Reconstructing intelligible speech from the human auditory cortex

Hassan Akbari[1,2], Bahar Khalighinejad[1,2], Jose L. Herrero[3,4], Ashesh D. Mehta[3,4], Nima Mesgarani[1,2]

[1]*Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY*

[2]*Department of Electrical Engineering, Columbia University, New York, NY*

[3]*Hofstra Northwell School of Medicine, Manhasset, NY, United States*

[4]*The Feinstein Institute for Medical Research, Manhasset, NY, United States*

Correspondence to: nima@ee.columbia.edu

Abstract

Auditory stimulus reconstruction is a technique that finds the best approximation of the acoustic stimulus from the population of evoked neural activity. Reconstructing speech from the human auditory cortex creates the possibility of a speech neuroprosthetic to establish a direct communication with the brain and has been shown to be possible in both overt and covert conditions. However, the low quality of the reconstructed speech has severely limited the utility of this method for brain-computer interface (BCI) applications. To advance the state-of-the-art in speech neuroprosthesis, we combined the recent advances in deep learning with the latest innovations in speech synthesis technologies to reconstruct closed-set intelligible speech from the human auditory cortex. We investigated the dependence of reconstruction accuracy on linear and nonlinear regression methods and the acoustic representation that is used as the target of reconstruction, including spectrogram and speech synthesis parameters. In addition, we compared the reconstruction accuracy from low and high neural frequency ranges. Our results show that a deep neural network model that directly estimates the parameters of a speech synthesizer from all neural frequencies achieves the highest subjective and objective scores on a digit recognition task, improving the intelligibility by 65% over the baseline. These results demonstrate the efficacy of deep learning and speech synthesis algorithms for designing the next generation of speech BCI systems, which not only can restore communications for paralyzed patients but also have the potential to transform human-computer interaction technologies.

Introduction

Auditory stimulus reconstruction is an inverse mapping technique that finds the best approximation of the acoustic stimulus from the population of evoked neural activity. Stimulus reconstruction was originally proposed as a method to study the representational properties of the neural population [1–5] because this method enables the intuitive interpretation of the neural responses in the stimulus domain. Reconstructing speech from the neural responses recorded from the human auditory cortex[6], however, opens up the possibility of using this technique as a speech brain-computer interface (BCI) to restore speech in severely paralyzed patients (for a review, see these references[7–9]). The ultimate goal of a speech neuroprosthesis is to create a direct communication pathway to the brain with the potential to benefit patients who have lost their ability to speak, which can result from a variety of clinical disorders leading to conditions such as locked-in syndrome[10,11]. The practicality of a neuroprosthetic device to restore speech communication was further supported by studies showing successful decoding of speech during both covert and overt conditions [12–16], which is a necessary feature of a speech BCI system. While previous

studies have established the feasibility of reconstructing speech from neural data, the quality of the reconstructed audio so far has been too low to merit subjective evaluation. For this reason, the reconstructed sounds in previous studies have been evaluated only using objective measures such as correlation or recognition accuracy. The low quality of the reconstructed sound is currently a major limiting factor in actualizing speech BCI systems[7].

The acoustic representation of the stimulus that is used as the decoding target can significantly impact the quality and accuracy of reconstructed sounds. Previous studies have used magnitude spectrogram (time-frequency representation)[3,17], speech envelope[18,19], spectrotemporal modulation frequencies[6,13,20], and discrete units such as phonemes and phonetic categories[8,21–23] and words[24,25]. Using discrete units can be advantageous by allowing for discriminative training. However, decoding discrete representations of speech such as phonemes eliminates the paralinguistic information such as speaker features, emotion, and intonation. In comparison, reconstructing continuous speech provides the possibility of real-time, continuous feedback that can be delivered to the user to promote coadaptation of the subject and the BCI algorithm[26,27] for enhanced accuracy. A natural choice is to directly estimate the parameters of a speech synthesizer from neural data, but this has not been attempted previously because the process requires a highly accurate estimation of several vocoder parameters, which is hard to achieve with traditional machine-learning techniques.

To advance the state-of-the-art in speech neuroprosthesis, we aimed to increase the intelligibility of the reconstructed speech by combining recent advances in deep learning[28] with the latest innovations in speech synthesis technologies. Deep learning models have recently become the dominant technique for acoustic and audio signal processing[29–32]. These models can improve reconstruction accuracy by imposing more complete constraints on the reconstructed audio by better modeling the statistical properties of the speech signal[3]. At the same time, nonlinear regression can invert the nonlinearly encoded speech features in neural data[33,34] more accurately.

We examined the effect of three factors on the reconstruction accuracy: 1) the regression technique (linear regression versus nonlinear deep neural network), 2) the representation of the speech intended for reconstruction (spectrogram versus speech vocoder parameters), and 3) the neural frequency range used for regression (low frequency versus high-gamma envelope) (Fig. 1A). Our results showed that a deep neural network model that uses all neural frequencies to directly estimate the parameters of a speech vocoder achieves the highest subjective and objective scores, both for intelligibility and the quality of reconstruction in a digit recognition task. These results represent an important step toward successful implementation of the next generation of speech BCI systems.

Neural recordings: We used invasive electrocorticography (ECoG) to measure neural activity from five neurosurgical patients undergoing treatment for epilepsy as they listened to continuous speech sounds. Two of the five subjects had high-density subdural grid electrodes implanted in the left hemisphere with coverage primarily over the superior temporal gyrus (STG), and four of the five subjects had depth electrodes with coverage of Heschl's gyrus (HG). All subjects had self-reported normal hearing. Subjects were presented with short continuous stories spoken by four speakers (two females, total duration: 30 minutes). To ensure that the subjects were engaged in the task, the stories were randomly paused, and the subjects were asked to repeat the last sentence.

Eight separate sentences (40 seconds) were each repeated eight times and were used as the test set for an objective evaluation of the reconstruction accuracy. In addition, three subjects listened to isolated digits from a publicly available corpus, TI-46[35]. We chose 40 digit sounds (zero to nine), spoken by four speakers (two females) that were not included in the training of the models. Reconstructed digits were used as the test set to evaluate subjective intelligibility and quality of the models. Two ranges of neural frequencies were used in the study. Low-frequency (0–50 Hz) components of the neural data were extracted by filtering the neural signals using a bandpass filter. The high-gamma (70–150 Hz) envelope[36] was extracted by filtering the neural signals and calculating the Hilbert envelope[37].

