NeuroImage 219 (2020) 116936

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Brain activity reflects the predictability of word sequences in listened continuous speech

Miika Koskinen [a,b,c,d,*], Mikko Kurimo [e], Joachim Gross [c,f], Aapo Hyvärinen [g], Riitta Hari [b,h]

[a] Medicum, Faculty of Medicine, P.O. Box 63, FI-00014, University of Helsinki, Finland
[b] Department of Neuroscience and Biomedical Engineering, P.O. Box 12200, FI-00076, Aalto University, Finland
[c] Institute of Neuroscience and Psychology, University of Glasgow, 58 Hillhead Street, Glasgow, G12 8QB, UK
[d] MEG Core, Aalto NeuroImaging, FI-00076, Aalto University, Finland
[e] Department of Signal Processing and Acoustics, P.O. Box 13000, FI-00076, Aalto University, Finland
[f] Institute for Biomagnetism and Biosignalanalysis, University of Muenster, 48149, Muenster, Germany
[g] Department of Computer Science, P.O. Box 68, FI-00014, University of Helsinki, Finland
[h] Department of Art, P.O. Box 31000, FI-00076, Aalto University, Finland

ARTICLE INFO

Keywords:
Speech perception
Continuous speech
MEG
Language model
N-gram
Speech–brain coupling
Naturalistic neuroscience
Cerebral cortex

ABSTRACT

Natural speech builds on contextual relations that can prompt predictions of upcoming utterances. To study the neural underpinnings of such predictive processing we asked 10 healthy adults to listen to a 1-h-long audiobook while their magnetoencephalographic (MEG) brain activity was recorded. We correlated the MEG signals with acoustic speech envelope, as well as with estimates of Bayesian word probability with and without the contextual word sequence (N-gram and Unigram, respectively), with a focus on time-lags. The MEG signals of auditory and sensorimotor cortices were strongly coupled to the speech envelope at the rates of syllables (4–8 Hz) and of prosody and intonation (0.5–2 Hz). The probability structure of word sequences, independently of the acoustical features, affected the $\leq$ 2-Hz signals extensively in auditory and rolandic regions, in precuneus, occipital cortices, and lateral and medial frontal regions. Fine-grained temporal progression patterns occurred across brain regions 100–1000 ms after word onsets. Although the acoustic effects were observed in both hemispheres, the contextual influences were statistically significantly lateralized to the left hemisphere. These results serve as a brain signature of the predictability of word sequences in listened continuous speech, confirming and extending previous results to demonstrate that deeply-learned knowledge and recent contextual information are employed dynamically and in a left-hemisphere-dominant manner in predicting the forthcoming words in natural speech.

1. Introduction

Humans predict the future and interprets the present based on prior experience. In the rich sensory environment of every-day life, the predictions take place continuously, for example during listening to ongoing natural speech—a very complex sensory input that healthy adults are highly experienced in. In natural conversations, the silent periods between the turns of two speakers are surprisingly short, on average 250 ms, across all languages and cultures (Stivers et al., 2009). So quick turn-takings imply predictive comprehension of the other participant's speech because purely reactive turn-taking would take considerably longer (Levinson, 2016). Thus, an active listener continuously updates the predictions of the forthcoming speech at several conceptual levels and temporal scales. Altogether, the alignment to other person's speech is a good example of smooth social interaction (Hari et al., 2015).

In continuous speech, information unfolds serially, building on context and resulting in complex sequential dependencies between phonemes, words, and sentences. These contextual dependencies comprise multiple time scales and are learned during healthy development. Listening to natural speech therefore leads to predictions of upcoming utterances, words, and even narratives. The corresponding brain mechanisms should preferably be studied in similar naturalistic conditions.

In the human brain, the temporal processing scales tend to be ordered so that the shortest time windows—from milliseconds to hundreds of milliseconds, or even seconds—occur close to early sensory areas whereas higher-order processing integrates information over much

* Corresponding author. PO Box 400, 00029, HUS, Finland.
E-mail address: miika.koskinen@hus.fi (M. Koskinen).

https://doi.org/10.1016/j.neuroimage.2020.116936
Received 19 June 2019; Received in revised form 24 April 2020; Accepted 7 May 2020
Available online 29 May 2020

longer time scales, up to tens of seconds in multiple brain areas (Hari et al., 2010; Hari and Parkkonen, 2015; Hasson et al., 2008; Jääskeläinen et al., 2011; Lerner et al., 2011; Stephens et al., 2013).

Given the complexity of natural speech and the multiple time scales of the associated context-dependencies, the identification of the brain correlates of word prediction becomes difficult both in principle and in practice. Functional magnetic resonance imaging (fMRI) has been successfully used to pinpoint brain areas affected by predictability of words in natural speech in the sense of entropy and surprisal (Willems et al., 2016). However, time-accurate brain-recording methods, such as magnetoencephalography (MEG), electroencephalography (EEG), or electrocorticography (ECoG), are needed for exploration of the underlying brain events at time scales shorter than 1 s. MEG oscillations below 10 Hz show a robust relationship to continuous auditory speech (e.g. Park et al., 2015) and may indicate predictive neural processing. For example, in macaque auditory cortex (Márton et al., 2019), frequencies below 10 Hz are involved in low-to-high cross-frequency interactions assumed to reflect top-down information processing in a predictive-coding framework. In humans, neural entrainment to speech rate tends to persist after an abrupt change of the speech rate which biases the perception of subsequent words (Kösem et al., 2018). Moreover, in strongly constraining contexts, 4–7-Hz EEG power is increased as a response to unexpected words (Rommers et al., 2017).

Because it is difficult to unravel the neural correlates of context-dependent predictive behavior in a naturalistic experimental setting, we first assumed that the discrepancy between expectations (predictions) and actual sensory input is visible in the brain signals occurring during speech listening. We then applied probabilistic language modeling to assign the conditional probability for each word given the previous words (and the prior probabilities of words in Finnish language). In other words, large probabilities were assigned to predictable words, and vice versa, and we analyzed brain activity by correlating the signals occurring after word onsets with the conditional word probabilities assigned by the model.

Here we focused on continuous low-frequency MEG fluctuations while looking for brain correlates of word predictability. We used 1-h recordings where subjects were attentively listening to narrative speech in an audiobook. Instead of spectral or coherence estimates often used in studies of high-frequency brain oscillations in specific time-windows, we chose a wavelet-transform that sensitively reveals changes in continuous brain activity. The resulting signals were then cross-correlated with the sequence of estimated (log) probabilities by the language model to provide detailed information of the lags of stimulus effects in MEG signals, thus complementing the often-used coherence computations with fixed lags.

Our two language models—one based on word probabilities (~word frequencies) and the other on conditional word probabilities given the recent context, i.e. the previous words in the sequence— have conceptually different interpretations, and we found that their brain correlates differ.

