Cortical signatures of heard and imagined speech envelopes

Siyi Deng, Ramesh Srinivasan & Michael D'Zmura

Department of Cognitive Sciences, University of California at Irvine

SSPA 3151, Irvine, CA 92697-5100, USA

Short title: Signatures of Heard and Imagined Speech

E-mail: sdeng@uci.edu

Working Paper: August 2013

Abstract

We performed an experiment with both heard and imagined speech to determine whether cortical signatures of heard speech can be used to identify imagined speech. Each trial in the experiment presented one of six possible spoken sentences; it was both heard and, immediately afterwards, produced in imagination. The analysis focused on the use of envelope following responses (EFRs) to identify sentences. Source imaging methods were used to find the cortical origins of EFRs to heard speech. Reconstructing the EEG from the strongest sources of the EFRs in parietal and temporal cortex improved the correlation between EEG and the amplitude envelope of the heard speech. Single-trial classification performance was statistically significant for two of eight subjects. Significant classification performance was found for all subjects when one used EEG data from multiple trials of the same sentence, concatenated to produce data of greater duration. Activities at the cortical sources determined for heard speech were estimated from EEG data recorded while speech was imagined, in order to classify the imagined speech. Classification performance improves as the duration of EEG data increases; about seven trials of the same sentence are required for classification of the imagined sentence to reach statistical significance. These results suggest imagining speech engages some of the cortical populations involved in perceiving speech, as suggested by models of speech perception and production.

Keywords

speech, envelope, EEG, source localization, classification

Introduction

When we listen to continuous natural speech, a cortical response which is phase-locked to the amplitude envelope of the speech can be recorded with EEG electrodes over temporal and parietal regions. This envelope following response (EFR) is believed to provide a direct measurement of auditory speech processing system activity (Lalor and Foxe 2010). The amplitude envelope of speech provides important information for comprehension (Shannon et al. 1995; Ahissar et al. 2001), bears information concerning syllable and word boundaries, and presents prosodic cues (Rosen 1992).

Recent studies show that analysis of EFRs can help to identify the sentences to which one is listening (Koskinen et al. 2012) or attending to in dichotic listening (Horton et al. 2013). These perceptual processes in speech perception may be intrinsically linked to motor processes in speech production, as suggested by models of speech perception and production (Hickok and Poeppel 2007). It is believed that the auditory imagery is linked to feedback (*efference copies*) used in speech motor planning (Tian and Poeppel 2011, 2013). This suggests that it may be possible to use the EFRs to identify imagined speech, and that these cortical signals would originate in the same areas as the EFRs to heard speech.

In a previous study of imagined speech, we showed that variations in rhythmic sequences of imagined phonemes can be discriminated from one another using EEG (Deng et al. 2010). However, even if EFRs are elicited when one imagines speech, the unknown onset and rate of production associated with imagined speech may make the EFR signal much harder to detect. Such temporal uncertainties in imagined speech production call for time-independent techniques to boost the signal-to-noise ratio of the potential responses of interest. Component analysis methods such as PCA and ICA are widely used to enhance spatial features and to remove artifacts (Delorme et al. 2007). Source imaging methods can also improve the signal-to-noise ratio of EEG by restricting the signals to the strongest cortical sources. Qin and colleagues (2005) have shown that a source analysis paradigm can help to separate signals generated by cortical regions of interest from background activity and to improve performance in motor imagery tasks. Similarly it has been shown that source analysis methods can improve the accuracy of brain-computer interfaces (Wentrup et al. 2005; He et al. 2013). Following these approaches, we performed a time course reconstruction of the activity at the locations of the strongest dipole sources that contribute to the EFR to heard speech and used the activity localized to these locations to classify both heard and imagined speech.

Materials and methods

Data recording and preprocessing

Six sentences were selected from the DARPA TIMIT corpus (Fisher et al. 1986). The sentences were chosen from among the 450 possible TIMIT SX sentences in a way that minimized the interclass correlations between sentence power envelopes. The sentences were: 1) *Steve collects rare and novel coins*; 2) *Herb's birthday occurs frequently on Thanksgiving*; 3) *Jane may earn more money by working hard*; 4) *The eastern coast is a place for pure pleasure and excitement*; 5) *Rock and roll music has a great rhythm*, and 6) *The government sought authorization of his citizenship*.

Each trial in the experiment lasted 10.5 sec. A 0.25-sec beep at trial onset was followed by a 5.0-sec heard cue period during which one of the six selected sentences was presented to the subject. Another 0.25-sec beep following the heard cue signaled to the subject the onset of the 5.0-sec imagine cue period during which the subject was asked to repeat the heard sentence in imagination. Each subject heard the cued sentences in his or her own voice. This was made possible by recording high-quality audio waveforms at 44.1KHz of the six selected sentences in each subject's own voice prior to the experiment. During the EEG experiment, subjects sat in a dimly-lit room with eyes closed, and the auditory stimuli were presented binaurally using a set of electrostatic earphones (STAX SR001-MK2) which interfere minimally with EEG recording.

Each of the six sentences appeared in 89 trials which were presented in a block-randomized fashion within the total sequence of 534 trials. The entire experiment lasted about 1.5 hours. EEG data were recorded at 1024 Hz during the entire time using a 128-channel Geodesic Sensor Net (EGI) with an Advanced Neural Technology amplifier. Of the 128 channels, four were identified as bad and excluded from further analyses. A separate channel on the stimulus computer's sound card was used to record the audio stimuli as a marker in the EEG recording to facilitate data segmentation.

After acquisition, the EEG was band-pass filtered offline to 1-50 Hz using a Butterworth filter. EEG from each trial was then segmented into two 5.0-sec long pieces using the audio marker data. One EEG segment was drawn from the heard cue period. We refer to these heard speech EEG data as "heard EEG" in what follows. The other "imagined EEG" segment was drawn from the 5.0-sec imagine cue period.

The speech envelope was computed, for each of the six selected sentences spoken by each subject, by band-pass filtering the Hilbert modulus of the waveform to 1-30 Hz. The envelopes were downsampled from 44100 to 1024 Hz to match the EEG sampling rate. Relatively silent periods at the beginning and end of each speech envelope were manually identified and removed from further consideration.

Analysis proceeded by randomly partitioning the trials for each of the six stimulus sentences into a training set and a testing set. This partitioning was performed to prevent circular inference and to facilitate the classification procedures described in the next few sections. The training set had a total of 270 trials (45 trials per sentence). The testing set had a total of 264 trials (44 trials per sentence) and was used only for testing classification performance. All other analyses (ICA, correlation, and source localization) were performed only on the training data.

Seven native English-speaking subjects participated in this study. All were right-handed males in the age range 21-26; all reported normal hearing.

Independent Component Analysis (ICA)

Pilot analyses as well as the previous literature (Delorme et al. 2007; Mantini et al. 2007; Horton et al. 2013) suggest that ICA methods can help to separate genuine EEG from artifacts and to improve the signal-to-noise ratio of the component of interest. For each subject's heard EEG, we first used principal component analysis (PCA) to reduce data dimensionality. The 24 components with the largest eigenvalues were retained. They account for more than 92% of the variance for each subject. These components for the heard EEG were then transformed into 24 independent components (ICs) using the information-maximization ICA algorithm (Bell and Sejnowski 1995). In matrix notation:

$$\mathrm{x} = \mathrm{Ac} \qquad \text{Eq. 1}$$

As described in the equation, the assumption of ICA is that at each time instant, the observed EEG x is the result of a linear mixing of independent components c . The underlying "mixing" matrix A is estimated by ICA algorithms, and its pseudo-inverse gives the "separation" matrix A^{+}, which is applied to the EEG data x to estimate the independent components c. In this study, the separation matrix A^{+} generated from the training data was used to transform the testing data into ICs of the same basis.

