

Cognition and Behavior

# Dynamic Time-Locking Mechanism in the Cortical Representation of Spoken Words

A. Nora,<sup>1,\*</sup> A. Faisal,<sup>1,\*</sup> J. Seol,<sup>1</sup> H. Renvall,<sup>1</sup> E. Formisano,<sup>2,3</sup> and R. Salmelin<sup>1</sup><https://doi.org/10.1523/ENEURO.0475-19.2020>

<sup>1</sup>Department of Neuroscience and Biomedical Engineering, and Aalto Neuroimaging, Aalto University, Espoo FI-00076, Finland, <sup>2</sup>Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht 6200 MD, The Netherlands, and <sup>3</sup>Maastricht Center for Systems Biology (MaCSBio), Maastricht University, Maastricht 6200 MD, The Netherlands

## Abstract

Human speech has a unique capacity to carry and communicate rich meanings. However, it is not known how the highly dynamic and variable perceptual signal is mapped to existing linguistic and semantic representations. In this novel approach, we used the natural acoustic variability of sounds and mapped them to magnetoencephalography (MEG) data using physiologically-inspired machine-learning models. We aimed at determining how well the models, differing in their representation of temporal information, serve to decode and reconstruct spoken words from MEG recordings in 16 healthy volunteers. We discovered that dynamic time-locking of the cortical activation to the unfolding speech input is crucial for the encoding of the acoustic-phonetic features of speech. In contrast, time-locking was not highlighted in cortical processing of non-speech environmental sounds that conveyed the same meanings as the spoken words, including human-made sounds with temporal modulation content similar to speech. The amplitude envelope of the spoken words was particularly well reconstructed based on cortical evoked responses. Our results indicate that speech is encoded cortically with especially high temporal fidelity. This speech tracking by evoked responses may partly reflect the same underlying neural mechanism as the frequently reported entrainment of the cortical oscillations to the amplitude envelope of speech. Furthermore, the phoneme content was reflected in cortical evoked responses simultaneously with the spectrotemporal features, pointing to an instantaneous transformation of the unfolding acoustic features into linguistic representations during speech processing.

**Key words:** auditory system; magnetoencephalography; neural decoding; speech processing

## Significance Statement

It has remained unclear how speech is processed differently from other sounds with comparable meanings and spectrotemporal characteristics. In this study, computational modeling of cortical responses to spoken words highlights the relevance of temporal tracking of spectrotemporal features especially for speech. This mechanism is likely pivotal for transforming the acoustic-phonetic features into linguistic representations.

## Introduction

Humans effortlessly recognize and react to natural sounds but are especially tuned to speech. Numerous studies have attempted to localize speech-specific processing stages in the brain (Price et al., 2005; Zatorre and Gandour, 2008; Schirmer et al., 2012), but while subtle

differences in the time sequence of activation for processing speech versus other meaningful sounds have been described (Salmelin, 2007; Renvall et al., 2012), it remains unclear whether and how they relate to possible unique computations in speech processing. A major challenge

Received November 17, 2019; accepted June 1, 2020; First published June 8, 2020.

The authors declare no competing financial interests.

Author contributions: A.N., A.F., J.S., H.R., E.F., and R.S. designed research; A.N., A.F., and J.S. performed research; A.F. contributed unpublished reagents/analytic tools; A.N., A.F., and J.S. analyzed data; A.N., A.F., J.S., H.R., E.F., and R.S. wrote the paper.

has been to describe how the brain matches highly variable acoustic signal to linguistic representations, when there is no one-to-one correspondence between the two. Superior temporal cortices show sensitivity to spectrotemporal features of speech that correspond to different phonemes (Chang et al., 2010; Mesgarani et al., 2014), and to temporal structure of speech sounds but not other sounds with similar acoustic content (Overath et al., 2015). Thus, detailed tracking of temporal modulations that distinguish between phonemes may be crucial for speech processing (Poehppel, 2003; Zatorre and Gandour, 2008).

Recently, temporal locking (“entrainment”) of cortical oscillations to the amplitude envelope of speech has emerged as a potential mechanism for mapping between multiscale spectrotemporal features and linguistic structures (Ding and Simon, 2014; Ding et al., 2016; Obleser and Kayser, 2019). Low-frequency cortical oscillatory activation to speech appears to be best modeled as a combination of spectrotemporal features and phonetic-feature labels, highlighting tracking of spectrotemporal detail within speech (Di Liberto et al., 2015). However, whether temporal tracking, by continuous oscillatory activation or evoked responses as used here, is particularly relevant for processing of speech, and less crucial for other types of sounds, remains an open question. Subcortical and cortical neurons, starting from the cochlea, show phase-locked encoding of both environmental and speech sounds (Worden and Marsh, 1968; Coffey et al., 2016), and oscillatory entrainment is observed also for other sounds than speech, for example music (Doelling and Poehppel, 2015). In contrast, recent studies have suggested that spectrotemporal modulations corresponding to different phonetic features might not be best represented in a linear, time-locked manner (Chi et al., 2005; Pasley et al., 2012). Indeed, accurate reconstruction of speech can be achieved from functional magnetic resonance imaging (fMRI) recordings by modeling the encoding of the speech signal as frequency-specific spectrotemporal modulation filters, despite the loss of temporal information due to the sluggish hemodynamics and poor temporal sampling of the blood oxygen level-dependent (BOLD) response (Santoro et al., 2014, 2017).

This work was supported by the Academy of Finland Grants 255349, 256887, 292552, and 315553 (to R.S.) and 277655 (to H.R.); the Finnish Cultural Foundation (H.R.); the Sigrid Jusélius Foundation (R.S.); Maastricht University (E.F.); the Dutch Province of Limburg (E.F.); the Netherlands Organization for Scientific Research (NOW) Grant 453-12-002 (to E.F.); the Doctoral Program Brain and Mind (A.N.), the Foundation for Aalto University Science and Technology (A.N.); and the Emil Aaltonen Foundation (A.N.).

\*A.N. and A.F. contributed equally to this work.

**Acknowledgements:** We thank Mia Liljeström and Pekka Laitio for providing customized code for the source parcellation, Ossi Lehtonen for assisting with the source modeling, Jan Kujala with assistance in sound feature analysis, Tiina Lindh-Knuutila for assistance with the corpus vectors, and Sasa Kivisaari for comments on this manuscript.

Correspondence should be addressed to A. Nora at [anni.nora@aalto.fi](mailto:anni.nora@aalto.fi).

<https://doi.org/10.1523/ENEURO.0475-19.2020>

Copyright © 2020 Nora et al.

This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium provided that the original work is properly attributed.

This suggests spatially distributed representations of temporal information, similarly to frequency content, and questions the crucial role of temporal tracking of the speech signal.

The question of whether there is something special in the cortical processing of speech has often been approached through careful acoustic matching of linguistic and nonlinguistic stimuli and comparison of their activation patterns in the brain (Scott et al., 2000; Peelle et al., 2013; Overath et al., 2015). Here, we take a different approach: Our stimuli are natural spoken words and environmental sounds and, instead of delimiting their acoustic properties, we make full use of their large natural variability. We evaluate the hypothesis that temporal tracking of the spectrotemporal content is particularly important in cortical encoding of spoken words by modeling and decoding the sounds, using multiple physiologically-inspired representations that differ in temporal detail. Cortical activation is measured with magnetoencephalography (MEG), which detects cortex-wide neuronal signaling on a millisecond scale. Decoding of the time-varying features of the sounds is achieved using a convolution model (Bialek et al., 1991; Mesgarani et al., 2009; Pasley et al., 2012; Faisal et al., 2015), with a new formulation that efficiently handles the high spatiotemporal dimensionality of MEG data. The convolution model assumes that the activation of neuronal populations follows closely in time the time sequence of stimulus features. It is compared with a regression model where no such time-locking is assumed (Mitchell et al., 2008; Sudre et al., 2012; Santoro et al., 2014). The ability of each model to decode and reconstruct the sounds, using variations in the MEG signal, reveals whether the underlying assumptions of the model are an accurate description of how the human brain encodes the stimulus features. Importantly, this approach does not require a direct comparison of cortical responses to speech versus other sounds. Earlier studies that have combined machine learning-based neural decoding models (Mitchell et al., 2008) with models of cortical processing of sounds (Chi et al., 2005) have demonstrated tonotopy as well as selectivity for specific spectrotemporal modulations in the auditory cortical areas (Santoro et al., 2014), and allowed successful reconstruction of spoken word acoustics based on fMRI and intracranial recordings (Pasley et al., 2012; Santoro et al., 2017; Akbari et al., 2019). Phoneme categories or articulatory features have been shown to be reflected in EEG, fMRI, and intracranial recordings and have been decoded with quite good accuracy (Formisano et al., 2008; Chang et al., 2010; Mesgarani et al., 2014; Di Liberto et al., 2015), while semantic classification of spoken words has so far been reported only for small sets of stimuli (Simanova et al., 2010; Chan et al., 2011; Correia et al., 2015) or based on dissimilarity to preceding context (Broderick et al., 2018). The current work is the first to address acoustic, phonological, and semantic levels of analysis in the same study and to systematically evaluate the same time-varying and non-time-varying decoding models for spoken words and environmental sounds that refer to corresponding meanings (e.g., the uttered word “cat,”

and the sound of a cat “meow”). This study thus offers essential new insights into cortical mechanisms that are potentially particularly relevant for the encoding of speech, as compared with other sounds.

## Materials and Methods

### Participants

The participants were 16 (eight male, eight female) right-handed Finnish-speaking volunteers aged 19–35 years (mean 24 years). Exclusion criteria were hearing disorders and developmental or acquired neurologic or language disorders. Participants signed a written consent form before the measurement, in agreement with the prior approval of the University’s research ethics committee. The originally planned sample size was 20 participants, but data from only 16 was collected due to practical reasons. While the sample size could be considered relatively small according to current neuroimaging standards, the observed effects were robust in all individuals, with effect size ranging from moderate to large.

### Stimuli and experimental design

The speech stimuli were 44 words from various semantic categories (Extended Data Fig. 1-1 and Table 1-1). To increase acoustic variability, each word within a category was spoken by a different speaker (eight speakers in total; four females, four males; two children/adolescents). Also, each speaker spoke one word in each semantic category, to enable eliminating the influence of speaker-specific acoustic features from category-level decoding. The speaker set was rotated across participants. Stimulus duration was on average 810 ms (SD 180 ms) for the spoken words and 920 ms (SD 230 ms) for the environmental sounds. The spoken words were composed of two to five syllables, with a few compound words included. The uniqueness point (i.e., estimated time of lexical selection) was on average 500 ms (range 300–890 ms, from first to fourth syllable). Because of the transparent nature of Finnish, the uniqueness point of a spoken word (point of divergence from all other words with a different word stem) corresponds to its orthographic uniqueness point, calculated here based on the same 1.5 billion-token Finnish Internet-derived text corpus that was used to create the semantic features (Kanerva and Ginter, 2014).

The environmental sounds were high-quality sounds chosen from Internet sound libraries and, using the Adobe Audition program, modified to readily identifiable mono sounds of ~1-s duration, with sampling frequency 44.1 kHz and bit rate of 16 bits. With careful selection of sounds from several categories, we sought to include as much acoustic variability as possible to extensively span the low-level acoustic feature space. The experimental stimuli were selected from these sounds based on a behavioral test with 12 participants (who did not participate in the MEG study): participants wrote down the name of each sound immediately after recognizing it. Environmental sounds with at least 80% consistent naming and response time <3 s were chosen as stimuli. The final 44 environmental sound stimuli belong to six categories: animals, human sounds, tools, vehicles,

musical instruments and others (six to eight items per category, see Extended Data Fig. 1-1 and Table 1-1; the spoken word stimuli were the noun labels of the final selection of environmental sounds). The others category included items that did not belong to any of the other above-mentioned categories. The same environmental sounds were used as stimuli across all participants.