In the analysis, we tagged each word by its language-model probability estimate provided by an N-gram language model that had been trained using a large Finnish text corpus. Instead of a typical word N-gram model, we applied a subword varigram model that fits better to the structure of the Finnish language (Hirsimäki et al., 2009). Finnish is a morphologically rich language which does not have prepositions nor articles. Instead, the words are typically composed of inflections, compounds, and several subsequent suffixes that make the words longer and more complex compared with English. Thus, it is not practical for Finnish to build the language model of word units, because then the vocabulary would need more than a million entries. Moreover, the length of the required subword sequences to cover inflated words varies considerably, which makes also the fixed-N models, often used in psycholinguistic literature and in neuroscientific and speech-recognition applications, unpractical for the Finnish language. The established solution used in speech recognition and machine translation in Finnish and in other morphologically rich languages is to build subword varigram models to capture the linguistic relations in sentences needed to estimate the word probabilities.

Brain mechanisms activated in naturalistic conditions would be ideally studied using analysis approaches that bypass the need for stimulus repetition and response averaging. We have previously applied canonical correlation analysis to maximize between-subjects correlation after optimizing spatial filtering of non-averaged MEG signals from single sensors (e.g. Lankinen et al., 2014; Koskinen et al., 2016). More recently, models including linear stimulus–response relationships have been used to relate e.g. speech envelope to continuous brain activity (Lalor and Foxe, 2010), word frequency and acoustic power to broad-band (1–40 Hz) MEG signals (Brodbeck et al., 2018), as well as semantic dissimilarities between words in a listened narrative to 1–8 Hz EEG activity (Broderick et al., 2018). While the models may represent estimates of ensemble averages, stimulus repetition can be avoided. Here we used an alternative method based on cross-correlations and language modelling to detect acoustic and predictability-related modulations of brain signals by single presentation of the listened narrative, without the need to average brain signals.

Our main goal was to differentiate the predictability effects from acoustic effects in the MEG signals. We thus computed correlations between source-level brain activity—at different time-lags and frequency bands—and the probability-estimate sequence of the words by using two language models to find out where and when the brain activity is consistently affected by the context of the words. Confirming previous findings with English language and N-gram modeling, our study demonstrates the relevance of varigram modeling for Finnish as a highly inflecting, morphologically rich language. We also show that different frequency components of the ongoing MEG are differentially related to these stimulus properties.

Preliminary results of these data have been presented in abstract form (Koskinen et al., 2016).

2. Materials and methods

2.1. Materials

Ten healthy adults (5 females, 5 males; mean age 33 yrs, range 22–62 yrs) participated after informed consent. The protocol was approved by the ethics committee of the Helsinki and Uusimaa Hospital District.

Whole-scalp MEG was acquired with 306-channel Elekta Neuromag™ neuromagnetometer (Elekta Oy, Helsinki, Finland) at the MEG Core of Aalto NeuroImaging, Aalto University, Finland. Participants were comfortably seated under the MEG dewar within the magnetically shielded room. They listened to selected parts of the classic novel *Välskärin kertomuksia* by Z. Topelius. Although the narrative is over hundred years old (Finnish translation published in 1896) and differs somewhat in vocabulary and style from modern Finnish, its language is fully comprehensible and pleasant to contemporary adult listeners. Notably, the computational language models that we used to explain a part of listeners' brain activity, were trained with contemporary texts (as were our participants, of course). The story, lasting 1-h in total and containing 7230 words, was played via a non-magnetic open-field audio speaker (Panphonics Ltd., Tampere, Finland) at comfortable loudness level. The story was presented only once but divided into two recording sessions carried out on separate days, with two ~15-min pieces played on each day; for more details, see Koskinen and Seppä (2014).

The recording passband of MEG was 0.03–330 Hz and the sampling rate 1000 Hz. Anatomical T1-weighted magnetic resonance images (MRIs) of the brain were available for six subjects from prior studies with permission (3-T scanner at the AMI Centre, AaltoNeuroimaging, Aalto University, Finland). For four subjects, we used a template brain model ('fsaverage', Freesurfer software; http://surfer.nmr.mgh.harvard.edu/), scaled to match individual digitized head shape (MNE Python Toolbox; Gramfort et al., 2013).

2.2. MEG preprocessing

MEG data were down-sampled to 200 Hz applying an antialiasing FIR low-pass filter. External interference signals coming from the outside of the head were reduced using spatial signal-space-separation (SSS) technique (Taulu et al., 2004). No further artifact processing was applied. MEG analysis was performed using custom Matlab and Python scripts, utilizing MNE Python and MNE Matlab Toolboxes.

The MEG signals were transformed by Mexican-hat wavelet to seven timescales, corresponding to e-folding times (Torrence and Campo, 1998) from 660 ms to 44 ms or center frequencies from 0.5 to 8 Hz, respectively. The Mexican-hat wavelet was chosen due to its temporal accuracy, minimal temporal blurring, avoidance of filter 'ringing' effects, and sensitivity to both transients and expectedly subtle deviations in (non-stationary) oscillatory rhythms that we were interested in.

Anatomical MR images were co-registered with MEG by digitized fiducial markers and head-shape points on the scalp. A single-layer boundary-element model (BEM) of the brain surface was constructed from anatomical MR images with FreeSurfer software. The minimum-norm source-current estimates (MNEs; Hämäläinen and Ilmoniemi, 1994) on the cortical surface were computed for wavelet-transformed data of 306 MEG channels with free current-dipole orientations and without depth weighting for 2562 sources per hemisphere. The noise covariance matrices were computed from empty-room data recorded on the same days as the participants' data. Source space time-series of the four separate recordings were concatenated and treated as continuous 1-h data. For three dipole orientations at each point in source space, the signal component (derived from principal component analysis) explaining the largest signal variance was used in the subsequent analysis.

2.3. Segmentation of the speech stream to words

The continuous speech was segmented into words and morphemes by applying the AaltoASR speech- recognition system (Hirsimäki et al., 2009) for forced alignment between speech and text. The best alignment was based on the system's acoustic phoneme models trained on speech data of hundreds of native Finnish speakers by using the stochastic hidden-Markov modeling framework.

2.4. Language models

2.4.1. N-grams and cloze probability

The probability of a word sequence $w_1, \ldots, w_M$ of length M can be expressed as

$$P(w_1, \ldots, w_M) = \prod_{k=1}^{M} P(w_k | w_{1\ldots k-1}) \quad (1)$$

by the chain rule. Probabilities in Eq. (1) are estimated from counting occurrences of word sequences in large text corpora. As counting gets impractical with long sequences, a preferable choice is to approximate the conditional probabilities in Eq. (1) by limiting the scrutiny to N words by

$$P(w_k | w_{1\ldots k-1}) \approx P(w_k \quad | w_{k-N+1}, \ldots, w_{k-1}). \quad (2)$$

Eq. (2) is a generative stochastic model of word sequences, called N-gram. The value of N is typically 1–5. As natural language data are typically sparse (i.e. the observed words or sequences are rare or non-existing even in large training sets), a common procedure is to smooth these probabilities by moving a part of the probability mass to unseen events and interpolating rare N-grams with N-grams of smaller N (Chen and Goodman, 1999). Evidently, Eq. (2) is related to the notion of cloze probability referring to the probability of a word given the context, as subjectively assessed by human observers. However, arguments have been presented about the biased nature of cloze relative to corpus probabilities (Smith and Levy, 2011).