Identification of the EFR in heard EEG

We computed the cross-correlation between the 24 IC time courses and each of the six speech envelopes in order to identify EFRs in the EEG segment where speech was heard. Only the training data were used in this computation. The values were normalized so that the auto-correlation at time lag zero had the value of one. The cross-correlation was computed for time lags between -1000 to 1000 msec, although later analysis focuses on results between 0 and 500 msec (Aiken et al. 2008). We categorized the results into matched and mismatched conditions based on whether the cross-correlation was computed between a heard EEG component and the envelope of the sentence actually heard ("matched") or between a heard EEG component and the envelope of some sentence other than the heard one ("mismatched"). The statistical significance of matched versus mismatched condition was tested at each time lag using a one-way ANOVA on the null hypothesis of equal means, with a Bonferroni correction for multiple comparisons across time lags and components. The subset of ICs that was statistically significant was used for further analysis:

$$\mathrm{x}_s = \mathrm{A}_s\mathrm{c}_\mathrm{s} \qquad \text{Eq. 2}$$

In Eq. 2, c_s is the significant subset of ICs, A_s is a modified mixing matrix with columns corresponding to significant ICs, and x_s is the remixed EEG resulting from significant ICs.

Source-reconstructed time course of the EFR in heard EEG

Source imaging techniques were applied to those ICs which were shown to be significantly correlated with speech envelopes in order to localize their cortical origins. Two to four significant ICs were found for each subject. Because some subjects did not consent to MRI scans, we used the MNI-152 Brain Atlas, a linear average of 152 T1 weighted brain scans of young adults created by the Montreal Neurological Institute (Fonov et al. 2009), to construct the volume conduction model that was used for all subjects.

The MNI brain was first manually segmented into brain, skull, and scalp regions using region-growth techniques (BrainVoyager™). Surface meshes for the brain-skull, skull-scalp, and scalp-air boundaries were then generated based on the segmentation results. 4761 dipole triplets (unrestricted dipole orientation is specified by independent x, y, and z moments) were placed on the cortical surface. The corresponding dipole forward matrix was computed using symmetric boundary element methods (OPENMEEG package, Kybic et al. 2005; Gramfort et al. 2010). Minimum L2 norm estimations of the inverse solutions (MNE) for the significant ICs were then computed with Tikhonov regularization. The forward model can be written as:

$$\mathrm{x}_s = \mathrm{Kd} + \mathrm{n} \qquad \text{Eq. 3}$$

In Eq. 3, K is the source forward matrix computed using BEM, d represents the underlying source dipole magnitudes, and n represents zero-mean Gaussian noise. The MNE solution is computed as:

$$\mathrm{d} = \mathrm{K}^{\mathrm{T}}\left(\mathrm{K}\mathrm{K}^{\mathrm{T}} + \alpha\mathrm{I}\right)^{-1}\mathrm{x}_s = \mathrm{G}\mathrm{x}_s = \mathrm{G}\mathrm{A}_s\mathrm{c}_s = \mathrm{H}\mathrm{c}_s \qquad \text{Eq. 4}$$

In Eq. 4, G is the inverse matrix, α is the regularization parameter estimated using the L-curve method (Hansen and O'Leary 1993), and $\mathrm{H} = \mathrm{G}\mathrm{A}_s$ represents the source contribution to each significant IC. Upon obtaining the inverse result, we manually identified several source dipoles at the local maxima of each solution, and combined the contribution of selected dipoles to get H_m. Thereafter, a single source time course c_m can be reconstructed from the selected dipoles as:

$$\mathrm{c}_m = (\mathrm{K}_m\mathrm{H}_m)^{+}\mathrm{x} \qquad \text{Eq. 5}$$

In Eq. 5, K_m is modified forward matrixes with columns correspond to the selected dipoles, and $(\mathrm{K}_m\mathrm{H}_m)^{+}$ represents the reconstruction matrix for the selected dipole sources.

The heard EEG data were then multiplied by the reconstruction matrix to generate the time course from selected dipoles. Because the source-reconstructed time course involves only representative dipoles at regions with maximal responses to the speech stimuli, one can expect such reconstruction to enhance the signal-to-noise ratio by leaving out regions of low signal-to-noise ratio. Again, the source localization and reconstruction analyses were based on the training data set alone.

We then computed the cross-correlations between speech envelopes and the source-reconstructed time courses. We also determined the maximally-correlated EEG channel and its cross-correlation in order to compare with source-reconstruction results

Classification of heard EEG

The responses of EEG components which are more correlated to matching speech envelopes than to mismatched ones can be used to classify trials according to the sentence heard during the trial. This approach is conceptually equivalent to constructing a linear filter using the pre-determined speech envelopes and time delays. Identification works by determining which sentence's speech envelope generates the greatest cross-correlation among the responsive EEG components. In practice, the maximum cross-correlation values within a response delay window of 0-500 msec are used to determine the matching speech envelope.

Classification performance was determined by applying the above procedure to data from 1) the single best EEG channel; 2) all EEG channels; 3) the single best ICA component, and 4) source-reconstructed time courses. For each of these four possibilities, the training-testing partitioning was randomized 500 times to generate a distribution of classification rates.

We also performed this analysis using progressively longer pieces of EEG data. One expects that testing data of longer duration strengthens the signal-to-noise ratio and so enhances classification performance. EEG data segments of longer duration were created by concatenating two or more single-trial segments for the same sentence class. These segments were picked randomly using a bootstrap resampling scheme which made use of all 44 single-trial testing segments. This bootstrap scheme ensures that 44 trials were used for classification when same-sentence segments were concatenated. Again, the training-testing partitioning was randomized 500 times to create classification rate distributions. The α

= 0.01 significance level of 0.206 for classification performance were determined by using the cumulative binomial distribution to estimate the likelihood of observing a given classification rate by chance, which is 1/6 in this experiment with six sentences.

Classification of imagined EEG

We used the dipole locations for heard EEG data to generate the source-reconstructed time courses for imagined EEG. The latter were determined by multiplying the imagined EEG data with the source reconstruction matrix described in the earlier section on source analysis. Classification of trials according to imagined sentence was performed by applying the procedures described in the previous section to these imagined EEG source-reconstructed time courses. We used the amplitude envelopes of corresponding heard sentences. We feel it reasonable to assume that the auditory imagery of the sentence produced in imagination has loudness properties which match closely the amplitude envelopes measured for the corresponding heard sentence. Classification was also performed on progressively longer pieces of imagined data by concatenating single trials, as described in the previous section.

Results

We show that source reconstruction methods, when applied to ICs of heard speech, let one use EFRs to identify heard sentences. Furthermore, classification of the imagined EEG time course reconstructed using the source forward matrix generated from heard EEG data reaches significance if data of sufficient duration are provided.

Identification of EFRs using ICA of heard speech EEG

As described earlier in Methods, we computed the cross-correlations between the speech envelopes of spoken sentences and the ICs of heard EEG. Figure 1 shows results for a representative subject. The left panel shows the cross-correlations between 24 ICs of heard EEG and the speech envelope of the sentence heard while the EEG was recorded ("matched"). The right panel shows the correlations between the same 24 ICs of heard EEG and the speech envelope of sentences which were chosen randomly from among the five sentences *not* heard while EEG was recorded ("mismatched"). A one-way ANOVA on the assumption of equal means between an IC of matched and mismatched conditions in the time window of 0-500 msec suggests that there are statistically significant effects for two of the ICs found for this particular subject. These significant ICs are shown in the left panel in solid (IC 23) and dashed (IC 20) lines. These two components remain significant after multiple-test corrections using the Bonferroni criterion.

These significant ICs were determined for all eight subjects. The time delays of maximum correlation were found at about 140 msec (for all subjects, μ=137 msec, σ=11 msec), which match well the values found in earlier EEG (Lalor and Foxe 2010) and MEG (Koskinen et al. 2012) studies. The significant ICs let one identify EFRs for each subject. As illustrated in Figure 2, the response time-course of the most

significant IC (for this subject, IC 23) found when listening to a particular sentence follows the loudness envelope of that sentence.