All sounds were filtered with an 8-kHz linear low-pass fast Fourier transform (FFT) filter (Blackman–Harris) and re-sampled at 16 kHz. Mean amplitudes of the stimuli were normalized such that the root-mean-square power of each stimulus was the same. Before the MEG measurement, the individual hearing threshold was determined for each participant, and the stimuli were delivered through plastic tubes and earpieces at 75 dB (sensation level).

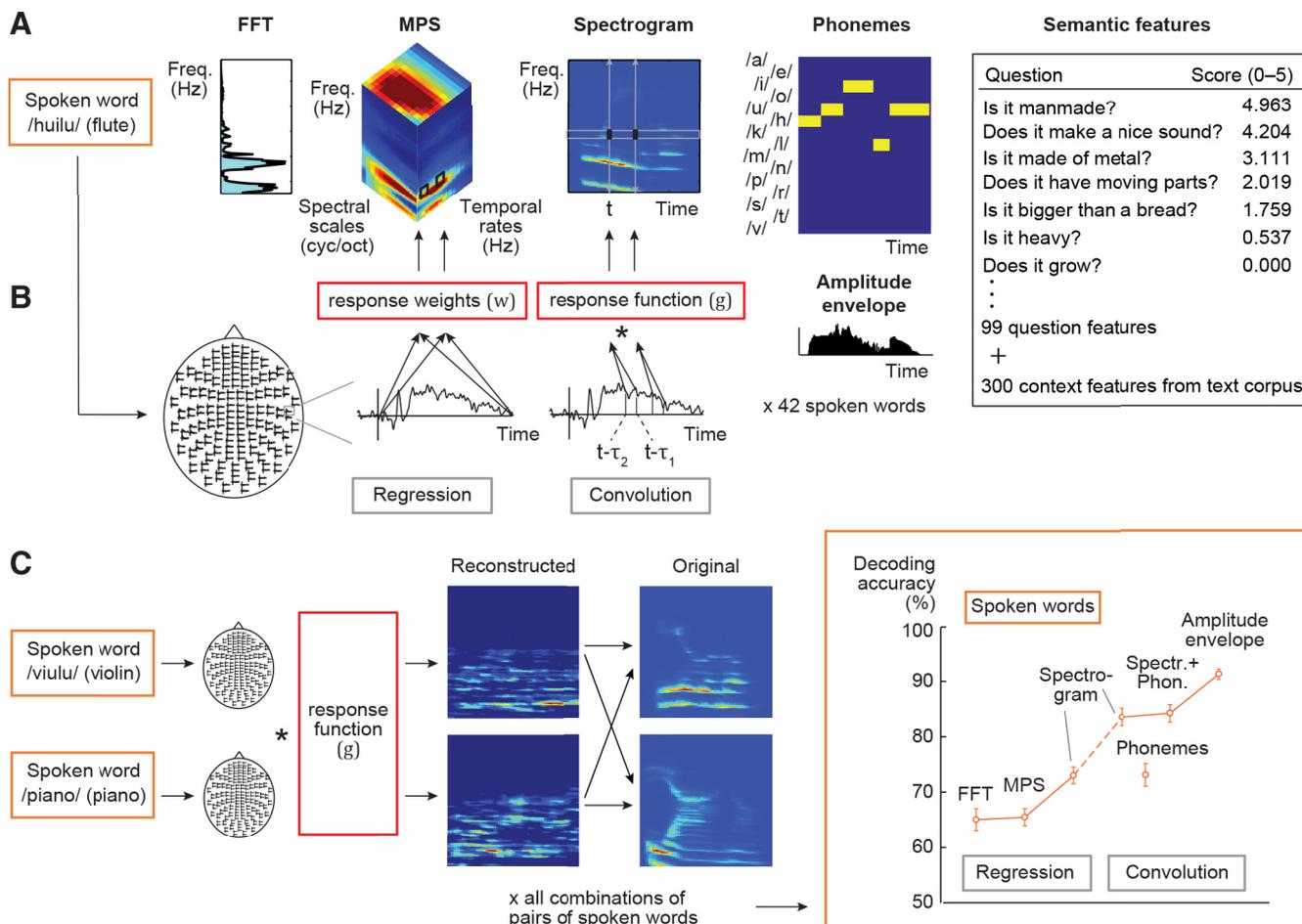
To reach a high enough MEG signal-to-noise ratio (SNR) per stimulus item for the machine learning approach, each stimulus was presented 20 times in a pseudorandom manner, such that two words spoken by the same speaker, or a spoken word and an environmental sound referring to the same meaning were not presented in a row. Event-related fields were calculated as an average of these 20 repetitions. To ensure concentration, participants performed a one-back task: They were instructed to listen carefully to each sound, think about the concept it refers to, and respond with a finger lift when two sounds with the same meaning were presented one after another (4% of trials). Response hand was alternated between participant pairs. In the task trials, another exemplar of a word or an environmental sound was presented as target stimulus (same word spoken by a different speaker or same environmental sound with a different acoustic form, e.g., a different kind of dog bark; the two stimulus types were not mixed, thus, e.g., the word cat was never followed by a cat sound).

Additional filler items, eight meaningful environmental sounds and nine spoken words, were included to make the sequence more variable. In addition, eight meaningless environmental sounds and eight pseudowords were presented 20 times each to increase participants’ attention. The meaningless environmental sounds were sounds from the abovementioned categories, which were further processed by reversing them in the time domain and, for some words, also scrambling them in 50- to 100-ms segments, to make them unidentifiable. The resulting sounds shared properties with the meaningful sounds from the same categories, but they were not identifiable in the behavioral pre-test. The pseudowords were minimal pairs to real Finnish words, following Finnish phonotactic rules. The task trials, filler sounds and meaningless sounds were excluded from the main analysis. However, we separately investigated pseudoword decoding with the different models to explore the possible contribution of lexical and semantic aspects in spoken word decoding.

### Acoustic, semantic, and phoneme features of the sounds

#### Acoustic features

We compared four sets of acoustic features that vary with respect to whether and how they incorporate information about the temporal modulations within the spoken



**Figure 1.** Different decoding models and their prediction accuracy. **A**, Stimulus features for a spoken word. The FFT model represents the frequency spectrum of the sound extracted in 128 frequency bands with logarithmically spaced center frequencies. The MPS model represents energy in four spectral scales (wide to narrow) and four temporal modulation rates (slow to fast), averaged over time. The sound spectrogram quantifies the time-evolving frequency content of the sound extracted in 128 frequency bands and 10-ms time windows by using short-term FFT. The amplitude envelope carries the temporal changes without frequency information (shown below the phoneme model). The phoneme sequence is the phoneme annotation of the word for each 10-ms time window. Semantic features were represented by scores of 99 questions (a few example questions shown) and a 300-dimensional vector trained with the co-occurrences of context words (words occurring near the stimulus words) in a large text corpus. **B**, Model estimation for the regression model (left) and the convolution model (right). A mapping between the cortical MEG responses (here illustrated on one sensor), and each stimulus feature was learned with kernel regression or kernel convolution model. As illustrated here, the regression model predicts, for example, power at each frequency-rate-scale point of the MPS by multiplying cortical responses at all (or selected) time points with unknown weights ( $w$ ). The convolution model predicts the amplitude at each time-frequency point of the spectrogram by convolving the time-sequence of cortical responses with an unknown spatiotemporal response function ( $g$ ). Specifically, values at each frequency band of a new spoken word are predicted at each time point  $t$  (moving from 0 to end of the sound) based on MEG responses in the time range from  $(t - \tau_2)$  to  $(t - \tau_1)$ , here illustrated for the lag window  $-\tau_2 = 100$  to  $-\tau_1 = 180$  ms at time points  $t$ . **C**, Model testing aimed to tell apart two left-out sounds by reconstructing the sound features (here, spectrogram) and correlating them with the original features. The procedure was repeated for all possible pairings of sounds. Predictive accuracy for the spoken words across all test sound pairs and 16 participants (mean  $\pm$  SEM) is shown on the right. The regression model was used for the decoding of non-time-varying features (FFT frequency bins/MPS rate-scale-frequency points/semantic question scores and corpus statistics), and the convolution model was used for decoding of time-varying features (spectrogram time-frequency points/phonemes at each time point/amplitude envelope); for a control analysis, the spectrogram was also decoded with the regression model. Predictive accuracy improved markedly when spoken words were modeled with the convolution spectrogram model, formalizing the concept that the neuronal population response follows closely in time the unfolding time-sequence of the sound acoustics. Even better performance was obtained when the spoken words were modeled using both the spectrogram and the phoneme sequence descriptions, and the best performance was obtained using the sounds' amplitude envelope.

words (Fig. 1A). The simplest set of features was the frequency spectrum, which is a non-time-varying representation of the stimulus power per frequency. FFT was calculated over the entire sound time window by using a filter bank of 128 frequency bands, with central frequencies of the bands ranging from 180 to 7246 Hz, uniformly distributed along a logarithmic frequency axis.

In addition to displaying an organization by frequency, the primary and secondary auditory cortices respond to different rates of temporal modulations at different spectral scales (Pasley et al., 2012; Santoro et al., 2014, 2017). These intricate multidimensional modulations within sounds are captured by our second set of features, the modulation power spectrum (MPS). The MPS [also called modulation transfer function (MTF), when describing neuronal filter properties; Elliott and Theunissen, 2009] represents energy in wide to narrow spectral scales and at slow to fast temporal modulation rates at different frequency bands (same as in FFT) for the entire duration of the sound (Chi et al., 2005). MPSs were calculated using the NSL toolbox (Chi et al., 2005) with modulation-selective filters spanning four spectral scales (0.5, 1, 2, and 4 cycles/octave) and four temporal modulation rates (1, 3, 9, and 27 Hz) within the sounds; these have been shown to capture the essential features of a broad range of natural sounds (Santoro et al., 2014). We chose a three-dimensional frequency-specific MPS where the time dimension of the MPS was averaged out, and upward-going and downward-going modulation rates were averaged together. This type of time-averaged modulation content has been shown to successfully decode environmental sounds from fMRI data (Santoro et al., 2014, 2017); computationally demanding time-varying MPS features are decodable from spatially limited intracranial recordings (Pasley et al., 2012).

For a time-evolving representation of the acoustic features, a spectrogram was generated from sounds using an auditory filter bank with 128 overlapping bandpass filter channels, with their central frequencies corresponding to the FFT frequency bands (following Pasley et al., 2012; <http://nsl.isr.umd.edu/downloads.html>). This filter bank mimics the representation of sound in the human cochlea (Chi et al., 2005). The time resolution and other parameters were optimized for retaining the characteristic temporal modulations of the sounds, while enabling extraction of the relevant spectral structure of the sounds. The sounds were divided into frames of 10 ms and integrated over 16-ms time windows.

Finally, we investigated a representation of the sounds that contains temporal modulation, but lacks fine spectral structure, by using the amplitude envelope of the sounds. The amplitude envelope was estimated through averaging the sound spectrogram across the 128 frequency bins in 10-ms windows, resulting in one feature vector.