2.4.2. Varigram model

As stated above, Finnish is a highly inflecting, morphologically rich language where the size of the vocabulary expands tremendously with different subsequent suffixes and compounds but with no prepositions nor articles. Thus, we built the language model using morpheme-based subword units (Hirsimäki et al., 2006). Because the length of the common morpheme sequences may in some cases extend over 10 units, traditional fixed-N models are impractical, and we applied variable-length N-grams called varigrams (Niesler and Woodland, 1999), where the value N is adaptive and depends on the actual word sequence. The purpose was to gain higher modeling accuracy by allowing larger N values for common sequences, as long as the increasing N provided relevant information. The N-gram probabilities in varigram model are smoothed and computed just like in the normal N-gram.

The models were implemented using the open source VariKN tool (Siivola et al., 2007) developed at Aalto University to estimate the optimal varigram model for a text corpus. The algorithm ensures that smoothing of the resulting varigram probabilities follows the state-of-the-art Kneser-Ney smoothing (Chen and Goodman, 1999). We trained the model for a standard written-style Finnish (Finnish Text Collection 2004; http://www.kielipankki.fi/).

Because the model was built from subword sequences, it also estimated the probability for the next subword. Importantly, the probability of a whole word can be composed as a product of the (conditional) probabilities of its subwords (Hirsimäki et al., 2006) similarly as the probability of a word sequence can be composed as a product of probabilities of single words. In the analysis, we compared the probability of the current word computed with and without the context of the previous words to find the contribution of the previous words in predicting the next word. In the varigram model built from subword units, we separated the contributions of the subwords of the current word (called here the unigram probability) from the full subword sequence covering also the previous words (called here the varigram probability).

2.5. Analysis

Fig. 1 shows schematically the starting point of our analysis. First, the transcripted audio narrative was associated word-by-word with the output of the language model (here the N-gram), the time course of which was then correlated with the resampled MEG time-series in the desired frequency band, separately for each individual.

We used the language models to test the hypothesis that the neural activity in brain areas related to speech processing—in addition to being temporally correlated to the acoustical features of the speech—is modulated by word probabilities. To this end, we computed group-level cluster statistics on cross-correlations between each individual's MEG signal (in specific frequency bands) and logarithmic word probability, estimated by the unigram language model that does not contain contextual cues. Importantly, in this analysis the effects that can be explained by acoustical features of the speech were removed by partial correlation analysis (see Section 2.6 below for details).

Next, having demonstrated that speech-related brain activity is sensitive to word probabilities, we tested the hypothesis that brain signals show the effect of (are correlated with) the sequence of conditional word probabilities. This analysis was carried out by incorporating model-based predictions (N-gram) of the next word given the prior word sequence. Importantly, and consistent with the previous analysis, correlations that can be explained by the speech envelope had been removed.

2.6. Clustering statistics

Down-sampled and wavelet-transformed MEG time-series were first segmented into 1000-ms epochs time-locked to all 7230 word onsets, and then the MEG amplitudes at a given delay between 0 and 1000 ms with respect to stimulus were correlated with the sequence of corresponding logarithmic conditional word probabilities. We used partial correlations

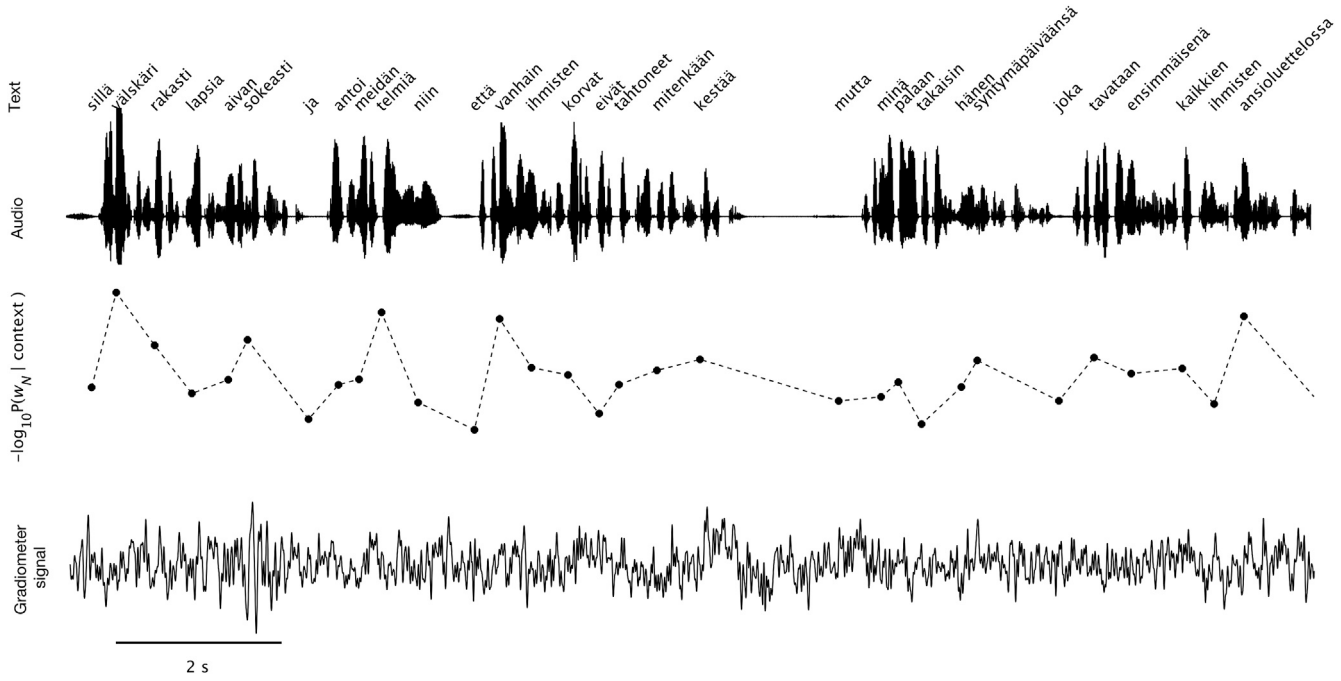

Fig. 1. Schematics of the analysis procedure. Top: The audio signal with the transcripted spoken text (in Finnish). Middle: Output of the N-gram language model; the trace was correlated with the resampled MEG time-series. Bottom: Wideband (0.5–40 Hz) planar-gradiometer MEG signal from one planar sensor over the left auditory cortex.

to control for the effects of speech envelope that reflects acoustic intensity changes, such as accents that emphasize syllables or words and may correlate with word informativeness directly and thereby inversely with word probabilities (e.g. Pan and Kathleen, 1999). For each source point, partial correlation was computed between the word probability time series by language models and MEG amplitude (wavelet transformed) time-series sampled at a specific lag relative to the word onsets. The lags from 0 to 1000 ms were separately investigated. Speech envelope was similarly sampled relative to word onsets, for lags between −200 ms and 1000 ms, resulting in the matrix used for controlling envelope effects in the partial correlation computations. The procedure was repeated separately for each subject, wavelet scale, and source point. For one participant, only ~75% of data were used due to MEG-recording artifacts.