Source time course further improves correlation

The significant ICs correspond to particular linear combinations of EEG electrodes across the scalp. Such topographies are illustrated for the same representative subject for two ICs in Figure 3 (top row). We use source localization techniques, described in Methods, to find the origins of these ICs on the cortical surface (Fig. 3, bottom row). By choosing dipoles which lie at local maxima of these source images, we emphasize the strongest sources of EFRs and, simultaneously, de-emphasize relatively non-informative EEG activity. In practice, we find that source reconstructed scalp potential time courses which correspond to these selected dipoles show increased cross-correlation with sentence envelopes. Figure 4 compares cross-correlations found between sentence envelopes and the time courses of 1) the best single EEG channel (dashed line); 2) the best IC (dotted line), and 3) the reconstructed sources (solid line). The greatest cross-correlation value is found when using source-reconstructed time courses. This improvement is consistent among all subjects.

Classification of heard sentences using EEG

Classification of trials according to the sentence heard was performed using four types of EEG response. As shown in Figure 5, classification performance found for the representative subject when using reconstructed source responses (rightmost box-and-whisker) surpassed the performances found when using the single best EEG channel, all EEG channels, or the single best IC. Furthermore, the median classification performance found for the distribution determined through 500 random training-testing partitioning of the data exceeded the $\alpha = 0.01$ significance level of 0.206.

Classification performance improves as more EEG data are provided. Figure 6 shows how classification becomes more accurate as the duration of EEG data increases. A single trial of data has a duration of about 2.8 sec, and provides the leftmost data point in this plot. As further trials are concatenated to produce simulated trials of greater duration, classification performance improves from its initial level of 0.23 to about 0.75 when as many as 44 testing trials are concatenated.

Classification of heard EEG as a function of duration is shown for all subjects in Figure 7. Each continuous curve shows classification rates for a single subject. The classification performance found for single trials is significant for two of eight subjects when using a significance level of 0.206 ($\alpha = 0.01$). Performance improves rapidly for all but one subject (bottommost curve), so that classification performance for seven of eight subjects is significant when three or more trials (about 8 sec) of data are provided. Peak classification performance for these seven subjects ranges between 0.46 and 0.75. These results show that given sufficient data, the EFR provided by source reconstruction can be used to successfully classify the heard speech sentences from EEG.

Classification of imagined speech reaches significance with sufficient data

The significant ICs found for heard EEG and used to generate source-reconstructed time courses can also be applied to the classification of imagined speech. Figure 8 shows for the representative subject the cross-correlations between EEG recorded while imagining speech (imagined EEG) and the time courses reconstructed using source information derived from heard speech (dotted line) and single best channel of imagined EEG (solid line). The correlations are not as pronounced as those found for heard speech responses (shown earlier in Figure 4). The EFR is not significant in raw EEG and is borderline significant in the case of source-reconstructed time courses of imagined speech.

The weak EFR underlies the poorer classification performance found for imagined EEG shown in Figure 9. As was shown for heard EEG, classification performance for imagined EEG improves as more EEG data are provided. Fully seven trials of data (of duration approximately 20 sec) are required for the significance level of 0.206 ($\alpha = 0.01$) to be reached for the best performing subject. Furthermore, classification performance reaches a maximum of about 0.30 when using all available data.

Classification of imagined EEG as a function of duration is shown for all subjects in Figure 10. Again, each continuous curve shows classification performance for a single subject in this plot. Subjects need between 7-11 trials of EEG data to reach a significant level of performance. Peak performances lie between 0.28 – 0.31 for all subjects. The subject with the lowest heard EEG classification performance (lowest curve in Fig. 7) does not differ strongly from the remaining subjects when imagined EEG data are analyzed. In summary, the results show that the EFR provided by source-reconstructed time courses using information derived from heard EEG can be used to identify the imagined speech sentences, provided that sufficient data are available.

Discussion

Results of the present study show that EEG recorded while one listens to natural speech contains traces of the speech amplitude envelope. Independent component analysis helps to extract this EFR signal from raw EEG. A source localization and reconstruction procedure which uses the ICA results further enhances the strength of this signal. We show that this source-reconstructed EFR can be used to successfully identify heard sentences. In addition, the source analysis results found for the heard data help one to reconstruct EFRs for imagined speech, and classification performance based on these EFRs reaches significance if sufficient data are available.

EEG has long been known to convey information concerning the responses to the envelopes of heard auditory stimuli (Dolphin and Mountain 1991; Dolphin 1997; Purcell et al. 2004), including speech streams (Aiken and Picton 2006, 2008). These EFR studies have tended to focus on the nature of the EFR as an evoked potential rather than on its use to identify speech from EEG. EEG studies which have focused on the identification of speech include early work by Suppes and colleagues (1997, 1998), who recorded EEG signals while subjects listened to or imagined small sets of words or sentences. Their classification results had limited success. More recently, Deng and colleagues (2010) showed that the rhythmic structure in speech like stimuli can be classified from both heard and imagined EEG data using time-frequency features based on Hilbert spectra analysis.

Magnetoencephalography (MEG) has also been used to study speech perception and its identification. Evidence in MEG has been found for cortical monitoring of both overt and covert speech production (Numminen and Curio 1999; Houde et al. 2002). The temporal envelopes of MEG responses in the theta band have been related to the syllabic structure of the heard sentences (Luo and Poeppel 2007). Envelope-correlated components were found by Ahissar and colleagues (2001) in MEG data recorded with natural speech stimuli using principal component analysis. Most recently, Koskinen and colleagues (2012) have used MEG EFRs to help identify the sentences to which one is listening.

In the present study, we use ICA to reduce data dimensionality and to create spatial filters in order to isolate brain signals relevant to speech perception. It is known that ICA algorithms are blind to the neurophysiological nature of the components generated (James and Hesse 2005). While significant ICs can be more strongly correlated with speech envelopes than are raw EEG time courses, and therefore can improve classification performance, the signal may be far from optimal. One possible way to further improve the quality of EFRs is to use *a priori* knowledge of the brain activity which underlies speech processing to guide the ICA algorithms (Chen et al. 2008). It has been shown that a reference signal can be used to constrain ICA algorithms to extract those components which approximate the reference time course (Lu and Rajapakse 2000). Spatial priors can also be incorporated in ICA algorithms so that brain patterns of interest will be emphasized in one or more output components (Hyvarinen et al. 2001).

The source localization and reconstruction procedure used in this paper offers an *a posteriori* filtering of the ICA results. The selection of dipoles deemed significant from the local maxima of the source images is subjective. Although this procedure improves classification performance beyond that found using just the ICA, the heuristic nature of this procedure seldom provides the optimal configuration of source dipoles for EFR extraction in classification tasks.

The strength of EFRs found in this EEG study seem unlikely to support the function of a brain-computer interface (BCI) which works in real time to identify heard or imagined speech streams. Volume conduction effects limit signal strength in EEG recordings. One alternative to EEG is the use of invasive methods. ECoG has proved to be very effective in experiments decoding using both heard (Nourski et al. 2009; Pasley et al. 2012; Kubanek et al. 2013) and imagined (Leuthardt et al. 2011) speech. Yet if non-invasive methods are preferred, our results suggest one may employ source imaging methods to improve the spatial resolution and to reduce volume conduction effects of EEG recording, as exemplified in the current study.

