Multisensory Role of Human Auditory Cortex," in *The Human Auditory Cortex*, D. Poeppel, T. Overath, A. N. Popper, and R. R. Fay, Eds. Springer New York, 2012, pp. 295–331.
[173] M. J. Crosse, G. M. Di Liberto, and E. C. Lalor, "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.*, vol. 36, no. 38, pp. 9888–9895, Sep. 2016.
[174] E. Sohoglu, J. E. Peelle, R. P. Carlyon, and M. H. Davis, "Predictive top-down integration of prior knowledge during speech perception," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 32, no. 25, pp. 8443–8453, Jun. 2012.
[175] A. Bendixen, M. Scharinger, A. Strauß, and J. Obleser, "Prediction in the service of comprehension: Modulated early brain responses to omitted speech segments," *Cortex*, vol. 53, pp. 9–26, Apr. 2014.

[176] S. Grimm and E. Schröger, "The processing of frequency deviations within sounds: evidence for the predictive nature of the Mismatch Negativity (MMN) system," *Restor. Neurol. Neurosci.*, vol. 25, no. 3–4, pp. 241–249, 2007.
[177] A. Tavano, S. Grimm, J. Costa-Faidella, L. Slabu, E. Schröger, and C. Escera, "Spectrotemporal processing drives fast access to memory traces for spoken words," *NeuroImage*, vol. 60, no. 4, pp. 2300–2308, May 2012.
[178] V. van Wassenhove, K. W. Grant, and D. Poeppel, "Visual speech speeds up the neural processing of auditory speech," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 4, pp. 1181–1186, Jan. 2005.
[179] A.-L. Giraud *et al.*, "Representation of the Temporal Envelope of Sounds in the Human Brain," *J. Neurophysiol.*, vol. 84, no. 3, pp. 1588–1598, Sep. 2000.
[180] E. M. Zion Golumbic *et al.*, "Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a 'Cocktail Party,'" *Neuron*, vol. 77, no. 5, pp. 980–991, Mar. 2013.
[181] N. Mesgarani, "Stimulus Reconstruction from Cortical Responses," in *Encyclopedia of Computational Neuroscience*, D. Jaeger and R. Jung, Eds. Springer New York, 2014, pp. 1–3.
[182] T. Naselaris, K. N. Kay, S. Nishimoto, and J. L. Gallant, "Encoding and decoding in fMRI," *NeuroImage*, vol. 56, no. 2, pp. 400–410, May 2011.
[183] S. V. David, J. B. Fritz, and S. A. Shamma, "Task reward structure shapes rapid receptive field plasticity in auditory cortex," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 6, pp. 2144–2149, Feb. 2012.
[184] N. Ding, M. Chatterjee, and J. Z. Simon, "Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure," *NeuroImage*, Nov. 2013.
[185] M. J. Henry and J. Obleser, "Frequency modulation entrains slow neural oscillations and optimizes human listening behavior," *Proc. Natl. Acad. Sci.*, vol. 109, no. 49, pp. 20095–20100, Nov. 2012.
[186] N. M. Weinberger, "Specific long-term memory traces in primary auditory cortex," *Nat. Rev. Neurosci.*, vol. 5, no. 4, pp. 279–290, Apr. 2004.
[187] N. Cowan, "On short and long auditory stores," *Psychol. Bull.*, vol. 96, no. 2, pp. 341–370, Sep. 1984.
[188] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends Cogn. Sci.*, vol. 12, no. 5, pp. 182–186, May 2008.
[189] K. C. Backer and C. Alain, "Attention to memory: orienting attention to sound object representations," *Psychol. Res.*, vol. 78, no. 3, pp. 439–452, 2014.
[190] K. C. Backer and C. Alain, "Orienting attention to sound object representations attenuates change deafness," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 38, no. 6, pp. 1554–1566, Dec. 2012.