The number of acoustic features were on average 87 in amplitude envelope (1 amplitude envelope  $\times$  87 time windows of 10 ms), 128 in FFT (128 frequency bands), 2048 in MPS (128 frequency bands  $\times$  four spectral scales  $\times$  four temporal modulation rates), and on average 11,136 in spectrogram (128 frequency bands  $\times$  87 time windows of

10 ms); for spectrogram and amplitude envelope, the length of the feature vectors varied with the sound length. To control for possible confounding effects of the varying lengths of feature vectors on the performance of the convolution model, the feature vectors for the two held-out test sounds (see below, Machine learning models and Performance evaluation, below) were always equalized to the length of the shorter one.

#### *Phoneme sequence*

Phoneme sequence of the words were obtained through their phonemic annotation, manually time-aligned to the stimulus wavefile using Praat software (Mesgarani et al., 2014; Di Liberto et al., 2015). This was a categorical representation, where each phoneme was set as 1 in those 10-ms time windows when it was present and otherwise as 0. There was no overlap, i.e., only one phoneme was marked “active” in each time window; it captures well the timing of phoneme onsets, but does not take coarticulation into account. Only phonemes with 10 or more instances in the stimulus set were included to reduce sparsity of the feature set, resulting in a set of 15 phonemes, each occurring 10–40 times. This representation had on average 1305 features.

#### *Combined spectrogram and phoneme features*

For investigating whether the spectrogram and phoneme decoding results were based on fully overlapping information, we combined these into a single feature representation of the spoken words, with on average 11,136 spectrogram features plus 1305 phoneme features; the number of features varied with sound length.

#### *Semantic features*

The semantic features were obtained by concatenating two sets of norms, one acquired through a questionnaire and the other using word co-occurrences in a large-scale text corpus. Question norms for the stimulus words were collected with a web-based survey. The questions in the survey were partly based on a previous study (Sudre et al., 2012) but modified to better represent the present selection of stimulus categories. Fifty-nine university students (32 female, 27 male, mean age 26 years; none of them participated in the present MEG study) answered 99 questions, presented in random order, about the semantic properties of each item on a scale from 0 to 5 (from definitely not to definitely yes); values for each question were averaged across the participants. Each item was thus described with a vector with 99 values.

For extracting the corpus statistics, the frequencies of co-occurrences of words in the immediate neighborhood (five words before and five words after) of each lemmatized stimulus word (compound markers added, some words changed from plural to singular) were calculated from a 1.5 billion-token Finnish Internet-derived text corpus (Kanerva and Ginter, 2014) using a continuous skip-gram Word2vec-algorithm using default parameters (Mikolov et al., 2013). The dimensionality of the trained vectors was 300, i.e., a vector with 300 values was used to describe each item.

### Similarity measures

To investigate how the different acoustic models (especially the MPS and spectrogram models) differ in the dissimilarity among the speech versus non-speech sounds, we calculated the pairwise distances (1 minus sample correlation) for all spoken words and environmental sounds (Extended Data Fig. 3-1B,D). We chose correlation as our similarity measure, as correlation between the original sound and its reconstructed model was also used in the evaluation of the classification and reconstruction performance. Thus, this measure serves as a baseline measure of discriminability of the items with the different models, and for the spectrogram model it describes the cumulative dissimilarity of all time points at each frequency band of the spectrogram. To calculate the distances (1 minus correlation), the different frequency bands were concatenated.

### Sound complexity measures

For estimating how the spectral distribution fluctuates across time in environmental sounds and spoken words we calculated, first, the variance of the spectral structure (spectral flatness measure SFM, an estimate of the number of distinct peaks in the frequency spectrum; Jayant and Noll, 1984) across each 10-ms segment within the spectrogram and, then, the spectral structure index (SSI; a measure of spectral variability defined as the SFM across time and suggested as a measure of sound complexity; Jayant and Noll, 1984). Larger values of SSI denote more variable spectral structure across time.

### MEG recording

Magnetic fields associated with neural current flow were recorded with a 306-channel whole-head neuro-magnetometer (Elekta Oy). The sensor array consists of 102 triple sensor elements, each with one magnetometer and two planar gradiometers. Planar gradiometers are most sensitive to neural activity close to the sensors and give maximum signal directly above a current source; magnetometers also pick up signals from more far away sources. The MEG signals were acquired at 1000 Hz, with a hardware bandpass filter set between 0.03 and 330 Hz. Eye movements and blink artifacts were monitored by two diagonally placed electrodes measuring electro-oculogram signal (EOG). The position of the participant's head within the MEG helmet was defined using five head position indicator coils. The locations of these coils, attached to the participant's scalp, were determined with respect to three anatomic landmarks (nasion and two preauricular reference points) with a 3D digitizer, and with respect to the sensor array by briefly feeding current to the coils during the measurement. Head movements were monitored continuously (Uutela et al., 1999). The MEG measurement lasted ~40 min, with breaks every 5 min.

### Anatomical MRI acquisition

Anatomical MRIs were obtained with a 3T MRI scanner (Magnetom Skyra, Siemens) for all 16 participants. The scan included a three-plane localizer and a T1-weighted anatomic image. To enable attribution of MEG activation

patterns to cortical loci, the MEG data were co-registered in the same coordinate system with the individual MR images.

### MEG preprocessing and source modeling

Spatiotemporal signal space separation (tSSS; Taulu and Simola, 2006) and movement compensation algorithms (Uutela et al., 1999) were applied offline to the raw data using MaxFilter software (Elekta Neuromag), to remove the effects of external interference and to compensate for head movements during the measurement. To obtain an estimate of the artifact signals caused by blinks or saccades, the MEG signals were averaged with respect to transient maxima in the EOG signal, principal component analysis (PCA) was performed on this average, and the corresponding magnetic field component was removed from the raw data (Uusitalo and Ilmoniemi, 1997).

The MEG data analysis focused on planar gradiometer channels. Single trials were averaged from 300 ms before to 2000 ms after the stimulus onset, rejecting trials contaminated by any remaining artifacts (signal strength exceeding 3000 fT/cm). On average  $19.7 \pm 0.6$  (mean  $\pm$  SD) artifact-free epochs (trials) per participant were gathered per item (maximum = 20). The averaged MEG responses were baseline-corrected to the 300-ms interval immediately preceding the stimulus onset. For visualization purposes only, the averaged data were further low-pass filtered with the Elekta Neuromag Xplotter software, with a center frequency of 40 Hz and filter width of 10 Hz.

An estimate of the underlying cortical sources was obtained with minimum norm estimates (MNEs; Hämäläinen and Ilmoniemi, 1994) using MNE Suite software package (Gramfort et al., 2014). For MNE analysis, the cortical surface of each participant was reconstructed from their individual MR images with Freesurfer software (Dale et al., 1999; Fischl et al., 1999). Each hemisphere was covered with ~5000 potential source locations. Currents oriented normal to the cortical surface were favored by weighting the transverse currents by a factor of 0.2, and depth-weighting was used to reduce the bias toward superficial sources (Lin et al., 2006). Noise-normalized MNEs [dynamical statistical parametric maps (dSPMs)] were calculated over the whole cortical area to estimate the SNR in each potential source location (Dale et al., 2000). Noise covariance matrix was estimated from the 300-ms prestimulus baseline periods across all trials. The source space (~10,000 vertices) was divided into 229 parcels of approximately equal size (101 in the left and 118 in the right hemisphere), using the Destrieux Atlas as a starting point (Destrieux et al., 2010). Medial and orbitofrontal areas were excluded, as MEG does not reliably measure activation in these areas. Before applying the parcellation, the cortical surface of each participant was morphed onto Freesurfer's average cortical surface template (fsaverage).

### Machine learning models

Each decoding model was trained and tested separately within each individual participant. Before model fitting, MEG sensor level responses were downsampled to 100 Hz (i.e., sampled at 10-ms resolution). All decoding

models were trained and tested separately for the spoken words and for environmental sounds (training and testing within spoken words or environmental sounds, not mixed together). Decoding was performed both on sensor-level and source-level. In the sensor space, the analysis was restricted to 28 planar gradiometer pairs over the bilateral auditory cortices for the acoustic and phoneme decoding, whereas signals from all 204 planar gradiometers were used for semantic decoding. For decoding of time-varying features, i.e., the spectrogram frequencies, amplitude envelope and phoneme sequences, we used a convolutional model. For decoding of non-time-varying features, i.e., semantic features, frequency spectrum (FFT), and time-averaged MPS, we used a regression model.

Before performing machine learning analysis, the stimulus features and MEG signals were standardized across all stimuli, such that the mean value was set to zero and standard deviation to unity. For the acoustic features, the FFT and spectrogram were normalized within each frequency band, and the MPS within each rate, scale, and frequency band. Each semantic feature was similarly standardized across all stimuli. The phoneme sequences were categorical variables (with values 0/1), and thus not standardized. The MEG signal power per sensor or source location was normalized within each 10-ms time window. When applied to both stimulus features and corresponding MEG responses, this procedure ensures that the absolute power (e.g., per frequency band) does not affect model estimation. Instead, the models use variation in the MEG signal power to reconstruct variation in each stimulus feature, across stimuli. Data standardization is a normal practice in statistical analysis and ensures that all the different data features are weighted equally in the analysis (Friedman et al., 2001).

In the training phase, each model learns a mapping between a stimulus feature set and the MEG data based on all but two sounds (Fig. 1A,B). The model is then evaluated on the remaining data from the same participant, i.e., used to tell apart the two held-out words/environmental sounds based on their decoded features (leave-two-out cross-validation; Fig. 1C). The training and testing steps are repeated for all possible sound pairs.

We had two machine-learning models. Time-sensitive decoding of the acoustic features was approached with lagged linear regression, which convolves the time-varying signal with a multivariate response function, and non-time-sensitive decoding with regular regression. In the following, for simplicity, we will refer to these models as convolution and regression, respectively. Both the regression and convolution models decode acoustic features from multivariate spatiotemporal neural patterns, but the regression model does not model the potential temporal tracking of stimulus features in the neural responses.

The convolution model searches for a suitable mapping (stimulus response function) between the time-varying neural response and the time-varying stimulus features (Bialek et al., 1991; Mesgarani et al., 2009; Pasley et al., 2012; Faisal et al., 2015). Here, we use a scalable formulation of the model that allows decoding of time-varying properties of the sounds ( $s$ ) by convolving the time-varying

MEG responses ( $r$ ) at brain location  $\times$  with the spatiotemporal response function  $g$ . The model decodes stimulus features  $\{\hat{s}_f$ , where  $f$  is a given stimulus feature, here the amplitude at one spectral frequency band or the value (0/1) of one phoneme] at time  $t$ , based on the MEG activation ( $r$ ) integrated from time  $(t - \tau_2)$  to  $(t - \tau_1)$ ;  $-\tau_1 > 0$ ,  $-\tau_2 \geq 0$  (Fig. 1B, where  $-\tau_1 = 180$  and  $-\tau_2 = 100$ ). Thus, the convolution model decodes spectrogram at time  $t$  using neural responses at time  $(t - \tau_2)$  and at successive 10-ms time points until time  $(t - \tau_1)$ . Thus, the lag values from stimulus to neural response are always positive (or zero), and the model incorporates the assumption that the neural activation follows (and never precedes) in time each time point of the stimulus that it is encoding. A short lag window entails that the activation of neuronal populations reflected in the MEG signal falls and rises closely following (i.e., time-locked to) the amplitude fluctuations within different frequency bands.