We applied data transformations to imagined EEG data which were derived from analysis of heard EEG data. While there is no reason to think that the procedure so constructed is optimal, we nevertheless were able to classify imagined speech given EEG data of sufficient duration. This is likely to have worked because auditory perceptual and imagery activity share a number of similarities. For example, the results of PET (e.g., Zatorre et al. 1996) and fMRI (e.g., Yoo et al. 2001) studies suggest that auditory perceptual and imagery processes share certain degree of functional and neurophysiological similarities. Zatorre and colleagues (1996) compared the neural substrate of listening and imagining songs and found that both produces similar patterns involving secondary auditory cortices, frontal-parietal lobes and supplementary motor area. These cortical areas are also activated in auditory verbal imagery experiments (Shergill et al. 2001). It is believed that the auditory imagery is linked to feedback (*efference copies*) used in speech motor planning (Hickok et al. 2011; Tian and Poeppel 2011, 2013). The localization results of our study indicate that posterior parietal region as well as the secondary auditory

cortices are critical to the phase-locking activity during speech perception. In addition, the fact that activities reconstructed from these cortical areas during speech imagination carry positive, albeit weak information about corresponding mental processes provides further evidence in favor of the hypothesis that shared neural mechanisms mediate speech perception and imagination.

The current experiment uses EEG to discriminate among a fixed set of six sentences with distinct amplitude envelopes. Classification based on EFRs is restricted only by the signal-to-noise ratio, so that with enough signal gain it should be possible to classify speech patterns of greater complexity. Should phonetic or syllabic tokens be reliably identified from one's brain activity due to increases in signal gain, an EEG-based imagined speech communication system can be developed. We believe that EFR offers a stepstone to decode verbal thoughts from EEG signals.

Acknowledgements

This work was supported by ARO 54228-LS-MUR.

References

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences* 98: 13367-13372, 2001.

Aiken SJ, Picton TW. Envelope following responses to natural vowels. *Audiology and Neurotology* 11: 213-232, 2006.

Aiken SJ, Picton TW. Human cortical responses to the speech envelope. *Ear and hearing* 29: 139-157, 2008.

Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. *Neural Comput* 7: 1129–1159, 1995.

Chen L, Xuan J, Wang C, Shih IM, Wang Y, Zhang Z, Clarke R. Knowledge-guided multi-scale independent component analysis for biomarker identification. *BMC bioinformatics* 9: 416, 2008.

Delorme A, Sejnowski T, Makeig S. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage* 34: 1443-1449, 2007.

Deng S, Srinivasan R, Lappas T, D'Zmura M. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of neural engineering* 7: 046006, 2010.

Dolphin WF. The envelope following response to multiple tone pair stimuli. *Hearing research* 110: 1-14, 1997.

Dolphin WF, Mountain DC. Scalp potentials follow the low frequency envelope of complex acoustic stimuli. *Proceedings of the IEEE Seventeenth Annual Northeast* 1991:215-216, 1991.

Fisher WM, Doddington GR, Goudie-Marshall KM. The DARPA Speech Recognition Research Database: Specifications and Status. *Proceedings of DARPA Workshop on Speech Recognition* 1986: 93–99, 1986.

Fonov VS, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, BDCG. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage* 54: 1053–8119, 2011.

Gramfort A, Papadopoulo T, Olivi E, Clerc M. OpenMEEG: opensource software for quasistatic bioelectromagnetics. *BioMedical Engineering OnLine* 45: 9, 2010.

Hansen PC, O'Leary DP. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* 14: 1487-1503, 1993.

He B, Gao S, Yuan H, Wolpaw JR. Brain–Computer Interfaces. *Neural Engineering*. Springer US. 2013

Hickok G, Houde J, Rong F. Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron* 69: 407-422, 2011.

Hickok G, Poeppel D. The cortical organization of speech processing. *Nature Reviews Neuroscience* 8: 393-402, 2007.

Horton Ct, D'Zmura M, Srinivasan R. Suppression of competing speech through entrainment of cortical oscillations. *Journal of neurophysiology* 109: 3082-3093, 2013.

Houde JF, Nagarajan SS, Sekihara K, Merzenich MM. Modulation of the auditory cortex during speech: an MEG study. *Journal of Cognitive Neuroscience* 15: 1125-1138. 2002.

Hyvärinen A. Complexity pursuit: separating interesting components from time series. *Neural Computation* 13: 883-898, 2001.

James CJ, Hesse CW. Independent component analysis for biomedical signals. *Physiological measurement* 26, R15, 2005.

Koskinen M, Viinikanoja J, Kurimo M, Klami A, Kaski S, Hari R. Identifying fragments of natural speech from the listener's MEG signals. *Hum. Brain Mapp.* 34: 1477–1489, 2012.

Kubanek J, Brunner P, Gunduz A, Poeppel D, Schalk G. The tracking of speech envelope in the human cortex. *PLOS One* 8: 1-9, 2013

Kybic J, Clerc M, Abboud T, Faugeras O, Keriven R, Papadopoulo T. A common formalism for the integral formulations of the forward EEG problem. *IEEE Transactions on Medical Imaging* 24:12-28, 2005.

Lalor EC, Foxe JJ. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience* 31: 189–193, 2010.

Leuthardt EC, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J, Schalk G. Using the electrocorticographic speech network to control a brain-computer interface in humans. *J. Neural Eng.* 8: 1-11. 2011.

Lu W, Rajapakse JC. ICA with reference. *Neurocomputing* 69: 2244-2257, 2006.

Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54: 1001-1010, 2007.

Mantini D, Perrucci MG, Cugini S, Ferretti A, Romani GL, Del Gratta C. Complete artifact removal for EEG recorded during continuous fMRI using independent component analysis. *Neuroimage* 34: 598-607, 2007.

Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Brugge JF. Temporal envelope of time-compressed speech represented in the human auditory cortex. *The Journal of Neuroscience* 29: 15564-15574, 2009.

Numminen J, Curio G. Differential effects of overt, covert and replayed speech on vowel-evoked responses of the human auditory cortex. *Neuroscience Letters* 272: 29-32, 1999.

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF. Reconstructing speech from human auditory cortex. *PLOS Biology* 10: 1-13, 2012.

Purcell DW, John SM, Schneider BA, Picton TW. Human temporal auditory acuity as assessed by envelope following responses. *J. Acoust. Soc. Am.* 116: 3581-3593, 2004.

Qin L, Ding L, He B. Motor imagery classification by means of source analysis for brain-computer interface applications. *Journal of Neural Engineering* 1: 135–141, 2004.

Rosen S. Temporal information in speech: acoustic, auditory and linguistic aspects. *Phil. Trans. R. Soc. Lond. B* 336: 367-373, 1992.

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M. Speech recognition with primarily temporal cues. *Science* 270: 303-304, 1995.

Shergill SS, Bullmore ET, Brammer MJ, Williams SCR, Murray RM, McGuire PK. A functional study of auditory verbal imagery. *Psychological Medicine* 31: 241-253, 2001.

Suppes P, Lu ZL, Han B. Brain wave recognition of words. *Proceedings of the National Academy of Science USA* 94: 14965-14969, 1997.

Suppes P, Han B, Lu ZL. Brain wave recognition of sentences. *Proceedings of the National Academy of Science USA* 95: 15861-15866, 1998.

Tian X, Poeppel D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology: Auditory Cognitive Neuroscience* 10:1-23, 2011.

Tian X, Poeppel D. The effect of imagination on stimulation: the functional specificity of efference copies in speech processing. *Journal of Cognitive Neuroscience* 2013.

Wentrup MG, Gramann K, Wascher E, Buss M. Eeg source localization for brain-computer-interfaces. *Proceedings. 2nd International IEEE EMBS Conference* 2005: 128-131, 2005

Yoo SS, Lee CU, Choi BG. Human brain mapping of auditory imagery: event-related functional MRI study. *Neuroreport* 12: 3045-3049, 2001.