[191] J. F. Zimmermann, M. Moscovitch, and C. Alain, "Attending to auditory memory," *Brain Res.*, vol. 1640, Part B, pp. 208–221, Jun. 2016.
[192] J. S. Snyder, C. M. Schwiedrzik, A. D. Vitela, and L. Melloni, "How previous experience shapes perception in different sensory modalities," *Front. Hum. Neurosci.*, vol. 9, Oct. 2015.

[193] A. Kleinschmidt, C. Büchel, C. Hutton, K. J. Friston, and R. S. J. Frackowiak, "The Neural Structures Expressing Perceptual Hysteresis in Visual Letter Recognition," *Neuron*, vol. 34, no. 4, pp. 659–666, May 2002.
[194] C. M. Schwiedrzik, C. C. Ruff, A. Lazar, F. C. Leitner, W. Singer, and L. Melloni, "Untangling perceptual memory: hysteresis and adaptation map into separate cortical networks," *Cereb. Cortex N. Y. N 1991*, vol. 24, no. 5, pp. 1152–1164, May 2014.
[195] J. Pearson and J. Brascamp, "Sensory memory for ambiguous vision," *Trends Cogn. Sci.*, vol. 12, no. 9, pp. 334–341, Sep. 2008.
[196] J. Kounios, S. A. Kotz, and P. J. Holcomb, "On the locus of the semantic satiation effect: evidence from event-related brain potentials," *Mem. Cognit.*, vol. 28, no. 8, pp. 1366–1377, Dec. 2000.
[197] M. Pilotti, J. S. Antrobus, and M. Duff, "The effect of presemantic acoustic adaptation on semantic 'satiation,'" *Mem. Cognit.*, vol. 25, no. 3, pp. 305–312, May 1997.
[198] J. S. Snyder and M. K. Gregg, "Memory for sound, with an ear toward hearing in complex auditory scenes," *Atten. Percept. Psychophys.*, vol. 73, no. 7, pp. 1993–2007, Oct. 2011.
[199] T. R. Agus, S. J. Thorpe, and D. Pressnitzer, "Rapid formation of robust auditory memories: insights from noise," *Neuron*, vol. 66, no. 4, pp. 610–618, May 2010.
[200] T. Andrillon, S. Kouider, T. Agus, and D. Pressnitzer, "Perceptual Learning of Acoustic Noise Generates Memory-Evoked Potentials," *Curr. Biol.*, vol. 25, no. 21, pp. 2823–2829, Nov. 2015.
[201] R. G. Crowder, "Imagery for musical timbre," *J. Exp. Psychol. Hum. Percept. Perform.*, vol. 15, no. 3, pp. 472–478, Aug. 1989.
[202] P. B. Birkett *et al.*, "Voice familiarity engages auditory cortex," *Neuroreport*, vol. 18, no. 13, pp. 1375–1378, Aug. 2007.
[203] B. R. Buchsbaum, A. Padmanabhan, and K. F. Berman, "The Neural Substrates of Recognition Memory for Verbal Information: Spanning the Divide between Short- and Long-term Memory," *J. Cogn. Neurosci.*, vol. 23, no. 4, pp. 978–991, Apr. 2011.
[204] N. I. Durlach and L. D. Braida, "Intensity Perception. I. Preliminary Theory of Intensity Resolution," *J. Acoust. Soc. Am.*, vol. 46, no. 2B, pp. 372–383, Aug. 1969.
[205] I. Winkler and N. Cowan, "From sensory to long-term memory: evidence from auditory memory reactivation studies," *Exp. Psychol.*, vol. 52, no. 1, pp. 3–20, 2005.
[206] T. L. Hubbard, "Auditory imagery: Empirical findings," *Psychol. Bull.*, vol. 136, no. 2, pp. 302–329, Mar. 2010.
[207] M. J. Intons-Peterson, "Components of auditory imagery," in *Auditory Imagery*, D. Reisberg, Ed. Psychology Press, 2014, pp. 45–72.