For comparison, MEG and speech envelope time-series were cross-correlated in a conventional manner using the maximum delay of 1000 ms for all the 1-h-long data.

The group statistics of correlations was computed using cluster-level statistics based on (non-parametric) permutation of paired t-tests for spatio-temporal data (Maris and Oostenveld, 2007; MNE Python toolbox 'spatio_temporal_cluster_1samp_test'), thereby addressing the multiple comparisons problem across space and time ($p < 0.05$) with Bonferroni correction over seven frequency bands. For statistical testing, (cross-) correlation coefficients at each source point were first morphed into the common anatomical template ('fsaverage', FreeSurfer software), spatially smoothened and Fisher-transformed, resulting in 1000-ms time-series at each 20484 vertices of the cortical surface model. Due to the arbitrariness of the sign in principal component analysis at the MEG preprocessing phase, we used the absolute values of the correlation coefficients. Paired contrasts were formed by subtraction between the correlation coefficients obtained in the experiment and control conditions. In the control condition, the correlation was calculated between the original MEG data and the probability sequence circularly shifted by approximately half of the number of words. The circular shifting here corresponded to switching the order of two halves of a sequence. Next, the N-gram – Unigram contrast was formed by subtracting the absolute values of the corresponding correlation coefficient timeseries.

We identified clusters contributing to rejection of null hypothesis in statistical analysis. Thus, the plotted values should not be directly interpreted as statistical significances of single source points (Sassenhagen and Draschkow, 2019). We marked sources and time-lags of the cluster as ones in a new sources-by-times matrix and zeros elsewhere. Deep structures below cingulate cortex were masked out. Time series in presented figures represent summation of this matrix along source-axis, divided by 10242, the number of vertices (sources) in one hemisphere. Corresponding anatomical maps represent the duration spanned by the cluster (code 'mne.stats.summarize_clusters_stc'; Gramfort et al., 2013). Instantaneous *t*-statistic maps are presented in Supplementary video animations.

Similar statistics was used for hemispheric comparison of the correlation coefficients where the contrasts were computed between symmetric vertex positions of the BEM model in two hemispheres. The closest corresponding vertices in contralateral hemisphere were found by mirroring the MNI coordinates of each BEM vertex to the contralateral side.

3. Results

3.1. Coupling between brain activity and speech envelope

Fig. 2 shows the results of speech–MEG coupling, that is the effects of the acoustic envelope, aimed to produce a baseline for our study of the contextual influences on brain activity. The spatial distribution and the delay-structure of clustering is illustrated for frequencies between 0.5 and 8 Hz. As expected, the main effects center around auditory cortices, rolandic sensorimotor cortices and frontal lobes in both hemispheres. The lower frequencies (0.5 and 1 Hz) displayed additional correlation clusters also in frontal lobes, including the frontal eye fields. The delays (shown in the right column) concentrated below 400 ms, and the speech–brain correlations decreased with increasing frequency. The Supplementary Videos "Envelope" display the temporal evolution of instantaneous *t*-statistic maps for the speech–brain coupling and suggests that the coupling varies between brain areas at different time-lags. The videos also show the involvement of medial brain structures, such as precuneus and medial prefrontal cortex. These results form the basis of

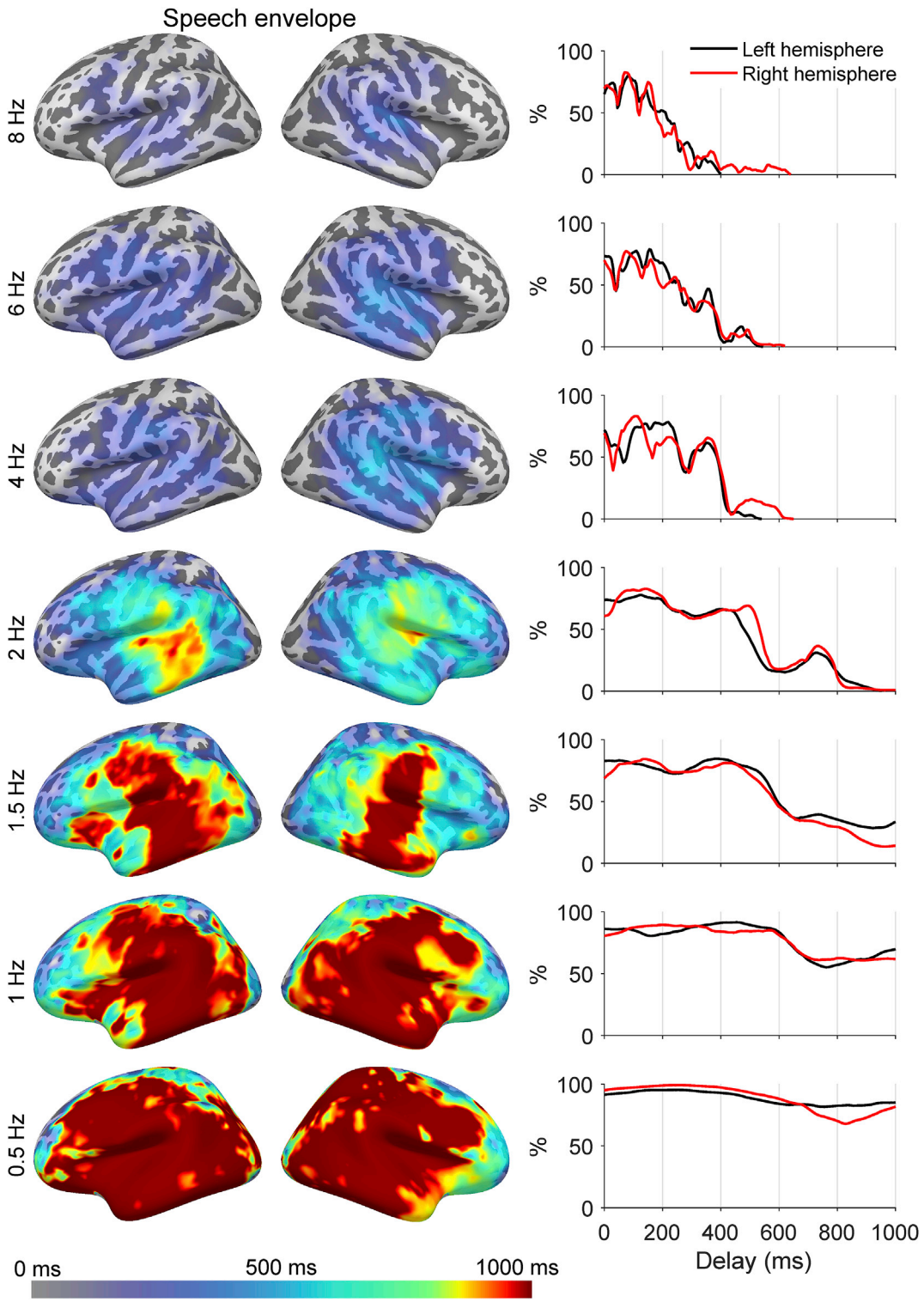

Fig. 2. Correlation of MEG signals with speech envelope. Cluster-level statistical analysis revealed a significant difference between conditions ($p < 0.05$, Bonferroni corrected). The colors code the time span (ms) of the cluster. On the right, the traces quantify the spatial extent of the cluster as the percentage of source points at each time-lag, separately for the left and right hemispheres (black and red lines, respectively).

speech–brain coupling, an acoustic baseline, with which we compare the effects of word probabilities.