Zatorre RJ, Halpern AR. Mental concerts: musical imagery and auditory cortex. *Neuron* 47: 9-12, 2005

Figures

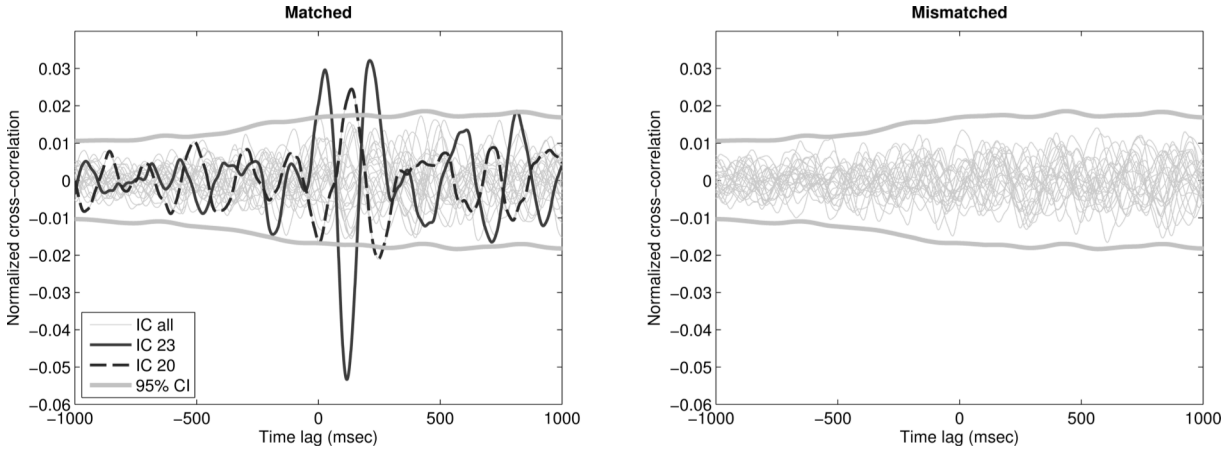

Figure 1. Time-lagged cross-correlation of heard speech EEG with matched and mismatched speech envelopes for a single subject. Left: result of matched cases, averaged across all 89 trials and six sentences. ICA components 23 and 20 have increased correlations at around 140 msec. Right: mismatched cases, averaged across 89 trials and six sentences. For each trial, a random mismatched envelope is used to compute the cross-correlation. The 95% Bonferroni-corrected confidence interval is calculated at each time frame from the distribution found by cross-correlating all six sentence envelopes, whether matched or not. This interval generates the two gray curves in each panel which span the entire time lag interval.

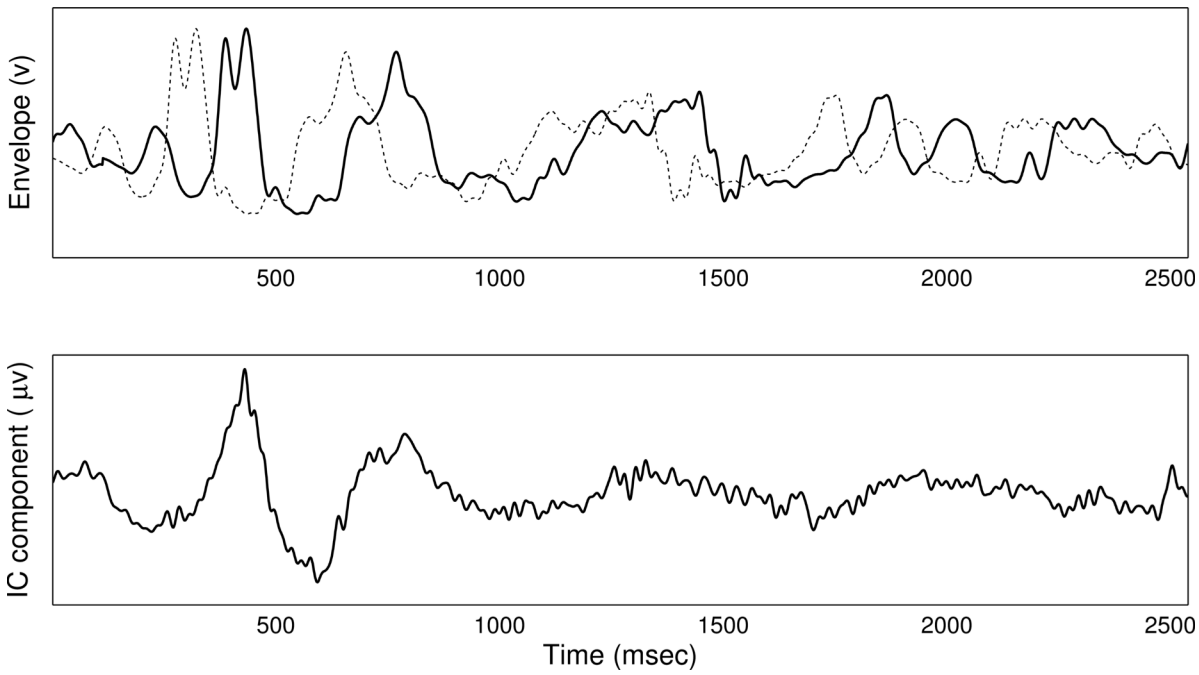

Figure 2. An example of the speech envelope correlated signal in EEG. Top panel: the speech envelope of one sentence ("Steve collects rare and novel coins") for the same subject as in Fig. 1. The light gray curve shows the original envelope, while the black curve is the envelope delayed by 120 msec. Bottom panel: the time course of the 23rd ICA component of the same subject. The correlation with the time-delayed speech envelope has the value 0.739.

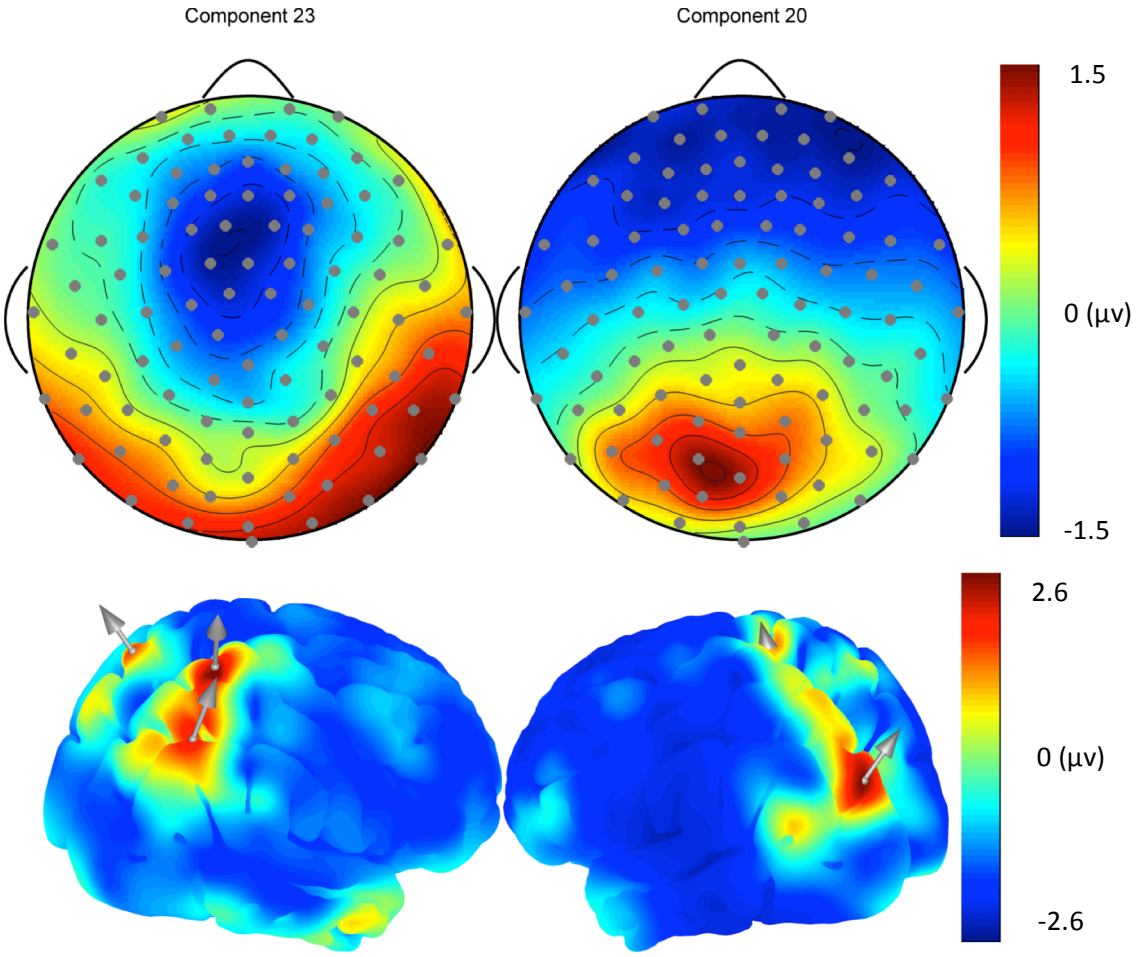

Figure 3. Source localization and reconstruction procedure for the same subject as in earlier Figures. Top row: the topographic weights of ICA component 23 (left) and 20 (right). Bottom row: source localization results for the respective components. The arrows represent the dipoles selected from the local maxima of the localization results for scalp potential time course reconstruction.