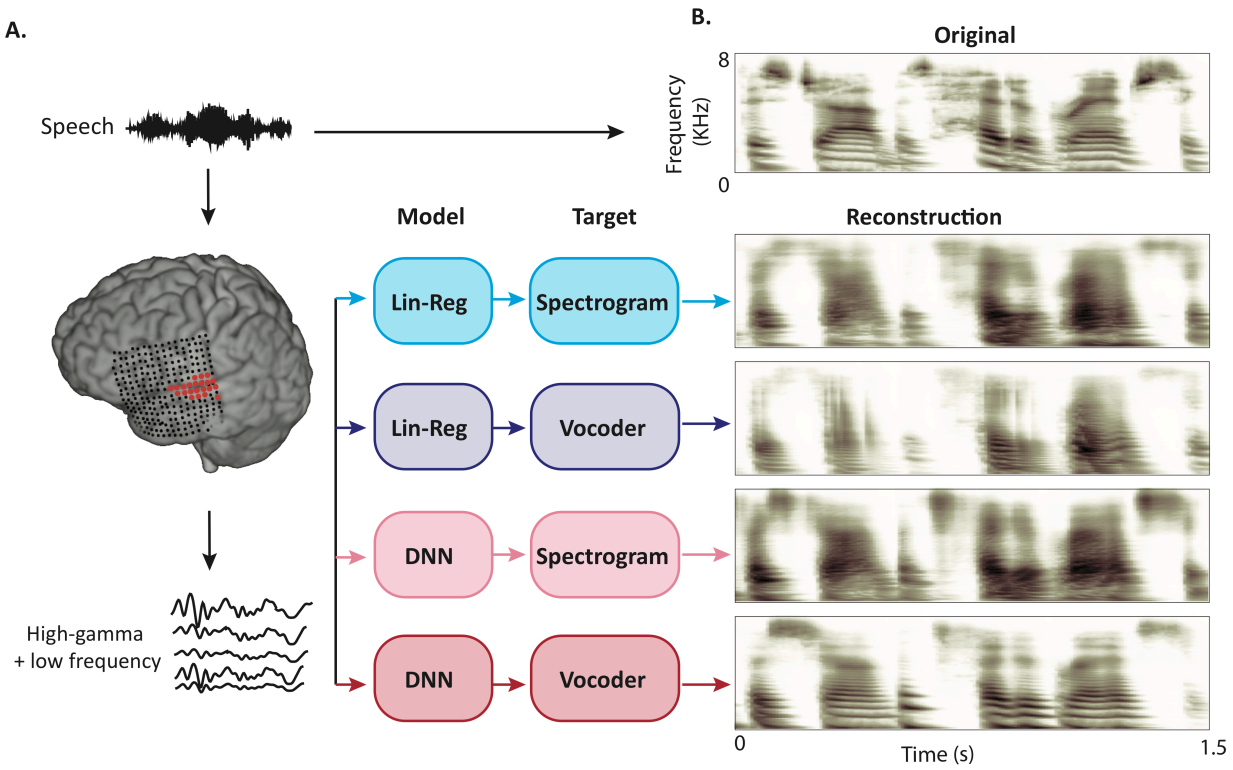

Figure 1. Schematic of the speech reconstruction method. (A) Subjects listened to natural speech sentences. The population of evoked neural activity in the auditory cortex of the listener was then used to reconstruct the speech stimulus. The responsive electrodes in an example subject are shown in red. High and low neural frequencies were extracted from the neural data. Two types of regression models and two types of speech representations were used, resulting in four combinations: linear regression to spectrogram (light blue), linear regression to vocoder (dark blue), DNN to spectrogram, and DNN to vocoder (dark red). (B) An original spectrogram of a speech sample is shown on top. The reconstructed spectrograms of the four models are shown below.

Regression models: The input to the regression models was a sliding window over the neural data with a duration of 300 ms (Fig. 2A). We compared the performance of linear and nonlinear regression models to estimate the acoustic representations from the neural signals. The linear regression found a linear mapping between the response of a population of neurons to the stimulus representation[3,6]. This method effectively assigns a spatiotemporal filter to each electrode estimated by minimizing the mean-squared-error (MSE) between the original and reconstructed stimulus.

The nonlinear regression model was implemented using a deep neural network (DNN). We designed a deep neural network architecture with two stages: 1) feature extraction and 2) feature summation networks[38–40] (Fig. 2A). In this framework, a high-dimensional representation of the input (neural responses) is first calculated, which results in mid-level features (output of the feature extraction network). These mid-level features are then input to the feature summation network to regress the output of the model (acoustic representation). The feature summation network in all cases was a two-layer fully connected network (FCN) with regularization, dropout[41], batch normalization[42], and nonlinearity between each layer. For feature extraction, we compared the efficacy of five different network architectures for spectrogram and vocoder reconstruction (Methods, Supp. Table 1 for details of each network). Specifically, we found that the fully connected network (FCN), in which no constraint was imposed on the connectivity of the nodes in each layer of the network to the previous layer, achieved the best performance for reconstructing the auditory spectrogram. However, the combination of the FCN and a locally connected network (LCN), which constrains the connectivity of each node to only a subset of nodes in the previous layer, achieved the highest performance for the vocoder representation (Supp. Tables 4, 5). In the combined FCN+LCN, the outputs of the two parallel networks are concatenated and used as the mid-level features (Fig. 2A).

Acoustic representations: We used two types of acoustic representation of the audio as the target for reconstruction: auditory spectrogram and speech vocoder. The auditory spectrogram was calculated using a model of the peripheral auditory system[43,44], which estimates a time-frequency representation of the acoustic signal on a tonotopic frequency axis. The reconstruction of the waveform from the auditory spectrogram is achieved using an iterative convex optimization procedure[43] because the phase of the signal is lost during this procedure.

For speech vocoder, we used a vocoder-based, high-quality speech synthesis algorithm (WOLRD[45]), which synthesizes speech from four main parameters: 1) spectral envelope, 2) *f0* or fundamental frequency, 3) band aperiodicity, and 4) a voiced-unvoiced (VUV) excitation label (Fig. 2B).