3.2. *Effects of word and conditional word probabilities*

Fig. 3 shows the spatial distributions of cross-correlations between MEG signals and word probabilities from the Unigram model (without contextual cues) on the left, from the N-gram model (with conditional word probabilities) in the middle, and their comparison on the right. As noted before, in these analyses the correlations that can be explained by the acoustical envelope have been removed.

Brain signals correlated statistically significantly with the Unigram word probability sequence in the 0.5-Hz range in bilateral temporal lobes and in left frontal and inferior rolandic areas (Fig. 3, left panel and

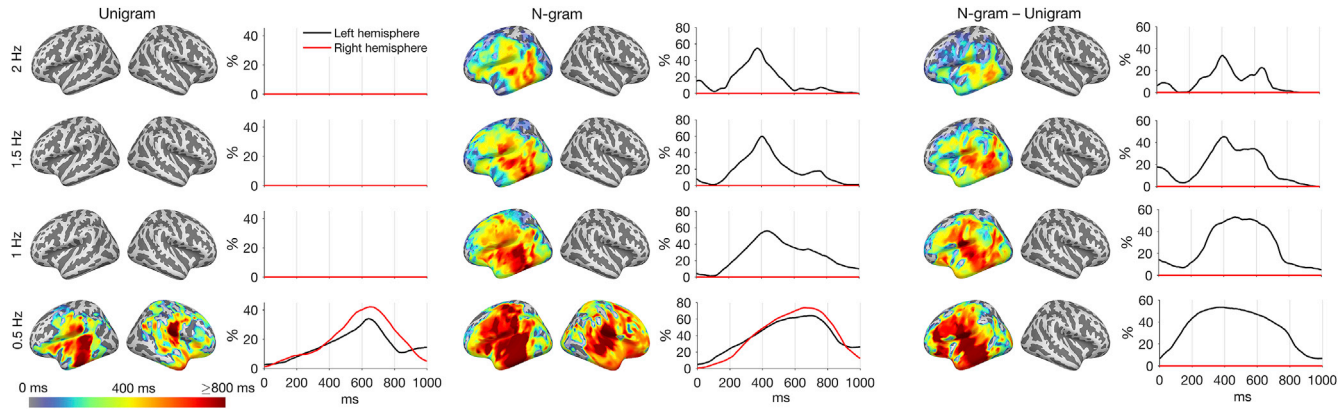

Fig. 3. Cortical activity in four frequency bands associated with estimated word probability sequence by Unigram (left) and N-gram (middle). Right: Speech–brain coupling associated with contextual effects, given as N-gram – Unigram contrast. Color-coding as in Fig. 2.

Supplementary Videos "Unigram"). The temporal progression of the correlations revealed differences in word probability effects between brain areas. From 200 to 300 ms, the correlation varied widely in the temporal lobes and rolandic areas, and later on (after 500 ms), lateral and medial prefrontal cortices and the right temporal pole regions got involved. Because the Unigram model produced statistically significant correlations only in the lowest frequency band (see below), we from this on focus on the N-gram results and on the differences between the N-gram and Unigram models.

Fig. 3 (middle panel) shows that the most consistent correlations between MEG signals and conditional word probabilities (N-gram) concentrate to the temporal lobes, lower rolandic areas and to frontal regions. Statistically significant correlations were in the 0.5-Hz range found in both hemispheres but at higher frequencies only in the left hemisphere. Hemispheric differences were statistically significant at 1.5 Hz and 2 Hz. The delay plots in Fig. 3, as well as the temporal dynamics in Fig. 4, show that the correlation effects were spatially most widespread at a delay of about 400 ms at 1–2 Hz, extending up to 700–800 ms at 0.5 and 1 Hz. Instantaneous correlations e.g. at 2 Hz (Supplementary Videos "N-gram") revealed fine-grained progression patterns typically starting from the left auditory cortex and insula, extending across the left temporal lobe, occipital and the lateral and medial frontal cortices, throughout the 1-s range of lags studied.

The difference between N-gram and unigram correlations (Fig. 3, right panel; Supplementary Videos "N-gram – Unigram") resulted in clustering only in the left hemisphere, involving widely the temporal lobe, the inferior rolandic areas, and parts of the frontal lobe. However, this hemispheric asymmetry when tested on the basis of correlation coefficients was not statistically significant. The most widespread context dependence occurred around 400 ms but with a wide tuning up to about 800 ms, depending on the frequency band. In superior and posterior temporal cortices, the difference was apparent already at ~100 ms. The medial frontal cortex was involved distinctively at 400 and 650 ms, and the posterior temporal cortex at 700 ms. Correlations disappeared after 850 ms.

4. Discussion

We found that when subjects listened to connected natural speech, their speech-related brain activity was modulated by the context of single words. Specifically, the MEG signals correlated with estimated context-related, conditional word probabilities computed by N-gram in temporal, frontal, parietal and occipital brain regions, including mesial cortex, at 2 Hz and below, revealing fine-grained spatiotemporal progression patterns of cortical activity. Importantly, the coupling delays differed between areas, revealing speech–brain coupling across most of the neocortex; the effects were visible from about 100 ms on throughout the studied 1000-ms analysis window, with prominent and widespread coupling around 400 ms. Before the analyses with language models, we had removed (via partial correlation) the effects of speech envelope as the most prominent confounding factor, so that the results represent the association between time-locked brain activity and the lexical-semantic probability structure of the listened narrative, while controlling for a large proportion of acoustical effects.

The evidence of predictive processing of on-going natural speech was obtained by using a novel combination of source-localized MEG data, language modeling, and statistical analysis. Our method—based on cross-correlations in specific frequency bands—allows to study the relationship