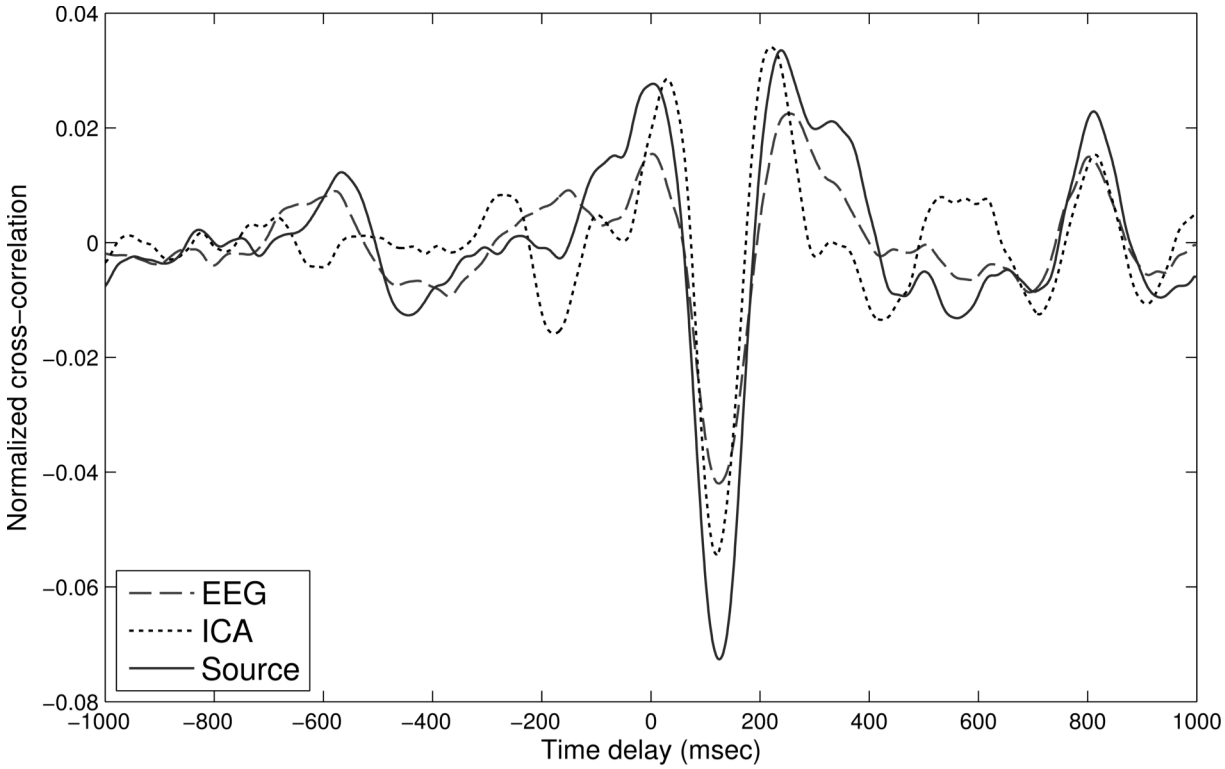

Figure 4. Comparison of the speech envelope cross-correlation found for the best EEG channel (dashed curve), best independent component (dotted curve) and the source-reconstructed time course (solid curve) for the representative subject. The dashed curve is the cross-correlation plot of the EEG channel with the maximum correlation with matched speech envelopes. The dotted curve is the cross-correlation plot found using the best independent component rather than the best EEG channel. The solid curve is the cross-correlation plot found for the source-reconstructed time course. The three plots show that the source-reconstructed time course has the highest correlation with matched spoken sentence envelopes. Each curve is the average of all 534 trials available (89 trials per sentence with six sentences).

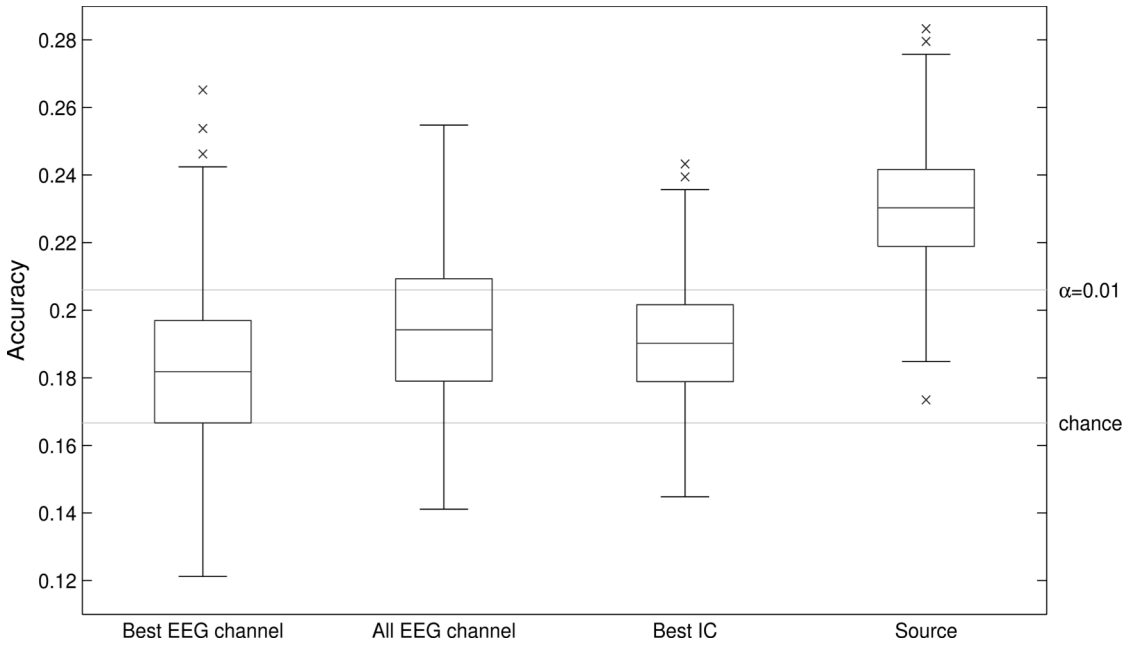

Figure 5. Classification performance using the best EEG channel (leftmost), the result of multivariate linear classification using all EEG channels (second from left), classification using the best independent component (third from left), and classification using the source-reconstructed time course (rightmost) for the single subject used in earlier Figures. Each box-and-whisker plot was generated from 500 random partitions for training-testing sets. Each box represents the 25th, 50th and 75th percentiles; the whiskers indicate data range. The x symbols mark outlier data points.