These parameters are then used to re-synthesize the speech waveform. This model can reconstruct high-quality speech and has been shown to outperform other methods including STRAIGHT[46]. The large numbers of the parameters in the vocoder (516 total) and the susceptibility of the synthesis quality on inaccurate estimation of parameters however pose a challenge. To remedy this, we first projected the sparse vocoder parameters onto a dense subspace in which the number of parameters can be reduced, which allows better training with a limited amount of data. We used a dimensionality reduction technique that relies on an autoencoder (AEC)[47](Fig. 2C), which compresses the vocoder parameters into a smaller space (encoder, 256 dimensions, Supp. Table 3) and subsequently recovers (decoder) the original vocoder parameters from the compressed features (Fig. 2C). The compressed features (also called bottleneck features) are used as the target for the reconstruction network. By adding noise to the bottleneck features before feeding them to the decoder during training, we can make the decoder more robust to unwanted variations in amplitude, which is necessary due to the noise inherently present in the neural signals. The autoencoder was trained on 80 hours of speech using a separate speech corpus (Wall Street Journal I[48]). During the test phase, we first reconstructed the bottleneck features from the neural data, and subsequently estimated the vocoder parameters using the decoder part of the autoencoder (Fig. 2C).

Figure 1B shows the example reconstructed spectrograms from each of the four combinations of the regression models (linear regression and DNN) and acoustic representation (auditory spectrogram and vocoder). As seen in this example, the only model that preserves the harmonic details of the speech stimulus is the DNN-vocoder model.

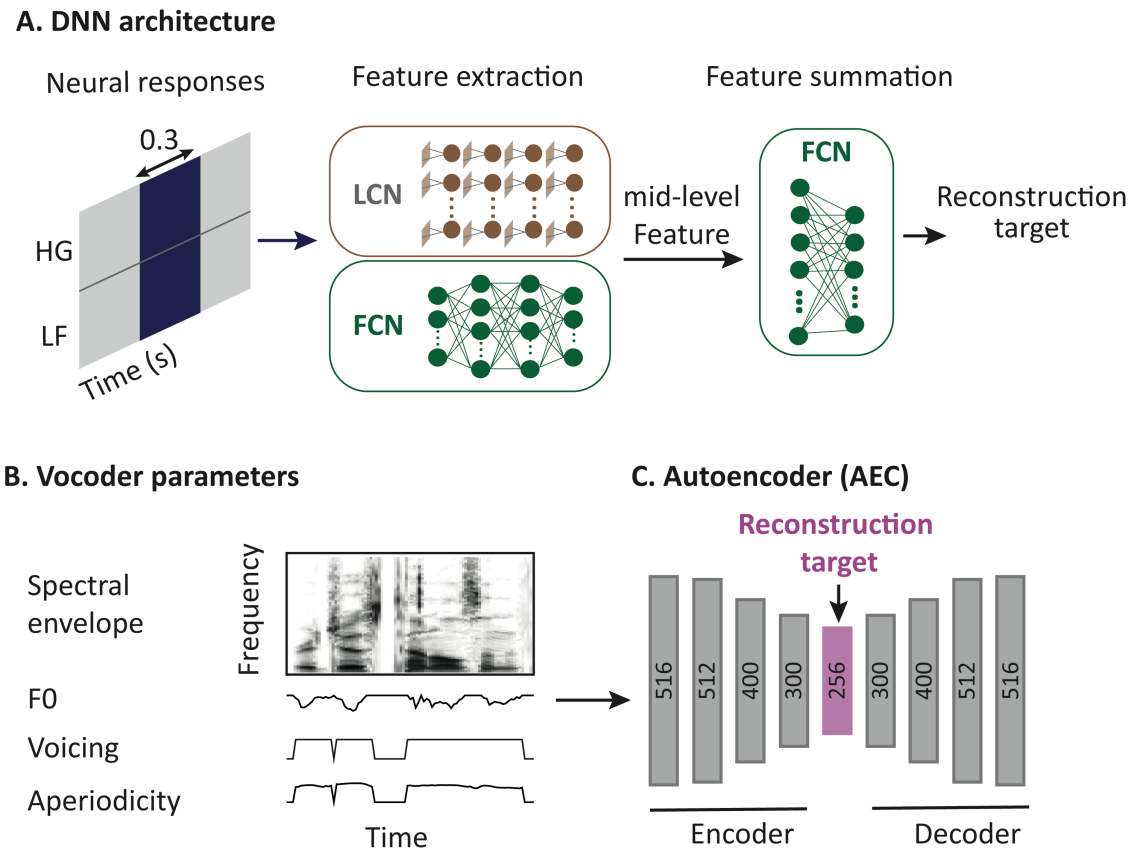

Figure 2. Deep neural network architecture (A) The input to all models was a 300 ms sliding window containing both low frequency (LF) and the high-gamma envelope (HG). The DNN architecture consists of two modules: feature extraction and feature summation networks. Feature extraction for spectrogram reconstruction was a fully connected neural network (FCN). For vocoder reconstruction, the feature extraction network consisted of an FCN concatenated with a locally connected network (LCN). The feature summation network is a two-layer fully connected neural network (FCN). (B) Vocoder parameters consist of spectral envelope, fundamental frequency (*f*0), voicing, and aperiodicity (total of 516 parameters). (C) An autoencoder with a bottleneck layer was used to reduce the 516 vocoder parameters to 256. The bottleneck features were then used as the target of reconstruction algorithms. The vocoder parameters were calculated from the reconstructed bottleneck features using the decoder part of the autoencoder network.

Subjective evaluation of the reconstruction accuracy: We used the reconstructed digit sounds to assess the subjective intelligibility and quality of the reconstructed audio. Forty unique tokens were reconstructed from each model, consisting of ten digits (zero to nine) that were spoken by two male and two female speakers. The speakers that uttered the digits were different from the speakers that were used in the training, and no digit sounds were included in the training of the networks. We asked 11 subjects with normal hearing to listen to the reconstructed digits from all four models (160 tokens total) in a random order. Each digit was heard only once. The subjects then reported the digits (zero to nine, or uncertain), rated the reconstruction quality using the mean opinion score (MOS[49], on a scale of 1 to 5), and reported the gender of the speaker (Fig. 3A).