The procedure is repeated independently for all feature time series (e.g., for each frequency band of the stimulus spectrogram), resulting in a reconstructed time series of amplitude changes of sound spectrogram, amplitude envelope or the phoneme sequence:

$$\hat{s}_f(t) = \sum_{\times} \sum_{\tau=\tau_2}^{\tau=\tau_1} g_f(\tau, \times) r(t - \tau, \times). \quad (1)$$

To estimate the unknown spatiotemporal response functions  $g_f$ , we use the dual representation of Equation 1 and impose an L2 prior on the unknown reconstruction weights,  $g_f(\tau, \times)$ , as follows. Using matrix notation, Equation 1 can be written as  $S_f = RG_f$ , where we define  $S_f \in \mathbb{R}^{(NT) \times 1}$ ,  $G_f \in \mathbb{R}^{(\tau X) \times 1}$ , and the response matrix  $R \in \mathbb{R}^{(NT) \times (\tau X)}$ , such that each row  $r_n(t)$  in  $R$  represents MEG response to a spoken word or environment sound  $n$  across all sensors  $\times$  and all time points sampled from  $(t - \tau_2)$  to  $(t - \tau_1)$ . The unknown  $G_f$  are estimated by minimizing the regularized sum of squared error between the original  $s_f$  and the reconstructed sound features  $\hat{s}_f$ , e.g., spectrogram:

$$\arg \min_{G_f} \sum_{n,t} \{s_f(n, t) - \hat{s}_f(n, t)\}^2 + \lambda_f \sum_{\times, \tau} g_f(\tau, \times)^2. \quad (2)$$

Minimizing this loss function leads to the maximum-a-posteriori (MAP) estimate for  $G_f$ :

$$\hat{G}_f = (R^T R + \lambda_f I)^{-1} R^T S_f.$$

This classical MAP estimate is not ideal for MEG studies where the number of conditions  $N$  are typically small compared with the dimensionality of neural responses since  $R^T R \in \mathbb{R}^{(\tau X) \times (\tau X)}$  (Mesgarani et al., 2009; Faisal et al., 2015). Therefore, similar to kernel ridge regression (Bishop, 2006; also known as dual representation of ridge regression), we obtain the MAP estimate of the convolution model using its dual representation where the inner product  $R^T R$  is replaced with corresponding Gram matrix  $RR^T$ :

$$\hat{G}_f = R^T (RR^T + \lambda_f I)^{-1} S_f.$$

To estimate the regularization parameter  $\lambda_f$  we use a grid of predefined values for the hyperparameter and

choose the optimum value that minimizes leave-one-out error within training data (Bishop, 2006). Given the lag parameters  $\tau_1$  and  $\tau_2$  the learned MAP estimates  $g_f(\tau, x)$  are used to decode the acoustic and phonetic features of a new, previously unencountered, test sound as follows:

$$\hat{S}_{f\{TEST\}}(t) = \sum_x \sum_{\tau=\tau_1}^{\tau=\tau_2} g_f(\tau, x) r_{\{TEST\}}(t - \tau, x). \quad (3)$$

The convolution model thus decodes the spectrogram at time  $t$  using neural responses at time  $(t - \tau_2), \dots, (t - \tau_1)$ . To obtain an overview of the model's performance, we used a lag window of 0–420 ms (delay from time point in the spectrogram to a range of time points in the MEG signal, i.e., from  $-\tau_2$  to  $-\tau_1$ , where  $\tau_1 = -420$  and  $\tau_2 = 0$ ). Next, we advanced the lag window in non-overlapping 80-ms steps (20–100, 100–180, ..., 340–420 ms) to investigate the dependence of the decoding accuracy on the lag timing. A relatively large step size (80 ms) was chosen due to the high computational load of the analysis and also motivated by previous research showing speech tracking at latencies of ~100–180 ms (Abrams et al., 2008; Aiken and Picton, 2008; Koskinen et al., 2013). The same convolution model was additionally used for categorical phoneme features to investigate whether spectrogram decoding reflects the emergence of categorical phoneme information and whether the time lag differs for decoding of these two feature sets. Note, however, that the model is not optimized for categorical variables, which may affect the decoding performance.

The time-averaged frequency/modulation content of the sounds, estimated with FFT/MPS, respectively, was decoded using a regression model, in which each feature of the stimuli is reconstructed based on all time points (or a selected time window) of the MEG responses (Fig. 1B; Sudre et al., 2012). In the regression model, the dependent (decoded/reconstructed) variables were the semantic or non-time-varying acoustic features, while the independent variables were the MEG responses  $r(t, x)$  at brain location  $x$  and time  $t$  from stimulus onset. Unknown weights  $w_f(t, x)$  and the L2 regularization parameters were learned in a similar fashion as for the convolution model using dual representation of the regression model (Sudre et al., 2012). Using the same notation as in Equation 1, the reconstruction value ( $\hat{S}_f$ ) for one semantic or non-time-varying acoustic feature  $f$  can be written as:

$$\hat{S}_f = \sum_x \sum_t w_f(t, x) r(t, x). \quad (4)$$

We first used MEG data at 0–1000 ms from stimulus onset to obtain an overview of model performance and then analyzed sensitivity of successive 50-ms time windows (0–50, 50–100 ms, etc.) in the MEG responses to see how the decoding of semantic/acoustic features varied with time. This resolution was considered sufficient to cover the early transient response components which were expected to be especially important for encoding the acoustic features (Salmelin, 2007). In the source space, the models were learned, for each participant, separately for the vertices within each cortical parcel, and visualized by averaging the decoding performance within

each cortical parcel across participants. We visualized the sensitivity of cortical areas in decoding stimulus features by averaging the decoding performance within each cortical parcel, across all participants.

### Performance evaluation

We adopted a two-stage evaluation scheme (Palatucci et al., 2009; Sudre et al., 2012). In the first stage, the convolution or the regression model was trained to learn the unknown weights using all but two held-out items/sounds. In the second stage, the learned weights were used to decode stimulus features (acoustic, phonemic, or semantic features) for the two held-out test items. The decoding was considered correct if the combined similarity between the true features of sounds ( $s_1$  and  $s_2$ ) and the decoded features of the sound ( $p_1$  and  $p_2$ ) were greater than the reverse labeling, i.e.,

$$\text{sim}(s_1, p_1) + \text{sim}(s_2, p_2) > \text{sim}(s_1, p_2) + \text{sim}(s_2, p_1), \quad (5)$$

where  $\text{sim}(s, p)$  is the Pearson correlation between original  $s$  and decoded features  $p$ . Here, all features were considered together (e.g., all frequency bands of the spectrogram). The evaluation scheme of Equation 5 was repeated for all possible combinations in a leave-two-out cross-validation approach as follows. For item-level decoding (for all acoustic and phoneme features), the 44 items (spoken words or environmental sounds) were divided into 42 training and two test sounds in all possible pair-wise combinations, leading to a total of 946 pair-wise tests. For category-level decoding (for semantic features), the two held-out test sounds were always chosen from two different semantic categories, resulting in 66 possible pair-wise tests. The evaluation was done separately for spoken words and environmental sounds. For spoken words, these test sounds were always chosen from the same speaker (i.e., within-speaker decoding); thus category-level decoding is not influenced by speaker-related acoustic differences. The reported participant-level decoding accuracy is an average over all pair-wise combinations of the held-out pairs.

### Statistical significance

Statistical significance was established by repeating each analysis with permuted data, and empirical  $p$  values were obtained for single participants. In each permutation run, data from one participant were chosen at random, and the item labels for the averaged evoked responses were randomly permuted across the different sounds (within spoken words or environmental sounds). This procedure was repeated 200 times for each convolution model and 1000 times for each regression model. For each permutation, the models were evaluated using all possible pairwise tests in a leave-two-out cross-validation scheme. Empirical  $p$  values were computed by calculating the number of times the permutation result was better than the observed decoding accuracy for each of the 16 participants. The  $p$  values across all 16 participants were combined using Fisher's method and corrected for

multiple comparisons (over time in the reported time courses) using false discovery rate (FDR) at an  $\alpha$  level of 0.01. Two-tailed Wilcoxon signed-rank test was used for comparing the decoding performance between different models and sound classes across all 16 participants.

### Data and code availability

Ethical restrictions imposed by the University's research ethics committee prevent the authors from making brain imaging data publicly available without restrictions, as these data cannot be fully anonymized. Making the data freely available under the Creative Commons license would not enable us to restrict their use to scientific purposes. However, the relevant pseudonymized data used in this study are available from the authors on reasonable request and with permission of the University's research ethics committee, for researchers aiming to reproduce the results. The machine-learning analysis has been performed with our custom-made MATLAB scripts, which are available in <https://aaltoimaginglanguage.github.io/speechness/>.

## Results

### Acoustic decoding of spoken words relies on neural responses that closely follow time-varying spectral features

The FFT features (non-time-varying frequency spectrum) were decoded reasonably well based on the MEG responses for spoken words using the regression model (average item-level decoding accuracy 65%,  $p < 10^{-14}$ ; Fig. 1C). Use of MPS and regression model resulted in similar performance (65%,  $p < 10^{-11}$ ; FFT vs MPS, two-tailed Wilcoxon signed-rank test,  $Z = 0.10$ ,  $p = 0.94$ ). In contrast, the time-sensitive convolution model using spectrogram was remarkably successful at decoding spoken words (83%,  $p < 10^{-16}$ ), with a significant improvement in comparison to the regression model that used time-averaged MPS ( $Z = 3.5$ ,  $p = 0.000031$ ; Fig. 1C); to obtain an overview of the convolution model's performance, we used a lag window of 0–420 ms (delay from time point in the spectrogram to a range of time points in the MEG signal).

The success of the spectrogram convolution model in decoding spoken words was not merely due to increased information content (e.g., number of features) of the time-varying (spectrogram) compared with the time-averaged (FFT/MPS) feature sets: when we used a regression model to predict the spectrogram frequencies of spoken words by including all time-points of the MEG response to predict each time point in the spectrogram separately, the convolution model continued to perform significantly better (spectrogram regression 73%, spectrogram convolution 83%;  $Z = 3.3$ ,  $p = 0.00031$ ; Fig. 1C). Thus, the salient improvement of decoding accuracy with the spectrogram convolution model suggests that the neuronal population activity in the auditory cortices closely follows the speech signal in time, to accurately encode the minute changes within spoken words.

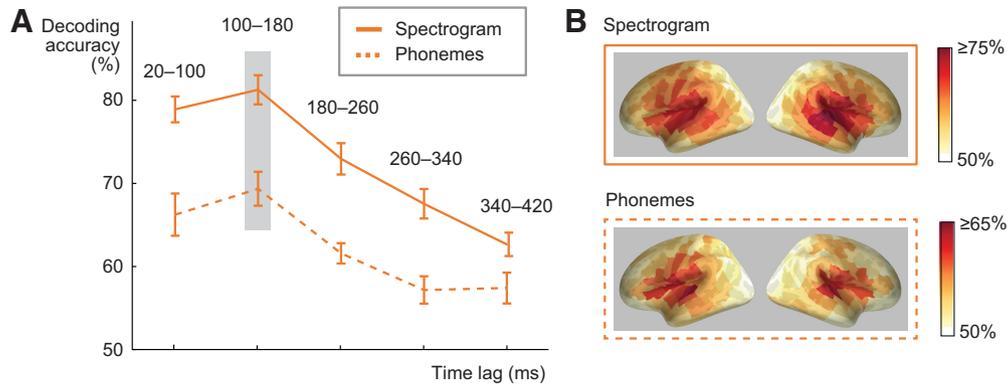
We then investigated whether changes in the amplitude envelope of the spoken words, corresponding to the slow temporal modulations within the speech rhythm, are important for the particularly successful decoding of spoken words, by separately decoding the amplitude envelope (spectrogram averaged across frequency) with the convolution model. The results reveal remarkably high classification performance for the spoken words (91%, significant difference from spectrogram decoding for speech,  $Z = 3.5$ ,  $p < 0.001$ ; Fig. 1C) and further highlight the importance of the temporal aspects of the stimulus for speech decoding.