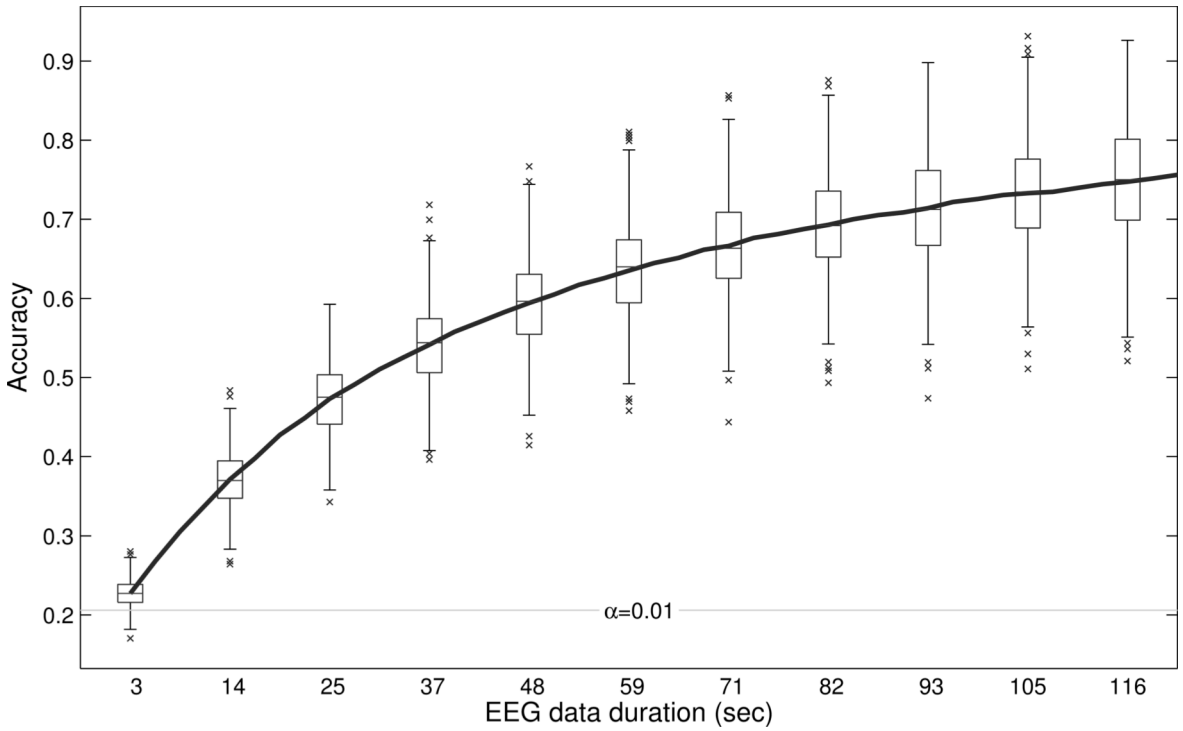

Figure 6. The classification performance found with increasingly longer durations of heard data for the single subject used in earlier Figures. Measurements for trials using the same sentence were concatenated to produce segments of longer duration. The black curve shows the median classification accuracy. The box-and-whisker plots are shown for every fourth concatenated trial. Each box-and-whisker plot was generated from 500 random partitions of training-testing sets. The boxes represent the 25th, 50th and 75th percentiles and the whiskers the extent of the data. The x symbols indicate outliers.

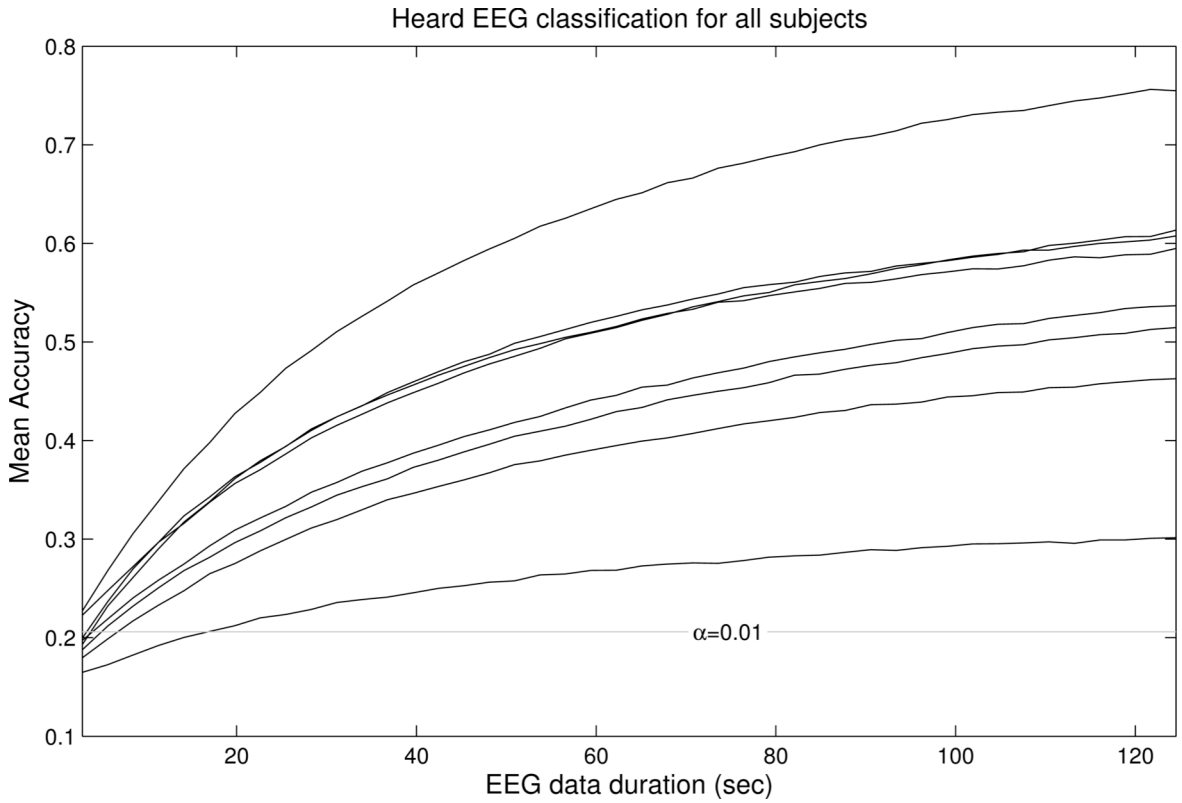

Figure 7. Classification performance as a function of heard EEG data duration for each of the eight subjects. Each curve shows the results for one subject. Measurements for trials using the same sentence were concatenated to produce segments of longer duration. Each data point on every curve represents the median performance from a distribution of 500 random partitions of training-testing sets for that subject. The mean classification values at the longest duration of 124.5 seconds, ranking for highest to lowest, are: 0.7548, 0.6134, 0.6076, 0.5948, 0.5367, 0.5146, 0.4627, and 0.3014.

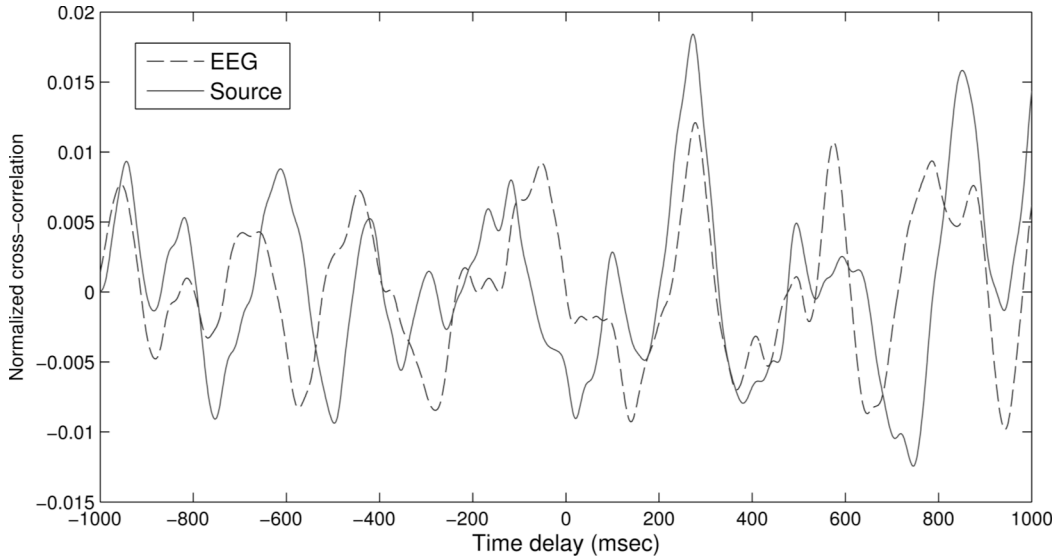

Figure 8. Comparison of the imagined speech envelope cross-correlation found for the best EEG channel (dashed curve) and for the source-reconstructed time course (solid curve) for the single subject used in the earlier Figures. The best EEG channel was taken as that with the maximum correlation with matched speech envelopes. The highest correlation is found for the source-reconstructed time course cross-correlation at a lag of about 280 msec. Each curve is the average of all 534 trials available (89 trials per sentences and six sentences).