Figure 3B shows the average reported intelligibility of the digits from the four reconstruction models. The DNN-vocoder combination achieved the best performance (75% accuracy), which is 67% higher than the baseline system (Lin-Spectrogram). Fig. 3B also shows that the reconstructions using DNN models are significantly better than the linear regression models (68.5% vs. 47.5%, paired t-test, $p<0.001$). Figure 3C shows that the subjects also rated the quality of the reconstruction significantly higher for the DNN-vocoder system than for the other three models (3.4 vs. 2.5, 2.3, and 2.1, unpaired t-test, $p<0.001$), meaning that the DNN-vocoder system sounds closest to natural speech. The subjects also accurately reported the gender of the speaker significantly higher than chance for the DNN-vocoder system (80%, t-test, $p<0.001$) while the performance for all other methods were at chance (Fig. 3D). This result indicates the importance of accurate reconstruction of pitch and harmonics frequencies for identifying speaker dependent information, which are best captured by the DNN-Voc model.

Finally, Figure 3E shows the confusion patterns in recognizing the digits for the four models, confirming again the advantage of the DNN based models, and the DNN vocoder in particular. As shown in Figure 3E, the discriminant acoustic features of the digit sounds are better preserved in the DNN-Voc model, enabling the listeners to correctly differentiate them from the other digits. Linear regression models, however, failed to preserve these cues, as seen by the high confusion among digit sounds. The confusion patterns also show that the errors were associated with the similarity of the digits (e.g., one and nine, four and five). This result suggests a possible strategy for enabling accurate discrimination in BCI applications by selecting target sounds with a sufficient acoustic distance between them. The audio samples from different models can be found online[50] and in the supplementary materials.

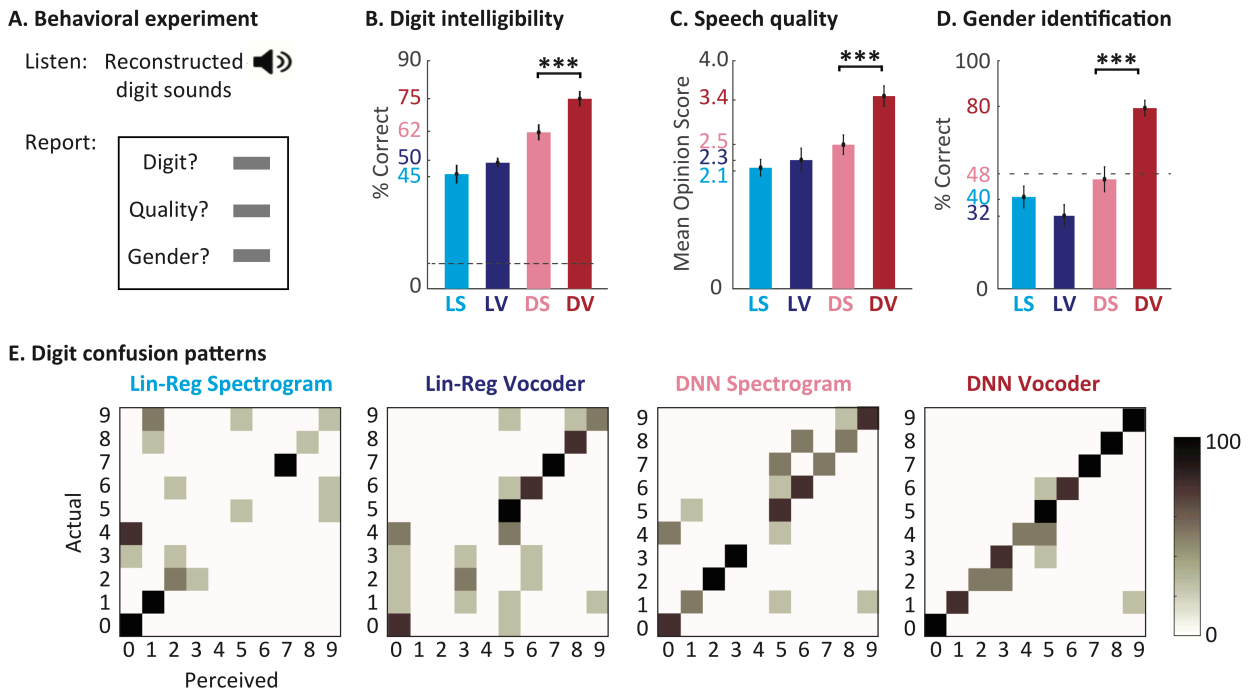

Figure 3. Subjective evaluation of the reconstruction accuracy. (A) The behavioral experiment design used to test the intelligibility and the quality of the reconstructed digits. Eleven subjects listened to digit sounds (zero to nine) spoken by two male and two female speakers. The subjects were asked to report the digit, the quality on the mean-opinion-scale, and the gender of the speaker. (B) The intelligibility score for each model defined as the percentage of correct digits reported by the subject. (C) The quality score on the MOS scale. (D) The speaker gender identification rate for each model. (E) The digit confusion patterns for each of the four models. The DNN vocoder shows the least amount of confusion among the digits.

Objective evaluation of reconstructed audio. We compared the objective reconstruction accuracy of reconstructed audio per subject using the extended short time objective intelligibility (ESTOI)[51] measure, which is commonly used for the intelligibility assessment of speech synthesis technologies. The ESTOI measures were calculated from continuous speech sentences in the test set. The average ESTOI of the reconstructed speech for all five subjects (Fig. 4A) confirms the results seen from the subjective tests, which confirm the superiority of DNN based models over the linear model, and that of vocoder reconstruction over the spectrogram ($p<0.001$, t-test). This pattern was consistent for each of the five subjects in this study, as shown in Fig. 4B alongside the electrode locations for each subject. While the overall reconstruction accuracy varies significantly across subjects, which is likely due to the difference in the coverage of the auditory cortical areas, the relative performance of the four models was the same in all subjects.