### Time-locked encoding of spoken words reflects acoustic-to-phoneme mapping

Leave-one-out-reconstruction of the spoken word waveforms with the spectrogram convolution model suggested preservation of acoustic properties characteristic of different phonemes (Audio 1; Extended Data Fig. 2-1). To test this hypothesis, phonemic annotation of the speech sounds was aligned to the stimulus time course (Mesgarani et al., 2014; Di Liberto et al., 2015), and the same convolution model was used for decoding these categorical phoneme features; please note that this model is not optimized to deal with categorical variables, which may affect model performance. These phoneme sequences (Fig. 1A) were decoded successfully with the convolution model (73%;  $p < 10^{-16}$ ; Fig. 1C). Furthermore, a representation of the spoken words that combined both the speech spectrogram and the sequence of phonemes performed even above the spectrogram alone (84%;  $Z = 2.5$ ,  $p = 0.01$ ).

To determine at what delay after each time point in a sound the brain responses reflect encoding of its acoustic and phoneme information, we investigated different lag windows between time points of the stimulus and the MEG response. For this analysis, we chose the spectrogram model, which allows for a reconstruction of the spoken words that retains the relevant sound features (in contrast to the amplitude envelope alone), and additionally analyzed whether phoneme information is represented at a similar lag. The lag window was advanced in non-overlapping 80-ms steps (20–100, 100–180, and so on until 340–420 ms). Both spectrogram and phoneme decoding with the convolution model performed best when MEG responses at a lagged window of 100–180 ms after each time point in the sound were used (81% at 100–180 ms vs 73% at 180–260 ms lag  $Z = 3.5$ ,  $p = 0.000031$ ; for phonemes 69% at 100–180 ms vs 62% at 180–260 ms lag  $Z = 2.4$ ,  $p = 0.016$ ; Fig. 2A). For a sanity check, we also calculated the decoding accuracy for speech spectrogram using a counterintuitive lagged window of –80–0 ms, i.e., evaluating if “past” neural response could predict “future” spectrogram. With this lag, decoding performance was at chance level (mean decoding accuracy across 16 subjects 55%, SEM = 1.85). The cortical areas contributing to successful performance of the acoustic and phoneme models concentrated in and around the left and right auditory cortices (Fig. 2B).

To explore the contribution of semantics in the acoustic-phonetic decoding of spoken words, we investigated



**Figure 2** Influence of convolution lag and cortical sources for spectrogram and phoneme decoding. **A**, Spectrogram and phoneme decoding accuracy at different lags between a time point in the stimulus and a time window in the MEG response (average across all 16 participants  $\pm$  SEM). Note that the lag window does not correspond to timing relative to the stimulus onset in the MEG evoked response. The best predictive accuracy for decoding spectrogram and phoneme sequence of spoken words was reached with a lag of 100–180 ms (significant difference to 180- to 260-ms lag,  $p=0.000031$  for spectrogram,  $p=0.016$  for phonemes). **B**, Cortical sources contributing to decoding of acoustic and phoneme features in spoken words with the convolution model at 100- to 180-ms lag (this time window showed best performance). Color scale denotes average decoding accuracy ( $>50\%$ ) across all 16 participants.

how these different models decoded pseudowords, by training the models with all the meaningful spoken words, and using all possible pairs of the eight pseudowords (28 combinations) for testing. Pseudoword decoding showed similar performance as decoding of the meaningful words (FFT 68%, MPS 66%, spectrogram convolution 86%, phoneme convolution 75%), with significant improvement for the spectrogram convolution compared with the MPS ( $Z=3.4$ ,  $p=0.00015$ ). The convolution model was able to decode the phoneme content of the pseudowords with a similar accuracy (75%) as for the real words. Thus, time-locked encoding does not seem to depend on lexical content, giving further support to the idea that it might reflect mapping of acoustic content to prelexical linguistic units.

### Decoding of environmental sound acoustics does not benefit from time-locked models

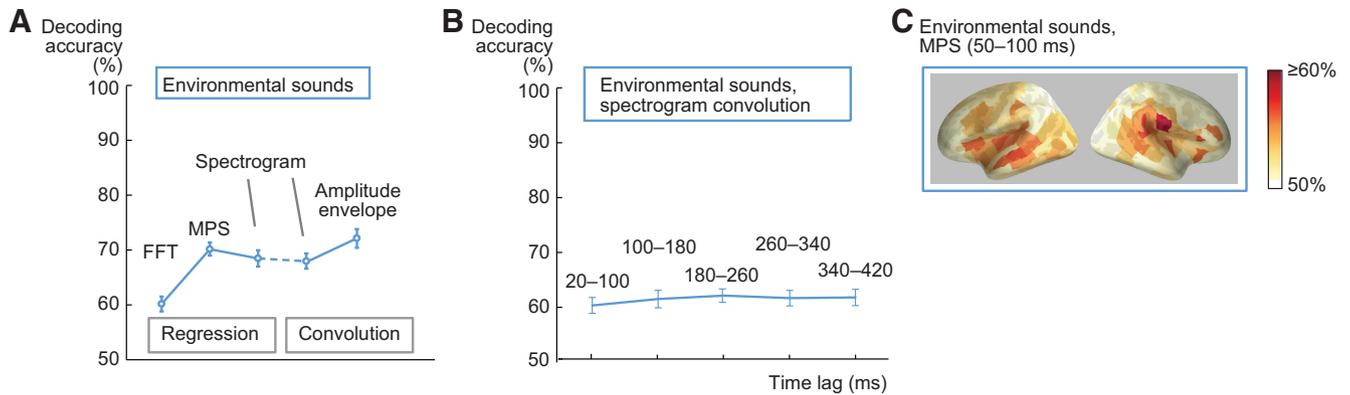
To investigate whether the convolution model performs better than time-averaged models for other sounds than speech, we separately applied the different acoustic models to a variety of environmental sounds (Extended Data Fig. 1-1 and Table 1-1). They conveyed the same meanings as the spoken words; thus, the conceptual endpoint of neural processing is presumed to be the same. Also, their cortical responses were similar to spoken words (Extended Data Fig. 1-2). However, the processing steps for accessing the meaning of environmental sounds are presumably different from for spoken words.

The FFT features were decoded based on the MEG responses for environmental sounds at a significant level with the regression model (60%,  $p < 0.0001$ ; Fig. 3A). Use of MPS together with regression model resulted in better performance than FFT (70%,  $p < 10^{-15}$ ; FFT vs MPS,  $Z=3.5$ ,  $p=0.000031$ ); such an improvement in decoding was not observed for spoken words (for spoken words FFT and MPS decoding were both at 65%). In contrast, the accuracy with spectrogram convolution (68%,

$p < 10^{-4}$ ) did not improve compared with the MPS result ( $Z=1.1$ ,  $p=0.27$ ), as was observed for words (spectrogram convolution decoding for words was at 83%, with significant difference between spoken words and environmental sounds:  $Z=3.5$ ,  $p < 0.0001$ ). Using a regression model to decode the spectrogram frequencies of environmental sounds resulted in similar decoding performance (68%,  $Z=0.18$ ,  $p=0.87$ ) and was also fairly similar to spoken word decoding (spectrogram regression 73%; no significant difference between spoken words and environmental sounds:  $Z=1.9$ ,  $p=0.058$ ). No time dependence was observed for environmental sounds using different lag windows (61% at 100–180 ms vs 62% at 180–260 ms lag;  $Z=0.052$ ,  $p=0.98$ ; Fig. 3B). The amplitude envelope decoding showed some improvement for the environmental sounds compared with the spectrogram convolution decoding ( $Z=2.3$ ,  $p=0.021$ ), but was still low (72%) compared with the remarkably high decoding performance for speech amplitude envelope (91%; significant difference between spoken words and environmental sounds:  $Z=3.5$ ,  $p < 0.0001$ ). Similarly to the convolution spectrogram decoding for speech, the cortical areas contributing to decoding environmental sound acoustics mainly concentrated in and around the left and right auditory cortices (Fig. 3C).

### Enhanced performance with the time-locked model for spoken words, but not environmental sounds, is not explained by possible confounds in the comparison of models

In visual inspection, all sounds showed typical auditory evoked responses at the sensor level, at good SNR (Extended Data Fig. 1-2A), with activation around the auditory cortices in the left and right temporal lobes (Extended Data Fig. 1-2B). Subtle differences in the averaged evoked responses between spoken words and environmental sounds were not investigated; only the



**Figure 3.** Comparison of different acoustic models for environmental sound decoding. **A**, Predictive accuracy for environmental sounds (blue) across all test sound pairs and 16 participants (mean  $\pm$  SEM). The regression model was used for the decoding of non-time-varying features (FFT frequency bins/MPS rate-scale-frequency points/semantic features), and the convolution model was used for decoding of time-varying features (spectrogram time-frequency points/amplitude envelope); for a control analysis, the spectrogram was also decoded with the regression model. **B**, Investigation of convolution lag on spectrogram decoding accuracy for environmental sounds (average across all 16 participants  $\pm$  SEM). **C**, Cortical sources contributing to decoding of acoustic features in environmental sounds with the MPS regression model at 50–100 ms after stimulus onset (performance was best with this model and time window). Color scale denotes average decoding accuracy (>50%) across all 16 participants.

variability among items (within spoken words or environmental sounds) affects the machine learning modeling.

However, as the sounds were a selection of natural sounds, there were several differences between the spoken words versus environmental sounds (Extended Data Figs. 3-1A, 3-2). Thus, we sought to rule out potential confounds arising from these differences that could be thought to influence the observed striking improvement of decoding performance with the convolution model for speech but lack of similar improvement for environmental sounds.