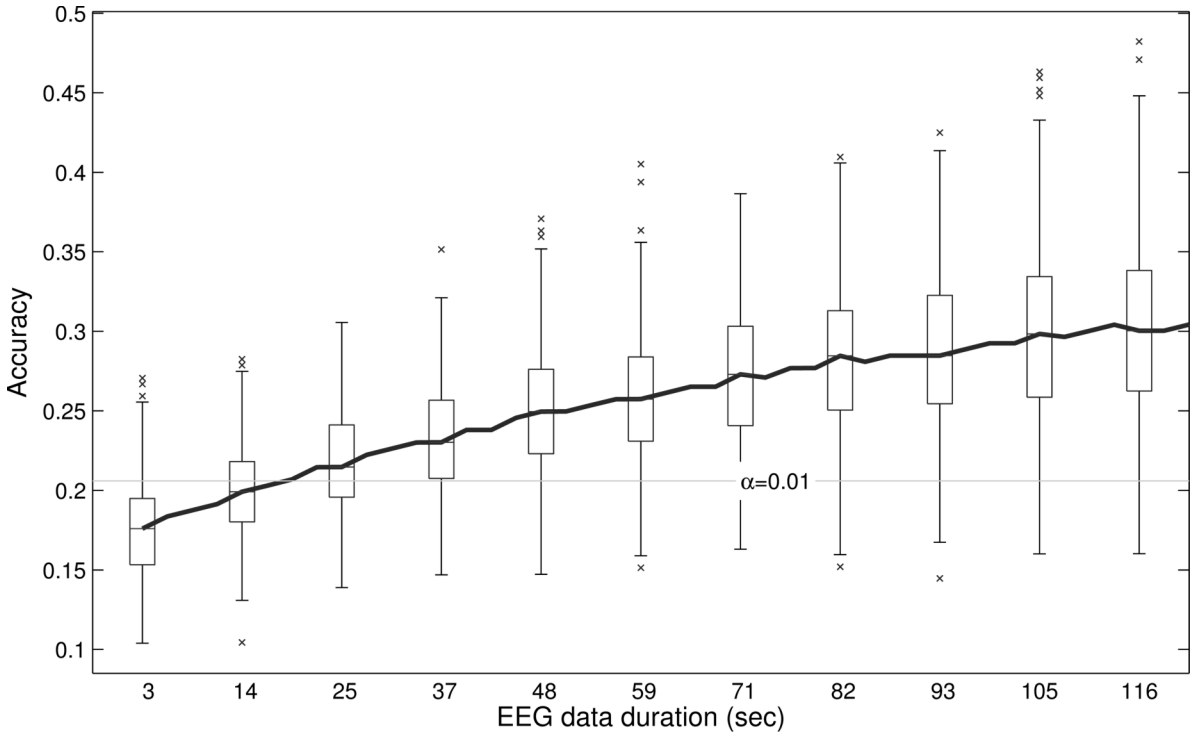

Figure 9. The classification performance found with increasingly longer durations of imagined data for the single subject used in earlier Figures. Measurements for trials using the same sentence were concatenated to produce segments of longer duration. The black curve shows the median classification accuracy. The box-and-whisker plots are shown for every fourth concatenated trial. Each box-and-whisker plot was generated from 500 random partitions of training-testing sets. The boxes represent the 25th, 50th and 75th percentiles and the whiskers the extent of the data. The x symbols indicate outliers.

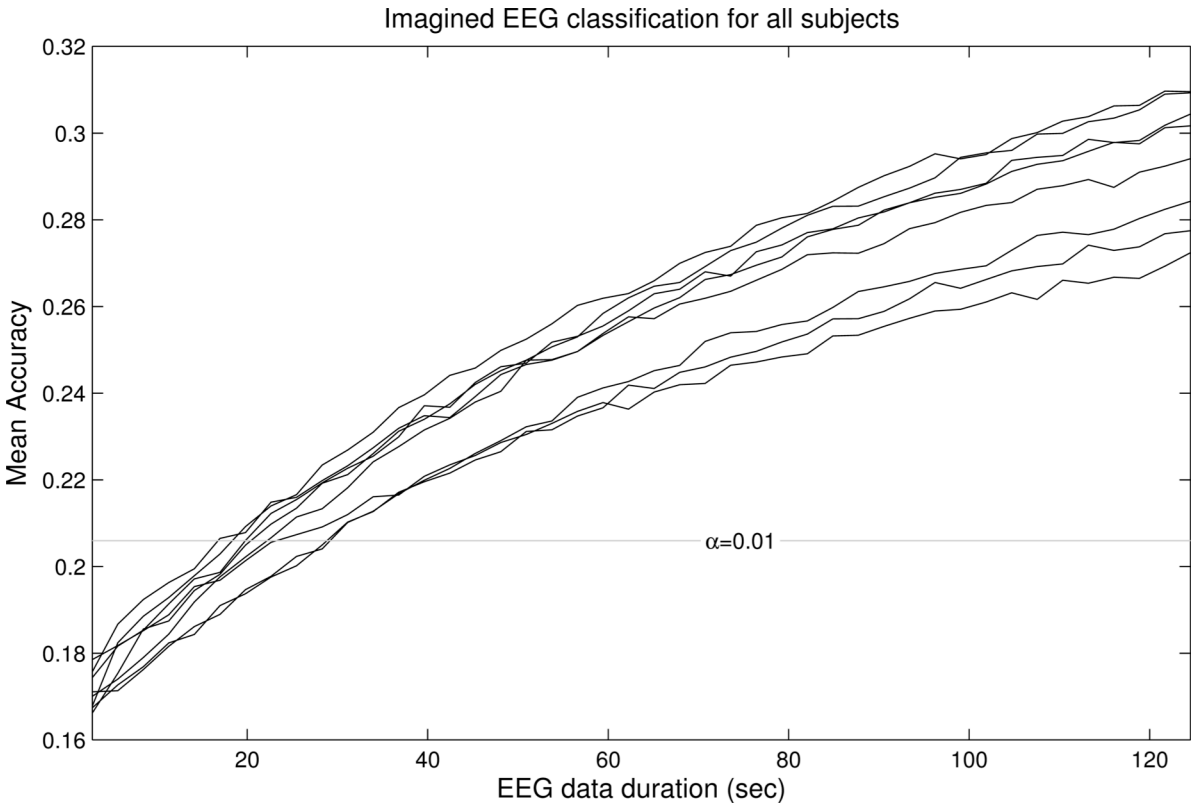

Figure 10. Classification performance as a function of imagined EEG data duration for each of the eight subjects. Measurements for trials using the same sentence were concatenated to produce segments of longer duration. Each curve shows the results for one subject. Each data point on every curve represents the median performance from a distribution of 500 random partitions of training-testing sets for that subject. The mean classification values at the longest duration of 124.5 seconds, ranking for highest to lowest, are: 0.3096, 0.3093, 0.3044, 0.3017, 0.2941, 0.2843, 0.2775, 0.2724.