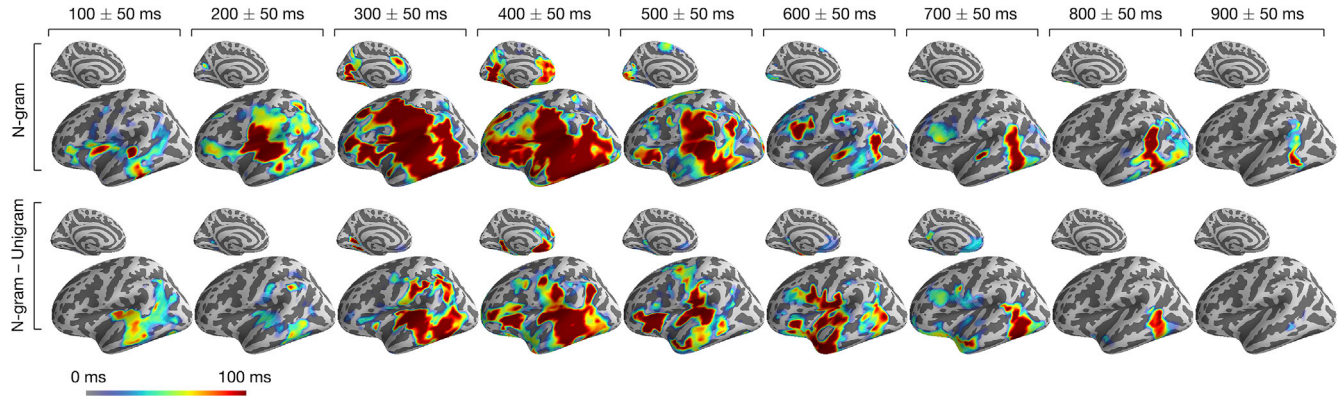

Fig. 4. Spatiotemporal progression patterns of predictability effects within the given time-lag at 2-Hz band, shown separately for the N-gram (upper part) and the N-gram – Unigram contrast (lower part). The small brain images display the mesial surfaces. The patterns were found only in the left hemisphere.

between stimulus features and frequency-specific brain responses at variable delays. We found that the acoustical features involved both hemispheres consistently across the frequency bands (Fig. 2). These patterns differed from language-model correlates. Importantly, even the brain correlates of two different word probability estimates were found to differ. Context-dependent word probabilities, N-grams, revealed left-hemisphere-lateralized effects in 1 Hz, 1.5 Hz and 2 Hz frequency bands (Fig. 3 and Supplementary videos), whereas the Unigram word probability estimates did not. In the lowest 0.5 Hz band, both estimates correlated with MEG signals in both hemispheres. Interestingly, the N-gram – Unigram contrast (specifically in 2 Hz), uncovering contextual effects, revealed a spatiotemporal progression pattern starting from superior and posterior temporal regions already around 100 ms, involving medial frontal cortex at 400 ms, towards temporal poles, lateral and orbitofrontal cortices, involving medial frontal cortex again at 650 ms, and disappearing after 850 ms at posterior temporal cortex. These results inform how short-term contextual information and deeply learned knowledge of language can dynamically and complementarily affect, at multiple levels of brain hierarchy, the processing of continuous speech.

4.1. Effects of acoustic speech envelope

Auditory cortices are known to react to listened speech with stimulus-locked activity (Luo and Poeppel, 2007). The phase of oscillatory auditory-cortex activity is temporally coupled to the speech envelope, especially at syllable rate of about 4–8 Hz (Bourguignon et al., 2013; Gross et al., 2013; Koskinen et al., 2013; Peelle et al., 2013). This coupling between brain activity and listened speech is related to how well the participants comprehend speech (Gross et al., 2013; Peelle et al., 2013).

The coupling between speech envelope and MEG signals, that we used as the baseline for finding out the context effects, thus reproduced previous reports of speech–brain coupling during listening to continuous speech (see above) but also additionally demonstrated that activity in rolandic and frontal brain areas—and thus not only of auditory areas—correlates with the speech envelope at frequencies corresponding to the syllable rate (4–8 Hz) and at lower frequencies corresponding to prosody and intonation (0.5–2 Hz). Thus, the rhythmic temporal structure of a continuous speech envelope was significantly and faithfully represented in the activity of bilateral temporal, frontal and sensorimotor areas, medial prefrontal regions, and precuneus.

4.2. Response delays

Our analysis approach enabled finding detailed spatiotemporal progression patterns of brain activity correlated to the speech signals, demonstrating that brain areas are differentially affected by the listened speech. Importantly, the dynamics was different for acoustic (correlation with speech envelope) and predictability effects. Specifically, correlation with acoustics showed most prominent coupling at short delays that decreased until about 400-ms delay for frequencies of 2–8 Hz. As expected, the slowly changing low-frequency oscillations below 1.5 Hz showed little temporal variation with delay.

The dominant delay of context effects at 400 ms corresponds to the peak latency of the N400 response observed in event-related EEG and MEG recordings to words with low cloze probabilities (i.e. probabilities that the word would smoothly complete a given sentence; Kutas and Hillyard, 1984) or to words with semantic violations or other “anomalies” (for reviews, see Kutas and Federmeier, 2011; Lau et al., 2008). Our results agree with the EEG findings of Broderick et al. (2018) interpreted to indicate that the “mapping function of semantic features to neural response shares traits with the N400”.

In our data, the probability sequences of words covaried with MEG signals but the correlations between MEG and the envelope and the two language models differed from each other both in spatial terms and in time-lags. Thus the relation between our results and conventional evoked responses remains unclear at this point, but the presented methodology and the observed progression patterns across a wide range of delays and in different brain areas provide information that would be difficult to obtain with typical evoked-response studies.

4.3. Predictions of future words

Generally, it is well established that information cumulated into memory is utilized predictively in human perception (Attneave, 1954; Friston, 2002; Mumford, 1992; Rao and Ballard, 1999; Schuls and Dickinson, 2000; Summerfield et al., 2006). It is likely that the rhythmicity of connected speech sets expectations for speech rate and that the sensitivity to speech features is controlled by changes in neuronal excitability (Peelle and Davis, 2012). Alignment of oscillations to stimuli might serve as a mechanism for attentional selection (Lakatos et al., 2008), modulated by top-down signals from left inferior cortex and sensorimotor cortices to optimize speech–brain coupling (Park et al., 2015).

Our work builds upon these and other prior studies on context-dependent sequence processing, now extending from sublexical features to the sequence of words, with respective timescales from milliseconds to seconds. For example, direct recordings from auditory cortex in the temporal lobe have shown that responses to transitions of sound segments (phonemes) within speech are modulated by the context, being different for English words versus non-words (Leonard et al., 2015). Moreover, scalp EEG recorded from subjects listening to audiobook passages has reflected, in addition to the acoustic features of the speech, also categorical phoneme-level speech processing (Di Liberto et al., 2015). The EEG frequencies below 9 Hz also reflected learned permissible phoneme sequences in English language (Di Liberto et al., 2019). However, these findings have recently been challenged through the use of more comprehensive acoustic feature models (Daube et al., 2019).

N-gram-based probability sequences of words in natural speech have been earlier associated with time-series of fMRI voxels, for example, in superior temporal gyri and anterior temporal poles of both hemispheres, in right amygdala, and in right inferior frontal sulcus (Willems et al., 2016). With respect to syntax and grammatical structures, syntactic complexity metrics have been shown to correlate with fMRI signals in anterior and posterior temporal lobe (Brennan et al., 2016).

Regarding continuous brain activity, our study relates to and extends recent intracranial recordings in humans and monkeys (Kikuchi et al., 2017) that order violations in a sequence of nonsense words (with respect to learned artificial grammar) result in transient low-frequency (4–8 Hz) oscillation coupling with >50 Hz (gamma) amplitude envelope in 450–700 ms latency in the auditory cortex. It is to be noted, however, that the predictability effects in language-related brain region may also be driven by the nested phrase structures of sentences, as has been suggested on the basis of intracranial recordings of high-gamma activity elicited by visual word-by-word presentation of sentences (Nelson et al., 2017).