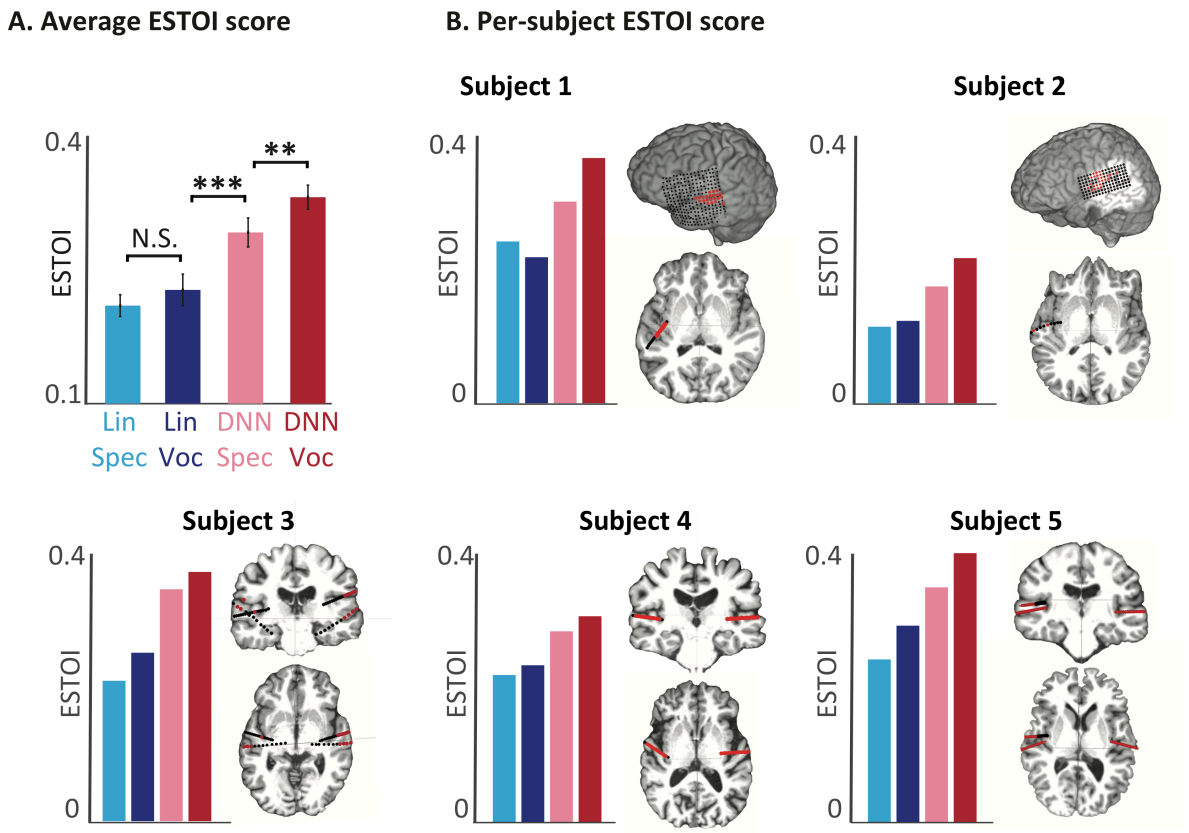

Figure 4. Objective intelligibly scores for different models. (A) The average ESTOI score based on all subjects for the four models. (B) Coverage and the location of the electrodes and ESTOI score for each of the five subjects. In all subjects, the ESTOI score of the DNN vocoder was higher than in the other models.

Reconstruction accuracy for low and high neural frequencies: There is increasing evidence that the low and high-frequency bands encode different and complementary information about the stimulus[52]. Considering that the sampling frequency of the reconstruction target is 100 Hz, we used 0–50 Hz as a low-frequency signal, and the envelope of high gamma (70–150 Hz) as high-frequency band information. To determine what frequency bands are best to include to achieve maximum reconstruction accuracy, we tested the reconstruction accuracy in three conditions, when the regression model uses only the high-gamma envelope, a low-frequency signal, or a combination of the two.

To simplify the comparison, we used only the DNN-spectrogram reconstruction model, as well as the correlation coefficients between reconstructed and original spectrograms for comparison. We found that the reconstruction accuracy from the combination of the two frequency bands significantly outperforms the reconstruction accuracy when only one of the frequency bands is used (Fig. 5A, $p<0.001$, t-test). This observation is consistent with the complementary encoding of the stimulus features in the low and high-frequency bands[53], which implicates the advantage of using the entire neural signal to achieve the best performance in speech neuroprosthesis applications, when it is practically possible.

Effect of the number of electrodes and duration of training data: The variability of the reconstruction accuracy across subjects (Fig. 4B) suggests an important role of neural coverage in improving the reconstruction[3,6]. In addition, because some of the noise signal across different electrodes is independent, reconstruction from a combination of electrodes may lead to a higher accuracy by finding a signal subspace less affected by the noise in the data[54]. To examine the effect of the number of electrodes on the reconstruction accuracy, we first combined the electrodes of all five subjects and randomly chose N electrodes (N = 1, 2, 4, 8, 16, 32, 64, 128), twenty times for training the individual networks. The average reconstruction accuracy for each N was then used for comparison. The results shown in Fig. 5B indicate that increasing the number of electrodes improves the reconstruction accuracy; however, the rate of improvement decreased significantly.

Finally, because the success of neural network models is largely attributed to training on large amounts of data[28], we examined the effect of training duration on reconstruction accuracy. We used 128 randomly chosen electrodes and trained several neural network models each on a segment of the training data as the duration of the segments was gradually increased from 10 to 30 minutes. This process was performed twenty times for each duration by choosing a random segment of the training data, and the reconstruction accuracy was averaged over the segments. As expected, the results show an increased

reconstruction accuracy as the duration of the training was increased (Fig. 5C), which indicates the importance of collecting a larger duration of training data when it is practically feasible.