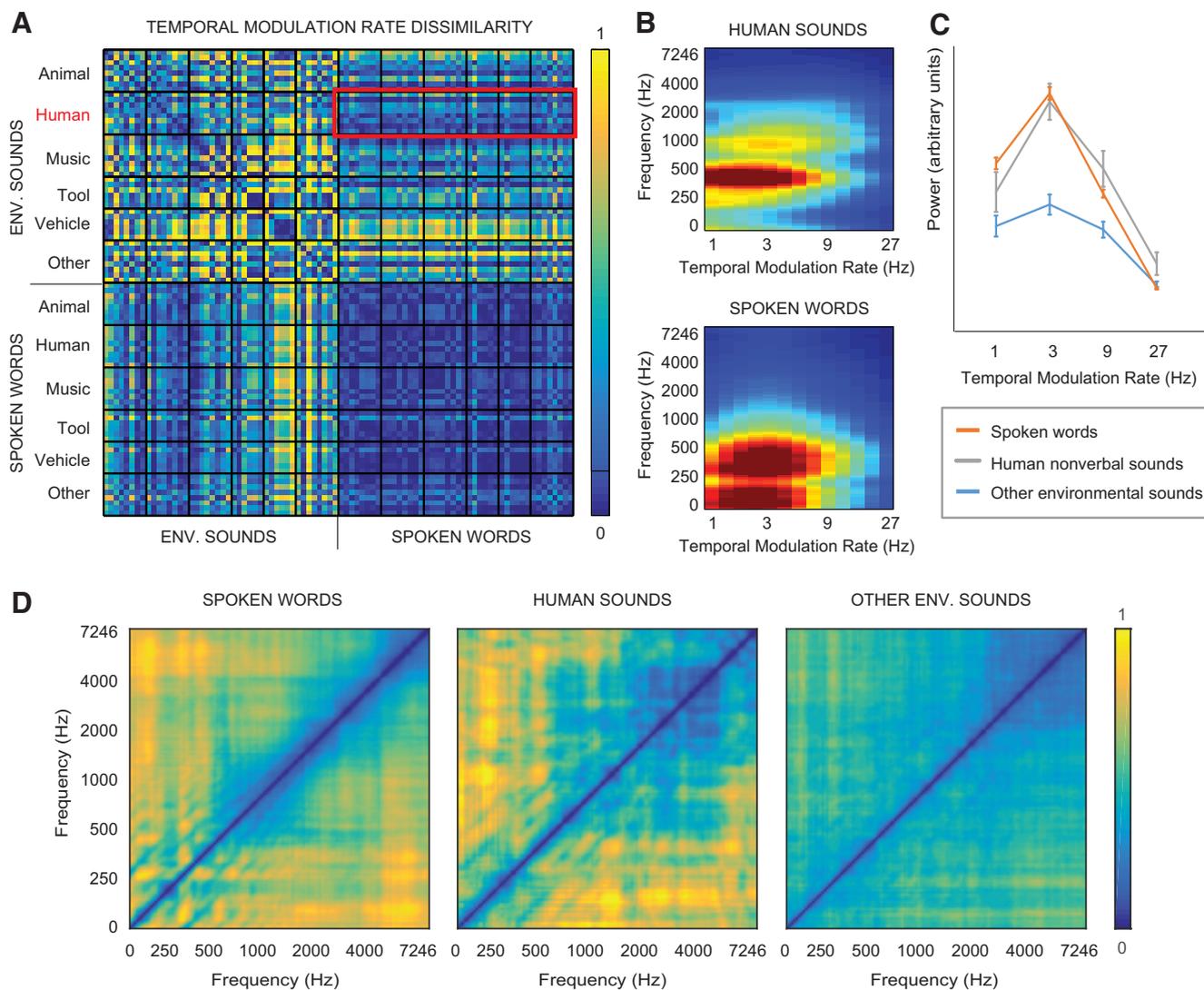
The benefit for spoken word but not environmental sound decoding was not due to larger dissimilarity among the spoken word stimuli with added temporal detail: while the spectrograms of spoken words do show larger dissimilarity than the MPSs (two-tailed  $t$  test:  $t_{(945)} = 139.5$ ,  $p < 0.001$ ), this is also true for the environmental sounds ( $t_{(945)} = 14.8$ ,  $p < 0.001$ ; Extended Data Fig. 3-1B). Notably, the dissimilarity among items, which typically leads to better decoding performance, is overall significantly greater for the environmental than speech stimuli, also with the spectrogram model ( $t_{(945)} = 22.5$ ,  $p < 0.001$ ). This confirms that increased stimulus dissimilarity alone cannot explain the significantly better classification performance for spoken words than environmental sounds.

We also verified that the difference for time-locked decoding between spoken words and environmental sounds was present already at the onset of the stimulus: The spectrograms were cut to 0–100 ms in length and decoded with the convolution model using a lag window of 100–180 ms. This resulted in an average decoding accuracy of 69% for spoken words and 58% for environmental sounds (significantly better for spoken words;  $Z = 3.16$ ,  $p = 0.002$ ). Thus, although the decoding performance is overall worse with only the start of the word/sounds, as we expected, a significant difference remains between

the spoken words and the environmental sounds (for the whole duration of the words and the same lag the decoding accuracies were 81% for speech and 61% for non-speech). This additional analysis shows that the enhanced decoding of words with the time-locked model is not explained by longer analysis time that would be needed to extract word meanings; instead, the time-locked model applies from the start of the word.

Furthermore, the benefits of the convolution model for modeling the cortical encoding of spoken words are not restricted to the leave-two-out classification task: direct leave-one-out reconstruction (correlation of the original and reconstructed features) demonstrated that spoken words were better reconstructed with the spectrogram convolution than MPS regression model (0.19 vs 0.10,  $Z = 3.5$ ,  $p < 0.001$ ; Extended Data Fig. 3-1C; for examples of reconstructed sounds, see Extended Data Fig. 2-1, Audio 1), whereas environmental sounds were better reconstructed with the MPS regression than spectrogram convolution (0.14 vs 0.08,  $Z = 3.5$ ,  $p = 0.001$ ; Extended Data Fig. 3-1C; for examples of reconstructions based on the spectrogram convolution model, see Extended Data Fig. 2-1, Audio 2). The amplitude envelope of spoken words was also remarkably well reconstructed in comparison to the amplitude envelope of environmental sounds (reconstruction accuracy for spoken words 0.67 and for environmental sounds 0.32;  $Z = 3.5$ ,  $p < 0.001$ ; Extended Data Fig. 3-1C). Moreover, correlating the pair-wise reconstructed spectrogram distances with the distances between original sound spectrograms shows that relatively more information is preserved in the reconstructions of spoken words compared with environmental sounds based on their respective cortical responses (Spearman correlation  $r = 0.41$ ,  $p < 0.001$  for spoken words,  $r = 0.22$ ,  $p < 0.001$  for environmental sounds; Extended Data Fig. 3-1D).

The improved performance of the time-locked encoding model for spoken words was also unlikely to be solely due



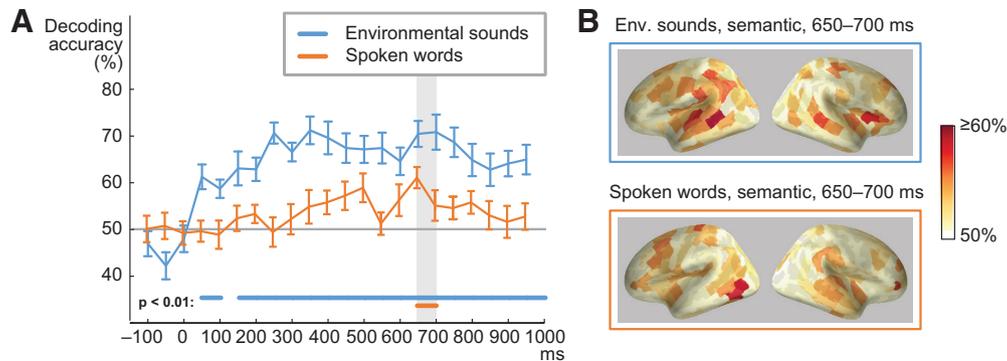
**Figure 4.** Comparison of human nonverbal sounds and spoken words. **A**, Dissimilarity (1 minus correlation) of the MPS temporal rates across the environmental sounds and spoken words. **B**, Average temporal modulation rate  $\times$  frequency representations for human sounds (top) and spoken words (bottom). **C**, Power (mean  $\pm$  SEM) at different temporal modulation rates for spoken words (red), human nonverbal sounds (gray) and other categories of environmental sounds (blue). **D**, Dissimilarity of different frequency bands of the sounds, indicating the degree of correlated temporal modulations (co-modulations), calculated separately for each item, and averaged over spoken words (left), human nonverbal sounds (middle), and other environmental sounds (right).

to different kinds of temporal properties of the words and other sounds. Specifically, speech has prominent slow (1–7 Hz) temporal modulations (Fig. 4A,B; Extended Data Figs. 3-1A, 3-2) that are important for its intelligibility (Elliott and Theunissen, 2009) and have been suggested to be represented in the auditory cortices through a linear coding scheme (Pasley et al., 2012). Environmental sounds produced by the human vocal tract (laughter, crying, etc.) are very similar to speech in terms of spectrotemporal characteristics and temporal modulation rates (with a small difference observed only for the highest rate ( $t_{(50)} = 4.5$ ,  $p = 0.004$ ; Bonferroni corrected for multiple comparisons; Fig. 4C). Spoken words showed a more variable spectral structure across time (larger values of SSI) than environmental sounds ( $t_{(86)} = 3.9$ ,  $p < 0.001$ ). However, the SSI did not differ between spoken words and human nonverbal sounds ( $t_{(50)} = 0.45$ ,  $p = 0.65$ ), which

were similar to each other also in their co-modulation properties (correlated temporal modulations between frequency channels; Fig. 4D). Yet, we saw no improvement in decoding of this subset of environmental sounds with the convolution model (MPS regression 68% vs spectrogram convolution 69%;  $Z = 0.31$ ,  $p = 0.84$ ). The results indicate that particularly spoken words, with their specific combination of spectrotemporal features, show remarkably improved decoding when neuronal population activity in the auditory cortices is modeled as tightly following the time-evolving acoustics.

#### Semantic features are successfully decoded for both spoken words and environmental sounds, but with different timing

To ensure that the sounds were processed up to their meaning, the participants' task was to identify immediate



**Figure 5.** Comparison of semantic decoding for spoken words and environmental sounds. **A**, Time course of regression-model based predictive accuracy in semantic features of environmental sounds and spoken words (16 participants, mean  $\pm$  SEM). The values on the x-axis represent the starting time of the 50-ms time windows in the MEG responses. Time 0 indicates the onset of the sound stimulus. Decoding of the semantic feature set was performed for successive 50-ms time windows for the whole stimulus duration, for both stimulus types. The time windows with statistically significant decoding ( $p < 0.01$ ) are indicated by thick horizontal lines above the x-axis. Gray solid line denotes chance level performance (50%). The gray bar represents the time window used for source level decoding. **B**, Source areas contributing to decoding of semantic features for environmental sounds and spoken words at 650–700 ms from stimulus onset. This time window had significant decoding performance for both classes of sounds.

repetitions of the same meaning, i.e., two different exemplars of a sound presented one after the other. The task was easier for the spoken words than for the environmental sounds (average percentage of hits for spoken words 94%, for environmental sounds 78%; two-tailed Wilcoxon signed-rank test,  $Z = 3.6$ ,  $p = 0.00003$ ). Reaction times (RTs) were also longer for the environmental sounds (average RT for hits 1.7 s) than the spoken words (average RT for hits 1.5 s; two-tailed Wilcoxon signed-rank test,  $Z = 3.5$ ,  $p = 0.001$ ). The same set of semantic features (Fig. 1A, right) was used for modeling the environmental sounds and spoken words. The semantic feature representations captured meanings of individual items and formed salient clusters of the five semantic stimulus categories. We focused on telling apart two sounds from two different categories based on their reconstructed features (combination of question features and text corpus features). The regression model was successful in decoding the semantic category of both environmental sounds (81%;  $p < 10^{-15}$ ) and spoken words (58%;  $p = 0.0015$ ), with significantly better accuracy for environmental sounds ( $Z = 3.5$ ;  $p = 0.000031$ ).

Semantic decoding reached significance for environmental sounds at 50–100 ms and remained significant ( $p < 0.01$ ) from 150 ms on until the end of the analysis window (Fig. 5A); we analyzed sensitivity of all successive 50-ms time windows in the MEG responses. In the same analysis for spoken words, best performance for decoding spoken word semantics was reached late and was significant only at 650–700 ms (Fig. 5A). In contrast, MPS-based acoustic decoding of environmental sounds was significant at 50–150 and 200–300 ms after stimulus onset for environmental sounds and at 250–300 ms for spoken words. Cortical sources contributing to semantic decoding are illustrated in Figure 5B (see also Extended Data Fig. 5-1).