Altogether our results extend previous studies on the relationship between brain activity and word context. In a statistical contrast of time windows corresponding to high and low word predictability, Armeni et al. (2019) reported significant differences at around 400–600 ms in the left hemisphere. While our results are in line with these findings, considerable differences exist between our approaches: the analysis of Armeni and coworkers was based on a few fixed delays (0, 200, 400, and 600 ms) while our cross-correlation approach captured the increasing delays for progressively more complex features (acoustics, word probability, context). Notably, spectral power inspected in an evoked response setting differs clearly from our approach where fluctuations of low-frequency signals were inspected in a naturalistic listening task. Similarly, using representational similarity analysis, Klimovich-Gray et al. (2019) showed left-hemisphere lateralized activation in frontal areas (BA45) that were sensitive to semantic context. Interestingly, Broderick et al. (2018) related EEG activity to semantic structure by

using a computational model of semantic dissimilarity combined with lagged regression. Similar to our results they revealed strongest correspondence at delays of 200–600 ms in centro-parietal EEG derivations. Our results (cf. Fig. 3) suggest that these effects are dominated by slow fluctuations below 3 Hz, widespread in the left hemisphere and involving mainly temporal and frontal brain areas. In the 0.5-Hz band, the effects were bilateral.

Our data also support previous findings that naturalistic stimuli can trigger wide-spread brain activation that would be more difficult to identify with strictly controlled stimuli (see Hamilton and Huth, 2018). The rich and complex characteristics of the natural continuous speech allowed us to inspect brain correlates of different stimulus characteristics—auditory envelope, word probability, and short-term contextual processing—in the very same experiment. The key idea was to use appropriate modelling as a proxy of hypothesized brain processing of language, in combination with automatic speech-recognition algorithms that enabled automatic estimation of a (conditional) probability for each of the several thousand words in the listened prose text. The applied models were explicitly tailored for Finnish language, and we expect similar approaches to be useful also for other highly inflecting, morphologically rich languages.

As a limitation, our study contained a small number of participants, although each listening to the audiobook for 1 h, i.e. considerably longer than what has been applied in previous studies. For source localization, some subjects did not have individual anatomical MRIs but the head model was assessed by a template, individually fitted to head size based on MEG-coil locations. Moreover, our reported results are based on group-level cluster statistics.

5. Conclusions

By using MEG recordings where the narrative stimulus, lasting for 1 h, was presented to a subject only once, we found that contextual information has a major impact on how words and sequential events are processed. We demonstrated that MEG low-frequency fluctuations convey predictability effects of speech input. In addition to deeply-learned knowledge of language use, recent prior information, accumulated in word sequences of natural speech, is dynamically and selectively deployed in moment-by-moment processing of the ongoing speech. Multiple brain areas in temporal lobes, but also in rolandic, prefrontal, lateral and medial frontal cortices, and in occipital cortex showed activity that could be explained to reflect predictive brain processing, with lags expanding from about 100 to 800 ms from the word onsets. Above 1 Hz, the predictability effects were left-hemisphere dominant. It was thus possible to noninvasively pick up neural signals that in different frequency bands were distinguishably sensitive to acoustic speech features, word-probability effects, and to local contextual information deployed in prospective, sequential information processing of natural continuous speech.

CRediT authorship contribution statement

Miika Koskinen: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing. **Mikko Kurimo:** Conceptualization, Methodology, Software, Writing - review & editing. **Joachim Gross:** Conceptualization, Methodology, Writing - review & editing. **Aapo Hyvärinen:** Conceptualization, Methodology, Writing - review & editing. **Riitta Hari:** Conceptualization, Supervision, Writing - review & editing.

Acknowledgements

Authors declare no conflict of interest. M. Koskinen was supported by Jane and Aatos Erkko Foundation and Aalto Brain Center (ABC), Aalto University, Finland. M. Kurimo was supported by Academy of Finland, Center of Excellence in Computational Inference (251170). J. Gross was supported by the Wellcome Trust (098433) and the DFG (GR 2024/5-1). A. Hyvärinen was supported by Academy of Finland, Centre-of-Excellence in Inverse Problems Research and Centre-of-Excellence Algorithmic Data Analysis. R. Hari was supported by the Finnish Cultural Foundation (Eminentia grant). M. Koskinen acknowledges discussions with Prof. Lauri Parkkonen.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.neuroimage.2020.116936.

References

Armeni, K., Willems, R.M., van der Bosch, A., Schoffelen, J.M., 2019. Frequency-specific brain dynamics related to prediction during language comprehension. Neuroimage 198, 283–295. https://doi.org/10.1016/j.neuroimage.2019.04.083.

Attneave, F., 1954. Some informational aspects of visual perception. Psychol. Rev. 61, 183–193.

Bourguignon, M., De Tiege, X., Op de Beeck, M., Ligot, N., Paquier, P., Van Bogaert, P., Goldman, S., Hari, R., Jousmäki, V., 2013. The pace of prosodic phrasing couples the reader's voice to the listener's cortex. Hum. Brain Mapp. 34, 314–326.

Brennan, J.R., Stabler, E.P., Van Wagenen, S.E., Luh, W.M., Hale, J.T., 2016. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. Brain Lang. 157–158, 81–94.

Brodbeck, C., Hong, E.L.E., Simon, J.Z., 2018. Rapid transformation from auditory to linguistic representations of continuous speech. Curr. Biol. 28, 3976–3983.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., Lalor, E.C., 2018. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr. Biol. 28, 803–809.

Chen, S., Goodman, J., 1999. An empirical study of smoothing techniques for language modeling. Comput. Speech Lang 13, 359–394.

Daube, C., Ince, R.A.A., Gross, J., 2019. Simple acoustic features can explain phoneme-based predictions of cortical responses to speech. Curr. Biol. 29, 1924–1937. https://doi.org/10.1016/j.cub.2019.04.067 e9.

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

Di Liberto, G.M., Wong, D., Melnik, G.A., de Cheveigné, A., 2019. Low-frequency cortical responses to natural speech reflect probabilistic phonotactics. Neuroimage 196, 237–247. https://doi.org/10.1016/j.neuroimage.2019.04.037.

Friston, K., 2002. Functional integration and inference in the brain. Prog. Neurobiol. 68, 113–143.

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

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., Garrod, S., 2013. Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol. 11, e1001752.

Hämäläinen, M., Ilmoniemi, R., 1994. Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. 32, 35–42.

Hamilton, L.S., Huth, A.G., 2018. The revolution will not be controlled: natural stimuli in speech neuroscience. Lang Cogn. Neurosci. https://doi.org/10.1080/23273798.2018.1499946.

Hari, R., Henriksson, L., Malinen, S., Parkkonen, L., 2015. Centrality of social interaction in human brain function. Perspective article. Neuron 88, 181–193. https://doi.org/10.1016/j.neuron.2015.09.022.

Hari, R., Parkkonen, L., 2015. The brain timewise: how timing shapes and supports brain function. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370, 20140170.