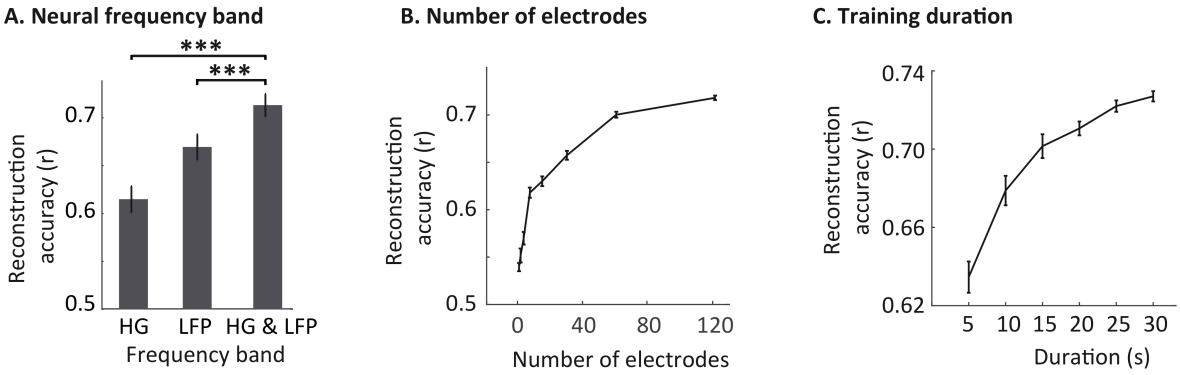

Figure 5. Effect of neural frequency range, number of electrodes, and stimulus duration on reconstruction accuracy. (A) The reconstruction accuracy based on high gamma, low frequency, and high gamma and low frequency combined. (B) The accuracy of reconstruction when the number of electrodes increases from one to 128. For each condition, 20 random subsets were chosen. (C) The accuracy of reconstruction when the duration of the training data increases. Each condition is the average of 20 random subsets.

Discussion:

We compared the performance of linear and nonlinear (DNN) regression models in reconstructing the spectrogram and vocoder representations of speech signals. We found that the combination of deep neural network models with vocoder parameters significantly outperformed the linear regression and spectrogram representations of speech, and resulted in 75% intelligibility scores on a closed-set, digit recognition task.

Our results are consistent with those of previous reconstruction studies that showed the importance of nonlinear techniques in neural decoding[55]. The previous methods have used support vector machines[13,56], linear discriminant analysis[57,58], linear regression[3,14,59], nonlinear embedding[6], and Bayes classifiers[15]. In recent years, deep learning[60] has shown tremendous success in many brain-computer interface technologies[61], and our study extended this trend by showing the benefit of deep learning in speech neuroprosthesis research[55].

We showed that the reconstruction accuracy depends on both the number of electrodes and the duration of the data that is available for training. This is consistent with the findings of studies showing the superior advantage of deep learning models over other techniques, particularly when the amount of training data is large[28]. We showed that the rate of improvement slows down as the number of electrodes increases. This could indicate the limited diversity of the neural responses in our recording which ultimately limits the added information that is gained from additional electrodes. Alternatively, increasing the number of electrodes also increases the complexity and the number of free parameters in the neural network model. Because the duration of our training data was limited, it is possible that more training data would be needed before the benefit of additional features becomes apparent. Because our experiments showed that increasing the amount of training data results in better reconstruction accuracy, methods that can be used to increase the amount of data that is available for the training of deep models are highly desirable, for example, when chronic recordings are possible in long-term implantable devices such as RNS [62].

We showed that the representation of the acoustic signal used as the target of reconstruction has an important role in the intelligibility and the quality of the reconstructed audio. We used a vocoder representation of speech, which extends the previous studies that used a magnitude spectrogram (time-frequency representation)[3,17], speech envelope[18,19], spectrotemporal modulation frequencies[6,13,20], and discrete units such as phonemes and phonetic categories[8,21–23] and words[24,25]. Reconstruction of the spectrogram, which we also used for comparison, inherently results in suboptimal audio quality because the phase of the spectrogram must be approximated. The discrete units such as phonemes enable

discriminative training by learning a direct map from the neural data to the class labels, which is typically more efficient than generative regression models[63]. The continuous nature of parameters in acoustic reconstruction however could prove advantageous for BCI applications because they provide a continuous feedback to the user[64], which is crucial for the subject and the BCI algorithm to coadapt to increase overall effectiveness[26,27]. Therefore, direct reconstruction of speech synthesis parameters is a natural choice. This choice however poses a challenge, since the vocoder quality is very sensitive to the quality of the decoding. As we have reported, reconstructing vocoder parameters resulted in both the worst (when used with linear regression) and the best (when used with DNN) results. Therefore, powerful modeling techniques such as deep learning are crucial as more inclusive representations of the speech signal are used for reconstruction and decoding applications. We proposed a solution to this problem by compressing the acoustic features into a low-dimensional space and using a decoder that is robust to the fluctuations of the input.

We found that the combination of low frequency and the envelope of high gamma results in higher reconstruction accuracy than each frequency band alone. This finding is consistent with those of studies that have shown the importance of an oscillatory phase[65] in addition to the neural firing rate, which is reflected in the high-gamma frequency band[66]. Combining both high and low frequencies not only enables access to the complementary information in each band[52,67] but also allows the decoder to use the information that is encoded in the interactions between the two bands, such as cross-frequency coupling[53]. Overall, we observed that better brain coverage, more training data, and combined neural frequency bands result in the best reconstruction accuracy, which can serve as an upper bound performance where practical limitations prevent the use of all possible factors, for example, where the brain coverage is small, or high-frequency neural signals are not accessible such as in noninvasive neuroimaging methods.

Finally, while previous studies have already shown the successful generalization of reconstruction techniques from overt to covert speech[12–16], further research is needed to devise system architectures and training procedures that can optimally fine-tune a model to perform and generalize well in both conditions. Furthermore, expanding from the closed-set intelligible speech in this work to continuous, open-set, natural intelligible speech requires additional research, which will undoubtedly benefit from a larger amount of training data, higher-resolution neural recording technologies[68], and the adaptation of regression models[69] and the subject to improve the BCI system[26,27].

In summary, we present a general framework that can be used for speech neuroprothesis technologies that can result in accurate and intelligible reconstructed speech from the human auditory

cortex. Our approach takes a step toward the next generation of human-computer interaction systems and more natural communication channels for patients suffering from paralysis and locked-in syndromes.

Materials and methods:

Participants and neural recording

Five patients with pharmacoresistent focal epilepsy were included in this study. All subjects underwent chronic intracranial encephalography (iEEG) monitoring at Northshore University Hospital to identify epileptogenic foci in the brain for later removal. Three subjects were implanted with only stereo-electroencephalographic (sEEG) depth arrays, one with a high-density grid, and one with both grid and depth electrodes (PMT, Chanhassen, MN, USA). The electrodes showing any sign of abnormal epileptiform discharges, as identified in the epileptologists' clinical reports, were excluded from the analysis. All included iEEG time series were manually inspected for signal quality and were free from interictal spikes. All research protocols were approved and monitored by the institutional review board at the Feinstein Institute for Medical Research, and informed written consent to participate in the research studies was obtained from each subject before electrode implantation. All research was performed in accordance with relevant guidelines and regulations.