## Discussion

### Cortical activation faithfully tracks the spectrotemporal detail of spoken words

The cortical processing of speech as compared with the processing of other sounds has remained a major open question in human neuroscience. Here, by combining time-sensitive brain imaging and advanced computational modeling, we discovered that the acoustic-phonetic content of natural, meaningful spoken words is encoded in a special manner, where the cortical evoked responses faithfully track in time the unfolding spectrotemporal structure and the amplitude envelope of the spoken words. This time-locked encoding was observed also for meaningless pseudowords. However, responses to environmental sounds, even human-made non-speech sounds with spectral and temporal modulations comparable to speech, did not show improved decoding with the dynamic time-locked mechanism, and were better reconstructed using the time-averaged spectral and temporal modulation content, suggesting that a time-averaged analysis is sufficient to reach their meanings.

Phase-locking of oscillatory activation to the amplitude envelope of sounds has recently gained wide interest as a potential mechanism for cortical encoding of speech (Ding and Simon, 2014; Ding et al., 2016). However, it has remained controversial whether entrainment is directly related to speech parsing: it might reflect several different underlying processes (Ding and Simon, 2014), some of which are not specific to speech nor to the human auditory system (Steinschneider et al., 2013). Based on the current results, cortical auditory evoked activation tracks fine-scale spectral detail of single spoken words in a temporally accurate manner, relying heavily on both the spectrotemporal and overall amplitude changes within the sound. We propose that this time-locking of evoked responses to an unfolding spoken word may partly reflect

the same phenomenon as the reported entrainment of oscillatory activity to continuous speech envelope. It has been suggested that phase-locking of oscillatory activation to the stimulus amplitude envelope may be driven by acoustic edges at syllable onsets (Doelling et al., 2014), and it can be modeled as a series of transient responses to changes in the stimulus (Aiken and Picton, 2008). Furthermore, the spectral content and temporal modulations within speech are non-independent, and envelope entrainment also depends on fine spectrotemporal detail (Ding et al., 2014). Thus, phase-locked oscillations can at least partly be thought of as reflecting the superposition of transient evoked responses that track the fine-scale spectrotemporal evolution of the speech stimulus (Alexandrou et al., 2018; Obleser and Kayser, 2019); this does not exclude the possible involvement of endogenous oscillatory activity. Here, for evoked responses to isolated spoken words the best decoding performance was obtained for the amplitude envelope of the sound (reconstructed with 67% accuracy), indicating that the cortical responses carry a lot of the characteristics of the envelope. However, the particularly high performance for the amplitude envelope compared with the spectrogram may be partly due to the reduced dimensionality of the decoded features (1 amplitude envelope vs 128 frequency bands, respectively), reducing the influence of noise. Also, the different acoustic features within speech are temporally coupled, and the on-line analysis of features in the spectrotemporal fine structure is modulated by cues in the envelope (Shamma et al., 2011). As an average of all the frequency channels, the amplitude envelope may carry the most prominent temporal events in the spoken words, and its decoding may thus highlight the most salient speech-tracking features of the MEG signal.

Speech comprehension has been shown to rely strongly on slow-rate temporal modulations of the speech envelope (Elliott and Theunissen, 2009). Notably, in our study, human vocalizations, the environmental sounds most similar to speech in that they contain prominent slow temporal modulations as well as similar co-modulations across frequencies, did not show any improvement in decoding with the time-sensitive spectrogram model. The dynamic mode of encoding uncovered here thus seems to take place for the particular combination of spectrotemporal features that are characteristic of speech. The high performance (83%) of the spectrogram convolution model, in which each time-frequency-point was reconstructed based on the MEG signal, showed that the fine spectral content of speech is tracked alongside the amplitude envelope at a high temporal resolution, and the spoken word spectrograms can be fairly accurately reconstructed based on this model. Furthermore, the difference in decoding performance is visible right from the onset of the sound, for the first 100 ms, when the sound is first recognized as speech. Our results, along with recent findings based on decoding sounds from fMRI responses (Santoro et al., 2017), indicate that human auditory cortical processing might be optimally tuned for encoding the spectrotemporal structure of speech.

The current results corroborate previous work highlighting the importance of temporal modulations within

speech (Overath et al., 2015), but they should not be interpreted to mean that auditory cortical activation would not be time-locked to non-speech sounds. Non-speech sounds also elicit prominent cortical auditory evoked responses time-locked to the acoustic changes within the sound, and phase-locking of cortical oscillations to the amplitude envelope is observed also for non-speech sounds (Doelling and Poeppel, 2015). The current results further show that the amplitude envelope of environmental sounds is tracked in cortical evoked responses, as it can be reconstructed at 32% accuracy with the convolution model. Thus, a part of the time-locking phenomenon seems not to be specific to speech. However, overall, the best reconstruction of the environmental sounds (14%) was achieved by modeling their time-averaged modulation content (MPS), suggesting that the time-averaged spectral and temporal modulation content may be the information most accurately represented in cortical activation and sufficient for accessing meanings of non-speech sounds. Indeed, the identification of common sounds in the environment may be based on the encoded summary statistics (McDermott et al., 2013; Santoro et al., 2014). Interestingly, the overall modulation content was a more accurate model of environmental sounds encoding than the frequency spectrum alone, whereas for speech there was no such improvement with modulation filters, consistent with a recent fMRI study (de Heer et al., 2017).

For mapping speech acoustics to linguistic representations, accurate dynamic encoding of the temporal evolution within each frequency band at each time point is needed. An interesting parallel can be found from animal studies, where single-unit recordings in the auditory cortex suggest frequency-bin-selective synchronization of neuronal population discharges to the temporal envelope of species-specific calls, which in many ways resemble human speech; thus, temporal envelope information within different frequency bands in behaviorally relevant vocalizations might be encoded cortically by coherent discharge patterns in distributed neuronal populations (Gehr et al., 2000; Nagarajan et al., 2002; Gourévitch and Eggermont, 2007). The current results suggest that, in humans, similar encoding mechanisms might have become especially important for speech, with extensive exposure to the language environment during development. Future studies should determine if similar time-locking might be observed with tasks requiring attention to fine-grained temporal detail or specialization to categorical perception in other sounds through experience, e.g., for instrumental sounds in musicians.

### **Accurate temporal encoding of the acoustic stream is necessary for lexical and semantic access of spoken words**

Despite the different acoustic content of spoken words and corresponding environmental sounds, their processing converged at the endpoint; the same set of semantic features was successful in decoding both classes of sounds. Spoken words showed relatively low semantic decoding performance, in line with previous studies (Simanova et al., 2010; Correia et al., 2015). One reason

for the different semantic decoding performance of spoken words and environmental sounds may have been the fact that environmental sounds are inherently less familiar and their semantic access may thus require a more active effort and longer time than for spoken words. This conclusion is also suggested by the performance in the one-back task (detecting two subsequent sounds with the same meaning), which was more effortful, with more errors and longer reaction times, for environmental sounds than spoken words. Nonetheless, our set of environmental sounds was tested for fast and consistent naming. Together, results of the behavioral task and the semantic decoding results indicate that attentional engagement was equal or even stronger for the environmental sounds than spoken words, and in fact more effort and time was needed to decipher the meanings of the environmental sounds. Thus, increased attention or longer identification times cannot explain the strikingly more successful acoustic decoding for spoken words than environmental sounds. Furthermore, the differential decoding performance was visible already during the first 100 ms of the sounds.

Semantic decoding of environmental sounds first reached significance at 50–100 ms, at the same time as acoustic decoding, and continued to be successful until the end of the analysis period, whereas semantic decoding of spoken words reached significance at 650–700 ms (this one significant time bin likely represents only a part of a wider time range). The current results are complementary to traditional experimental paradigms of semantic processing, where the violation of semantic expectation for spoken words and sounds shows in cortical responses at around 200–700 ms (Aramaki et al., 2010; Frey et al., 2014; Hendrickson et al., 2015). Comparable latencies (200–700 ms) were also recently observed in decoding the similarity of spoken word meanings based on preceding content in continuous speech, using EEG (Broderick et al., 2018). The current decoding results indicate that for environmental sounds the acoustic features may be informative of the physical sources (i.e., meaning) of the sound early on, but that the identification may be extended in time or its timing and duration may be variable between sounds, whereas for spoken words identification seems to occur within a more distinct and narrow time window, when the correct word candidate has been identified in the lexicon. The identification of spoken words involves the comparison of the incoming input signal to the lexical representations stored in memory in a continuous manner and selection of the correct lexical candidate (Marslen-Wilson, 1987). For our stimulus words, which contained compound words, the uniqueness point and selection of one lexical candidate for semantic access was fairly late (on average 500 ms), which may have contributed to the late time window for the semantic decoding.

Both phoneme and acoustic information could be decoded from the MEG signal best at a latency of 100–180 ms, suggesting that those representations might exist simultaneously in the cortical activation sequence. It has to be noted, that our decoding model was not optimally designed for decoding categorical variables such as the phoneme labels which probably explains the overall lower

decoding accuracy for phoneme labels (73%). However, the similarity of the optimal lags in spectrogram and phoneme decoding is not affected by this confound. Several other studies also indicate that the optimal temporal integration window for parsing speech acoustics into linguistic units might lie within this range. The amplitude envelope of continuous speech is tracked by phase-locked cortical oscillations at 100- to 180-ms latency (Abrams et al., 2008; Aiken and Picton, 2008; Koskinen et al., 2013), and this has also been shown to be approximately the time frame in which the categorical neural organization for phonemes transiently emerges (Chang et al., 2010). Also, studies with isolated speech syllables have demonstrated that the evoked responses can be seen as a combination of transient “impulse responses” to the onsets of constituent phonemes, with latencies of ~100–200 ms (Ostroff et al., 1998; Tremblay et al., 2003). However, the performance of the convolution model for spectrogram and phoneme decoding was fairly high also at 20- to 100-ms lag, suggesting that the acoustic-phonetic content of the words might in fact be tracked over multiple different integration windows. This may reflect the encoding of phonemes with different temporal characteristics (Khalighinejad et al., 2017), but may also be related to simultaneous phoneme and syllable level encoding, similarly to what has been suggested in entrainment of nested cortical oscillations (Ding and Simon, 2014; Ding et al., 2016). However, obtaining temporally and spatially distinct neural response fields for each phoneme was not the goal of this study, and forward models would be more useful for this purpose.

Categorical perception of phonemes is a well-established phenomenon in behavioral studies, and its development is the basis of tuning to mother tongue in early infancy. However, the neural underpinnings of this phenomenon are still debated. Previous neuroimaging studies have shown that the superior temporal areas are tuned to phoneme categories, which rely on integration of several spectral and temporal cues (Formisano et al., 2008; Mesgarani et al., 2014). In the current results, spoken word decoding improved, compared with the spectrogram, when we included a combination of the spectrogram and the phoneme sequence of the word. These results echo those of a recent EEG study, where articulatory features of phonemes and the spectrogram together were the best model for decoding continuous speech based on phase-locked cortical oscillatory activation (Di Liberto et al., 2015). These results together might be interpreted to suggest pre-lexical categorical representations that are separate from the analysis of acoustic properties of the speech signal. However, a recent MEG study (Daube et al., 2019) suggested that the gain in predictive power of the phonemic/articulatory features may be explained by acoustic features of the phonemes: When including Gabor patterns of the sounds (different spectrotemporal modulations akin to our MPS model, with a time series) or even only the phoneme onset timings, performance gain was similar as when the categorical articulatory features were combined with the spectrogram. In that study, the best model was a relatively simple acoustic feature space that focused on acoustic edges. This might suggest that purely acoustic models are sufficient in explaining MEG/EEG

responses to speech. Thus, it remains unclear what the successful decoding of categorical phonemes or phonemic features across studies tells us about prelexical stages of speech processing, and further research is needed.