Hari, R., Parkkonen, L., Nangini, C., 2010. The brain in time: insights from neuromagnetic recordings. Ann. N. Y. Acad. Sci. 1191, 89–109.

Hasson, U., Yang, E., Vallines, I., Heeger, D., Rubin, N., 2008. A hierarchy of temporal receptive windows in human cortex. J. Neurosci. 28, 2539–2550.

Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J., 2006. Unlimited vocabulary speech recognition with morph language models applied to Finnish. Comput. Speech Lang 20, 515–541.

Hirsimäki, T., Pylkkönen, J., Kurimo, M., 2009. Importance of high-order N-gram models in morph-based speech recognition. IEEE Trans. Audio Speech Lang. Process. 17, 724–732.

Jääskeläinen, I.P., Ahveninen, J., Andermann, M.L., Belliveau, J.W., Raij, T., Sams, M., 2011. Short-term plasticity as a neural mechanism supporting memory and attentional functions. Brain Res. 1422, 66–81.

Kikuchi, Y., Attaheri, A., Wilson, B., Rhone, A.E., Nourski, K.V., Gander, P.E., Kovach, C.K., Kawasaki, H., Griffiths, T.D., Howard, M.A., Petkov, C.I., 2017. Sequence learning modulates neural responses and oscillatory coupling in human and monkey auditory cortex. PLoS Biol. 15, e2000219.

Klimovich-Gray, A., Tyler, L.K., Randall, B., Kocagoncu, E., Devereux, B., Marslen-Wilson, W.D., 2019. Balancing prediction and sensory input in speech comprehension: the spatiotemporal dynamics of word recognition in context. J. Neurosci. 39, 519–527.

Koskinen, M., Seppä, M., 2014. Uncovering cortical MEG responses to listened audiobook stories. Neuroimage 100, 263–270.

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.
Kösem, A., Bosker, H.R., Takashima, A., Meyer, A., Jensen, O., Hagoort, P., 2018. Neural entrainment determines the words we hear. Curr. Biol. 28, 2867–2875.
Koskinen, M., Kurimo, M., Hyvärinen, A., Hari, R., 2016. Predictive brain processing of listened audio narrative: MEG evidence, Poster Presentation. Biomag2016, Seoul, South Korea.
Kutas, M., Federmeier, K.D., 2011. Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annu. Rev. Psychol. 62, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123.
Kutas, M., Hillyard, S.A., 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307, 161–163.
Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., Schroeder, C.E., 2008. Entrainment of neuronal oscillations as a mechanism of attentional selection. Science 320, 110–113.
Lalor, E.C., Foxe, J.J., 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur. J. Neurosci. 31, 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x, 2010 Jan.
Lankinen, K., Saari, J., Hari, M., Koskinen, M., 2014. Intersubject consistency of cortical MEG signals during movie viewing. NeuroImage 92, 217–224. https://doi.org/10.1016/j.neuroimage.2014.02.004.
Lau, E.F., Phillips, C., Poeppel, D., 2008. A cortical network for semantics (de) constructing the N400. Nat. Rev. Neurosci. 9, 920–933.
Leonard, M.K., Bouchard, K.E., Tang, C., Chang, E.F., 2015. Dynamic encoding of speech sequence probability in human temporal cortex. J. Neurosci. 35, 7203–7214.
Lerner, Y., Honey, C., Silbert, L., Hasson, U., 2011. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. 31, 2906–2915.
Levinson, S.C., 2016. Turn-taking in human communication – origins and implications for language processing. Trends Cognit. Sci. 20, 6–14. https://doi.org/10.1016/j.tics.2015.10.010.
Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001–1010.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190.
Márton, C.D., Fukushima, M., Camalier, C.R., Schultz, S.R., Averbeck, B.B., 2019. Signature patterns for top-down and bottom-up information processing via cross-frequency coupling in macaque auditory cortex. eNeuro 6. https://doi.org/10.1523/ENEURO.0467-18.2019.
Mumford, D., 1992. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biol. Cybern. 66, 241–251.
Nelson, M.J., El Karoui, I., Giber, K., Yang, X., Cohen, L., Koopman, H., Cash, S.S., Naccache, L., Hale, J.T., Pallier, C., Dehaene, S., 2017. Neurophysiological dynamics of phrase-structure building during sentence processing. Proc. Natl. Acad. Sci. U.S.A. 114, E3669–E3678.
Niesler, T., Woodland, P., 1999. Variable-length category *n*-gram language models. Comput. Speech Lang 13, 99–124.
Pan, S., Kathleen, R.M., 1999. Word informativeness and automatic pitch accent modeling. Proc. EMNLP VLC'99, 148–157.
Park, H., Ince, R., Schyns, P.G., Thut, G., Gross, J., 2015. Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr. Biol. 25, 1649–1653.
Peelle, J., Davis, M., 2012. Neural oscillations carry speech rhythm through to comprehension. Front. Psychol. 3, 320.
Peelle, J.E., Gross, J., Davis, M.H., 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebr. Cortex 23, 1378–1387.
Rao, R., Ballard, D., 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat. Neurosci. 2, 79–87.
Rommers, J., Dickson, D.S., Norton, J.J., Wlotko, E.W., Federmeier, K.D., 2017. Alpha and theta band dynamics related to sentential constraint and word expectancy. Lang. Cogn. Neurosci. 32, 576–589.
Sassenhagen, J., Draschkow, D., 2019. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology 56 (6). https://doi.org/10.1111/psyp.13335 e13335.
Schultz, W., Dickinson, A., 2000. Neuronal coding of prediction errors. Annu. Rev. Neurosci. 23, 473–500.
Siivola, V., Creutz, M., Kurimo, M., 2007. Morfessor and VariKN machine learning tools for speech and language technology. In: Proceedings of the 8th International Conference on Speech Communication and Technology. INTERSPEECH'07, pp. 1549–1552.
Smith, N.J, Levy, R, 2011. Cloze but no cigar: the complex relationship between cloze, corpus, and subjective probabilities in language processing. In: Proceedings of the 33rd Annual Conference of the Cognitive Science Society, pp. 1637–1642.
Stephens, G., Honey, C., Hasson, U., 2013. A place for time: the spatiotemporal structure of neural dynamics during natural audition. J. Neurophysiol. 110, 2019–2026.
Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., et al., 2009. Universals and cultural variation in turn-taking in conversation. Proc. Natl. Acad. Sci. U.S.A. 106, 10587–10592.
Summerfield, C., Egner, T., Greene, M., Koechlin, E., Mangels, J., Hirsch, J., 2006. Predictive codes for forthcoming perception in the frontal cortex. Science 314, 1311–1314.
Taulu, S., Kajola, M., Simola, J., 2004. Suppression of interference and artifacts by the signal space separation method. Brain Topogr. 16, 269–275.
Torrence, C., Compo, G.A., 1998. Practical guide to wavelet analysis. Bull. Am. Meteorol. Soc. 79, 61–78.
Willems, R.M., Frank, S.L., Nijhof, A.D., Hagoort, P., van den Bosch, A., 2016. Prediction during natural language comprehension. Cerebr. Cortex 26, 2506–2516.