Intracranial EEG (iEEG) signals were acquired continuously at 3 kHz per channel (16-bit precision, range ± 8 mV, DC) using a data acquisition module (Tucker-Davis Technologies, Alachua, FL, USA). Either subdural or skull electrodes were used as references, as dictated by recording quality at the bedside after online visualization of the spectrogram of the signal. Speech signals were recorded simultaneously with the iEEG for subsequent offline analysis. The amplitude of the high-gamma response (75–150 Hz) was extracted using the Hilbert transform[37] and was resampled to 100 Hz. The high-gamma responses were normalized based on the responses recorded during a 2-minute silence interval before each recording.

Brain maps

The electrode positions were mapped to brain anatomy using registration of the post-implant computed tomography (CT) to the pre-implant MRI via the post-op MRI[70]. After coregistration, the electrodes were identified on the post-implantation CT scan using BioImage Suite[71]. Following coregistration, the subdural grid and strip electrodes were snapped to the closest point on the reconstructed brain surface of the pre-implantation MRI. We used the FreeSurfer automated cortical parcellation[72] to identify the anatomical regions in which each electrode contact was located within approximately 3 mm resolution (the maximum parcellation error of a given electrode to a parcellated area was < 5 voxels/mm). We used Destrieux's parcellation because it provides higher specificity in the ventral and lateral aspects of the medial lobe[73]. The automated parcellation results for each electrode were closely inspected by a neurosurgeon using the patient's coregistered post-implant MRI.

Stimulus

The speech materials included continuous speech stories recorded in-house by four voice actors and actresses (duration: 30 min, 11,025 Hz sampling rate). Eight of the sentences (40 seconds) were used for objective tests and were presented to the patients eight times to improve the signal to noise ratio. The digit sounds were taken from the TI-46 corpus[35]. Two female (f2 and f8) and two male (m2 and m5) speakers were chosen from the corpus, and one token per digit and speaker was used (total of 40 unique tokens). Each digit was repeated six times to improve the signal to noise ratio of the neural responses. The speakers that uttered the digits were different from the speakers that narrated the stories.

Acoustic representation

The auditory spectrogram representation of speech was calculated from a model of the peripheral auditory system[34]. The model consists of three stages: 1) a cochlear filter bank consisting of 128 constant-Q filters equally spaced on a logarithmic axis, 2) a hair cell stage consisting of a low-pass filter and a nonlinear compression function, and 3) a lateral inhibitory network, consisting of a first-order derivative along the spectral axis. Finally, the envelope of each frequency band was calculated to obtain a time-frequency representation simulating the pattern of activity on the auditory nerve[74]. The final spectrogram has a sampling frequency of 100 Hz. The audio signal was reconstructed from the auditory spectrogram using an iterative convex optimization procedure[43]. For the vocoder-based speech synthesizer, we used the WORLD[45] (D4C edition) system. In this model, four major speech parameters were estimated, from which the speech waveform was synthesized: 1) spectral envelope, 2) *f*0 or fundamental frequency, 3) band aperiodicity, and 4) voiced-unvoiced (VUV) excitation label. The dimension of each parameter was automatically calculated by the vocoder method and was based on the window size and the sampling frequency of the waveform (16 KHz).

DNN architecture

We used a common deep neural network architecture that consists of two stages: feature extraction and feature summation[38–40] (Fig. 2A). In this framework, a high-dimensional representation of the input is first calculated (feature extraction), which is then used to regress the output of the model (feature summation). In all models examined, the feature summation step consisted of a two-layer fully connected network (FCN) with L2 regularization, dropout[41], batch normalization[42], and nonlinearity in each layer.

We study five different architectures for the feature extraction part of the network: the fully connected network (FCN, also known as the multilayer perceptron or MLP), the locally connected network (LCN) [75], convolutional neural network (CNN) [76], FCN+CNN, and FCN+LCN (for details of each architecture see Supp. Table 1). In the combined networks, we concatenated the output of two parallel paths, which were fed into the summation network. For FCN, the windowed neural responses were flattened and fed to a multilayer FCN. However, in LCN and CNN, all the extracted features were of the same size as the input, meaning that we did not use flattening, strided convolution, or downsampling prior to the input layer or between the two consecutive layers. Instead, the final output of the multilayer LCN or CNN was flattened prior to feeding the output into the feature summation network.

The optimal network structure was found separately for the spectrogram and vocoder parameters using an ablation study. For spectrogram reconstruction, we directly regressed the 128 frequency bands using a multilayer FCN model for feature extraction (Supp. Table 5). This architecture, however, was not plausible for reconstructing vocoder parameters due to the high-dimensionality and statistical variability of the vocoder parameters. To remedy this, we used a deep autoencoder network (AEC) [47] to find a compact representation of the 516-dimensional vocoder parameters (512 spectral envelope, pitch, voiced-unvoiced, and band periodicity) [45]. We confirmed that decoding the AEC features performed significantly better than decoding the vocoder parameters directly (Supp. Table 2). The structure for the proposed deep AEC is illustrated in Figure 2D. To carry out decoding, we used a multilayer FCN, in which the number of the nodes changed in a descending (encoder) and then ascending order (decoder) (Fig. 2C)(supp. Table 6). The bottleneck layer of such a network (or the output of the encoder part of the pre-trained AEC) can be used as a low-dimensional reconstruction target by employing the neural network model, from which the vocoder parameters can be estimated using the decoder part of the AEC. We chose the number of nodes in the bottleneck layer to be 256, because it maximized both the objective reconstruction accuracy (Supp. Table 3), and the subjective assessment of the reconstructed sound. To increase the robustness to unwanted variations in the encoded features, we used two methods in the bottleneck layer: 1) the hyperbolic tangent function (tanh) was used as a nonlinearity to control the range of the encoded features, and 2) Gaussian noise was added during training prior to feeding into the first layer of the decoder part to make the decoder robust enough to unwanted changes in amplitude resulting from noises in neural responses. We confirmed that using additive Gaussian noise in the bottleneck instead of dropout performed significantly better (paired t-test, $p<0.001$). It is important that we use the same nonlinearity as the bottleneck (tanh) in the output of the main network, since the estimations should

be in the same range and space as those in which they were originally coded. The best network architecture for decoding the vocoder parameters was found to be the FCN+LCN network (Supp. Table 4).