In any case, the current results are compatible with the idea that linguistic representations emerge directly from tuning to the complex spectrotemporal acoustic features characteristic of different phonemes (Mesgarani et al., 2014; de Heer et al., 2017) and that the quasi-rhythmic changes in the amplitude envelope at syllable boundaries might be crucial in guiding this process. As noted above, the processing step where the continuous acoustic signal interfaces with phonological representations is only required for speech, and this is where time-locked encoding for speech appears to come into play. Decoding performance was similar for pseudowords and real words, indicating that this mode of encoding is related to the low-level parsing of the acoustic signal, and might not be directly affected by the lexical-semantic status of the speech utterance. However, this warrants further examination. Speech processing is interactive, such that top-down influences from semantic and syntactic context are expected to influence low-level acoustic processing, particularly with sentence-level and narrative speech. Indeed, a recent fMRI study showed that not only the spectral content and phonemic features, but also semantic features within narrative speech are represented in highly overlapping regions early in the acoustic processing hierarchy (de Heer et al., 2017), indicating top-down influences of semantic content on acoustic and phonological processing. In the current study, cortical sources contributing to acoustic and phoneme decoding were both concentrated around bilateral auditory cortices and did not show clear lateralization, consistent with previous studies (de Heer et al., 2017; Brodbeck et al., 2018) and compatible with the view that acoustic-phonetic processing is bilaterally implemented (Poeppel, 2003).

### Future directions

The present finding of time-locked encoding for speech, but not other sounds, deepens the understanding of the computations required for mapping between acoustic and linguistic representations. The current findings raise the question of what specific aspects within sounds are crucial for cueing the brain into using this special mode of encoding. Future work could investigate the contribution of different statistical properties within speech acoustics by using synthesized stimuli, the possible effect of experimental task to boost the use of time-locked or time-averaged mode in sound processing, and the role of top-down semantic contributions using real-life like auditory environments. The rapidly computable convolution model for high-dimensional MEG/EEG signals can be further developed for decoding of a time-varying modulation representation, which might even better model the time-locked cortical encoding of speech (Pasley et al., 2012). Finally, the present findings may in the future help to bridge the gap between investigations of cortical temporal tracking of continuous speech and isolated linguistic stimuli.

## References

- Abrams DA, Nicol T, Zecker S, Kraus N (2008) Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *J Neurosci* 28:3958–3965.
- Aiken SJ, Picton TW (2008) Human cortical responses to the speech envelope. *Ear Hear* 29:139–157.
- Akbari H, Khalighinejad B, Herrero JL, Mehta AD, Mesgarani N (2019) Towards reconstructing intelligible speech from the human auditory cortex. *Sci Rep* 9:874.
- Alexandrou AM, Saarinen T, Kujala J, Salmelin R (2018) Cortical entrainment: what we can learn from studying naturalistic speech perception. *J Cogn Neurosci* 30:1704–1719.
- Aramaki M, Marie C, Kronland-Martinet R, Ystad S, Besson M (2010) Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. *J Cogn Neurosci* 22:2555–2569.
- Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D (1991) Reading a neural code. *Science* 252:1854–1857.
- Bishop C (2006) Pattern recognition and machine learning. Berlin: Springer.
- Brodbeck C, Presacco A, Simon JZ (2018) Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. *Neuroimage* 172:162–174.
- Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr Biol* 28:803–809.
- Chan AM, Halgren E, Marinkovic K, Cash SS (2011) Decoding word and category-specific spatiotemporal representations from MEG and EEG. *Neuroimage* 54:3028–3039.
- Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13:1428–1433.
- Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* 118:887–906.
- Coffey EB, Herholz SC, Chepesiuk AM, Baillet S, Zatorre RJ (2016) Cortical contributions to the auditory frequency-following response revealed by MEG. *Nat Commun* 7:11070.
- Correia JM, Jansma B, Hausfeld L, Kikkert S, Bonte M (2015) EEG decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Front Psychol* 6:71.
- Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuron* 9:179–194.
- Dale AM, Liu AK, Fischl BR, Buckner RL, Belliveau JW, Lewine JD, Halgren E (2000) Dynamic statistical parametric mapping: combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron* 26:55–67.
- Daube C, Ince RAA, Gross J (2019) Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. *Curr Biol* 29:1924–1937.
- de Heer WA, Huth AG, Griffiths TL, Gallant JL, Theunissen FE (2017) The hierarchical cortical organization of human speech processing. *J Neurosci* 37:6539–6557.
- Destrieux C, Fischl B, Dale A, Halgren E (2010) Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53:1–15.
- Di Liberto GM, O’Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr Biol* 25:2457–2465.
- Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. *Front Hum Neurosci* 8:1–7.
- Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage* 88:41–46.
- Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 19:158–164.

- Doelling KB, Poeppel D (2015) Cortical entrainment to music and its modulation by expertise. *Proc Natl Acad Sci USA* 112:E6233–E6242.
- Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *Neuroimage* 85:761–768.
- Elliott TM, Theunissen FE (2009) The modulation transfer function for speech intelligibility. *PLoS Comput Biol* 5:e1000302.
- Faisal A, Nora A, Seol J, Renvall H, Salmelin R (2015) Kernel Convolution Model for Decoding Sounds from Time-Varying Neural Responses. *International Workshop on Pattern Recognition in NeuroImaging*, pp 49–52. 10–12 June 2015, Stanford, CA.
- Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *Neuroimage* 9:195–207.
- Formisano E, De Martino F, Bonte M, Goebel R (2008) Who” is saying “what”? Brain-based decoding of human voice and speech. *Science* 322:970–973.
- Frey A, Aramaki M, Besson M (2014) Conceptual priming for realistic auditory scenes and for auditory words. *Brain Cogn* 84:141–152.
- Friedman JH, Tibshirani R, Hastie T (2001) The elements of statistical learning: data mining, inference, and prediction, Vol 1. Berlin: Springer.
- Gehr DD, Komiya H, Eggermont JJ (2000) Neuronal responses in cat primary auditory cortex to natural and altered species-specific calls. *Hearing Res* 150:27–42.
- Gourévitch B, Eggermont JJ (2007) Spatial representation of neural responses to natural and altered conspecific vocalizations in cat auditory cortex. *J Neurophysiol* 97:144–158.
- Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS (2014) Mne software for processing MEG and EEG data. *Neuroimage* 86:446–460.
- Hendrickson K, Walenski M, Friend M, Love T (2015) The organization of words and environmental sounds in memory. *Neuropsychologia* 69:67–76.
- Hämäläinen M, Ilmoniemi RJ (1994) Interpreting magnetic fields of the brain: minimum norm estimates. *Med Biol Eng Comput* 32:35–42.
- Jayant NS, Noll P (1984) Digital coding of waveforms. Upper Saddle River: Prentice Hall.
- Kanerva J, Ginter F (2014) Post-hoc manipulations of vector space models with application to semantic role labeling. *Proceedings of the 2nd Workshop on Continuous Vector Space Models and Their Compositionality (CVSC)*, pp 1–10. 26–30 April, 2014, Gothenburg, Sweden.
- Khalighinejad B, Cruzatto da Silva G, Mesgarani N (2017) Dynamic encoding of acoustic features in neural responses to continuous speech. *J Neurosci* 37:2176–2185.
- Koskinen M, Viinikanoja J, Kurimo M, Klami A, Kaski S, Hari R (2013) Identifying fragments of natural speech from the listener’s MEG signals. *Hum Brain Mapp* 34:1477–1489.
- Lin FH, Witzel T, Ahlfors SP, Stufflebeam SM, Belliveau JW, Hämäläinen MS (2006) Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. *Neuroimage* 31:160–171.
- Marslen-Wilson W (1987) Functional parallelism in spoken word recognition. *Cognition* 25:71–102.
- McDermott JH, Schemitsch M, Simoncelli EP (2013) Summary statistics in auditory perception. *Nat Neurosci* 16:493–498.
- Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J Neurophysiol* 102:3329–3339.
- Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. *Science* 343:1006–1010.
- Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013) Distributed representations of words and phrases and their compositionality. arXiv 1310.4546.
- Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–1195.
- Nagarajan SS, Cheung SW, Bedenbaugh P, Beitel RE, Schreiner CE, Merzenich MM (2002) Representation of spectral and temporal envelope of twitter vocalizations in common marmoset primary auditory cortex. *J Neurophysiol* 87:1723–1737.
- Obleser J, Kayser C (2019) Neural entrainment and attentional selection in the listening brain. *Trends Cogn Sci* 23:913–926.
- Ostroff JM, Martin BA, Boothroyd A (1998) Cortical evoked response to acoustic change within a syllable. *Ear Hear* 19:290–297.
- Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 18:903–911.
- Palatucci M, Pomerleau D, Hinton GE, Mitchell TM (2009) Zero-shot learning with semantic output codes. *Adv Neural Inf* 22:1410–1418.
- Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing speech from human auditory cortex. *PLoS Biol* 10:e1001251.
- Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cereb Cortex* 23:1378–1387.
- Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time. *Speech Commun* 41:245–255.
- Price C, Thierry G, Griffiths T (2005) Speech-specific auditory processing: where is it? *Trends Cogn Sci* 9:271–276.
- Renvall H, Formisano E, Parviainen T, Bonte M, Vihla M, Salmelin R (2012) Parametric merging of MEG and fMRI reveals spatiotemporal differences in cortical processing of spoken words and environmental sounds in background noise. *Cereb Cortex* 22:132–143.
- Salmelin R (2007) Clinical neurophysiology of language: the MEG approach. *Clin Neurophysiol* 118:237–254.
- Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput Biol* 10:e1003412.
- Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, Yacoub E, Formisano E (2017) Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc Natl Acad Sci USA* 114:4799–4804.
- Schirmer A, Fox PM, Grandjean D (2012) On the spatial organization of sound processing in the human temporal lobe: a meta-analysis. *Neuroimage* 63:137–147.
- Scott S, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. *Brain* 123:2400–2406.
- Shamma SA, Elhilali M, Micheyl C (2011) Temporal coherence and attention in auditory scene analysis. *Trends Neurosci* 34:114–123.
- Simanova I, van Gerven M, Oostenveld R, Hagoort P (2010) Identifying object categories from event-related EEG: toward decoding of conceptual representations. *PLoS One* 5:e14465.
- Steinschneider M, Nourski KV, Fishman YI (2013) Representation of speech in human auditory cortex: is it special? *Hearing Res* 305:57–73.
- Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, Salmelin R, Mitchell T (2012) Tracking neural coding of perceptual and semantic features of concrete nouns. *Neuroimage* 62:451–463.
- Taulu S, Simola J (2006) Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Phys Med Biol* 51:1759–1768.
- Tremblay KL, Friesen L, Martin BA, Wright R (2003) Test-retest reliability of cortical evoked potentials using naturally produced speech sounds. *Ear Hear* 24:225–232.

- Uusitalo M, Ilmoniemi RJ (1997) Signal-space projection method for separating MEG or EEG into components. *Med Biol Eng Comput* 35:135–140.
- Uutela K, Hämäläinen M, Somersalo E (1999) Visualization of magnetoencephalographic data using minimum current estimates. *Neuroimage* 10:173–180.
- Worden FG, Marsh JT (1968) Frequency-following (microphonic-like) neural responses evoked by sound. *Electroencephalogr Clin Neurophysiol* 25:42–52.
- Zatorre RJ, Gandour JT (2008) Neural specializations for speech and pitch: moving beyond the dichotomies. *Philos Trans R Soc Lond B Biol Sci* 363:1087–1104.