[208] F. Bailes, "The prevalence and nature of imagined music in the everyday lives of music students," *Psychol. Music*, vol. 35, no. 4, pp. 555–570, Oct. 2007.
[209] M. Meyer, S. Elmer, S. Baumann, and L. Jancke, "Short-term plasticity in the auditory system: Differential neural responses to perception and imagery of

speech and music," *Restor. Neurol. Neurosci.*, vol. 25, no. 3/4, pp. 411–431, May 2007.
[210] R. J. Zatorre, A. R. Halpern, and M. Bouffard, "Mental Reversal of Imagined Melodies: A Role for the Posterior Parietal Cortex," *J. Cogn. Neurosci.*, vol. 22, no. 4, pp. 775–789, Apr. 2009.
[211] N. Bunzeck, T. Wuestenberg, K. Lutz, H.-J. Heinze, and L. Jancke, "Scanning silence: Mental imagery of complex sounds," *NeuroImage*, vol. 26, no. 4, pp. 1119–1127, Jul. 2005.
[212] B. R. Buchsbaum, R. K. Olsen, P. Koch, and K. F. Berman, "Human dorsal and ventral auditory streams subserve rehearsal-based and echoic processes during verbal working memory," *Neuron*, vol. 48, no. 4, pp. 687–697, Nov. 2005.
[213] J. Kaiser, "Dynamics of auditory working memory," *Front. Psychol.*, vol. 6, May 2015.
[214] M. E. Wheeler, S. E. Petersen, and R. L. Buckner, "Memory's echo: vivid remembering reactivates sensory-specific cortex," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 20, pp. 11125–11129, Sep. 2000.
[215] P. K. McGuire, D. A. Silbersweig, R. M. Murray, A. S. David, R. S. Frackowiak, and C. D. Frith, "Functional anatomy of inner speech and auditory verbal imagery," *Psychol. Med.*, vol. 26, no. 1, pp. 29–38, Jan. 1996.
[216] S. S. Shergill, E. T. Bullmore, M. J. Brammer, S. C. Williams, R. M. Murray, and P. K. McGuire, "A functional study of auditory verbal imagery," *Psychol. Med.*, vol. 31, no. 2, pp. 241–253, Feb. 2001.
[217] P. Janata, "Brain electrical activity evoked by mental formation of auditory expectations and images," *Brain Topogr.*, vol. 13, no. 3, pp. 169–193, 2001.
[218] P. Janata and K. Paroo, "Acuity of auditory images in pitch and time," *Percept. Psychophys.*, vol. 68, no. 5, pp. 829–844, Jul. 2006.
[219] R. Näätänen and I. Winkler, "The concept of auditory stimulus representation in cognitive neuroscience," *Psychol. Bull.*, vol. 125, no. 6, pp. 826–859, Nov. 1999.
[220] S. J. Kayser, R. A. A. Ince, J. Gross, and C. Kayser, "Irregular Speech Rate Dissociates Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha," *J. Neurosci. Off. J. Soc. Neurosci.*, vol. 35, no. 44, pp. 14691–14701, Nov. 2015.
[221] A. Pouget, J. M. Beck, W. J. Ma, and P. E. Latham, "Probabilistic brains: knowns and unknowns," *Nat. Neurosci.*, vol. 16, no. 9, pp. 1170–1178, 2013.
[222] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Phys. Rev. Lett.*, vol. 73, no. 7, p. 951, 1994.
[223] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans. Neural Netw.*, vol. 10, no. 3, pp. 626–634, May 1999.
[224] J. Cohen, *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J. : L. Erlbaum Associates, 1988.
[225] J. S. Bendat and A. G. Piersol, "The Hilbert Transform," in *Random Data: Analysis and Measurement Procedures*, 4th edition., John Wiley & Sons, Inc., 2010, pp. 473–503.