DNN training and cross validation

The networks were implemented in Keras with a Tensorflow backend[77]. Initialization of the weights was performed using a previously proposed method[77], which was specifically developed for deep multilayer networks with rectified linear units (ReLUs) as their nonlinearities. It has been shown that using this method helps such networks converge faster. We used batch normalization [42], nonlinearity, and a dropout of p=0.3 [41] between each layer. We applied an *L2* penalty (with a multiplier weight set to 0.001) on the weights of all the layers in all types of networks (including the AEC). However, we found that using additive Gaussian noise in the bottleneck of the AEC instead of dropout and regularization performed significantly better (paired t-test, $p<0.001$). We used three types of nonlinearities in the networks: 1) LeakyReLU[78] for all layers of AEC except the bottleneck and for all layers of the feature extraction part of the main network, 2) tanh for the output layer of the main network and the bottleneck of the AEC, and 3) the exponential linear unit (ELU) [79] for the feature summation network. Each epoch of training had a batch size of 256, and optimization was performed using Adam[80] with an initial learning rate of 0.0001, which was reduced by a factor of two if the validation loss did not improve in four consecutive epochs. Network training was achieved in 150 epochs and was performed for each subject separately. The loss function was a combination of MSE and Pearson's correlation coefficient for each sample:

$$\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2 - \frac{\sum_i (y_i - \bar{y}_i)(\hat{y}_i - \bar{\hat{y}}_i)}{\sqrt{\sum_i (y_i - \bar{y}_i)^2 \sum_i (y_i - \bar{y}_i)^2}}$$

in which y is the actual label (spectrogram or vocoder features) for that sample and $\hat{y}$ is the reconstruction from the output layer of the network. The maximum time-lag used was $\tau_{max} = 300$ ms. Because both CNN and LCN use small receptive fields that take local patterns into account, we retained the spatial organization of the electrode sites in the input to the network, meaning that the electrodes that were close to each other in the brain were arranged to be close together in the input data matrix.

Cross validation

We trained both the LR model and the DNN models using cross validation. We used the speech stories for training all models, and used repeated sentences (separate set from the stories) and digit sounds for testing. No digit sound was included in the training, and the speakers that uttered the digits were different

from those that read the stories. The autoencoder network (AEC) was trained on a separate speech corpus (Wall Street Journal, WSJ, 80 hours of read speech) [48].

Subjective and objective evaluations

We assessed the intelligibility of the reconstructed speech using both subjective and objective tests. For subjective assessment, 11 participants with normal hearing listened to the reconstructed digits using headphones in a quiet environment. Each participant listened to 160 tokens including 10 digits, four speakers, and four models. The participants were asked to report the digit or to select unsure if the digit was not intelligible. In addition, the participants reported the quality of the reconstructed speech using a mean opinion score (MOS): 1 (bad), 2 (poor), 3 (fair), 4 (good), and 5 (excellent). The participants also reported the gender of the speaker. For objective evaluation, we used the ESTOI measure[51] which is a monaural intelligibility prediction algorithm commonly used in speech enhancement and synthesis research.

Data availability

The data that support the findings of this study are available upon request from the corresponding author [NM].

Code availability

The codes for performing phoneme analysis, calculating high-gamma envelope, and reconstructing the spectrogram are available at *http://naplab.ee.columbia.edu/naplib.html* [81].

Acknowledgments

We thank James O'Sullivan for providing helpful comments on the manuscript. This work was funded by a grant from the National Institutes of Health, NIDCD, DC014279 and the Pew Charitable Trusts, Pew Biomedical Scholars Program.

Author Contributions

B.K., J.H., N.M., A.M. collected the data. H.A., B.K., N.M., designed the experiment, evaluated the results and wrote the manuscript. All authors commented on the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R. & Warland, D. Reading a neural code. *Science (80-.).* **252,** 1854–1857 (1991).
2. Rieke, F., Bodnar, D. A. & Bialek, W. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc Biol Sci* **262,** 259–265 (1995).
3. Mesgarani, N., David, S. V. S. V., Fritz, J. B. J. B. & Shamma, S. A. S. A. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *J Neurophysiol* **102,** 3329–3339 (2009).
4. Stanley, G. B., Li, F. F. & Dan, Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J Neurosci* **19,** 8036–8042 (1999).
5. Ramirez, A. D. *et al.* Incorporating naturalistic correlation structure improves spectrogram reconstruction from neuronal activity in the songbird auditory midbrain. *J. Neurosci.* **31,** 3828–3842 (2011).
6. Pasley, B. N. B. N. *et al.* Reconstructing speech from human auditory cortex. *PLoS Biol.* **10,** (2012).
7. Chakrabarti, S., Sandberg, H. M., Brumberg, J. S. & Krusienski, D. J. Progress in speech decoding from the electrocorticogram. *Biomed. Eng. Lett.* **5,** 10–21 (2015).
8. Herff, C. & Schultz, T. Automatic speech recognition from neural signals: a focused review. *Front. Neurosci.* **10,** 429 (2016).
9. Iljina, O. *et al.* Neurolinguistic and machine-learning perspectives on direct speech BCIs for restoration of naturalistic communication. *Brain-Computer Interfaces* **4,** 186–199 (2017).
10. Laureys, S. *et al.* The locked-in syndrome: what is it like to be conscious but paralyzed and voiceless? *Prog. Brain Res.* **150,** 495–611 (2005).
11. Sellers, E. W., Ryan, D. B. & Hauser, C. K. Noninvasive brain-computer interface enables communication after brainstem stroke. *Sci. Transl. Med.* **6,** 257re7-257re7 (2014).
12. Tian, X. & Poeppel, D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Front. Psychol.* **1,** 166 (2010).
13. Martin, S. *et al.* Word pair classification during imagined speech using direct brain recordings. *Sci. Rep.* **6,** 25803 (2016).
14. Leuthardt, E. C. *et al.* Using the electrocorticographic speech network to control a brain–computer interface in humans. *J. Neural Eng.* **8,** 36004 (2011).
15. Pei, X., Barbour, D. L., Leuthardt, E. C. & Schalk, G. Decoding vowels and consonants in spoken

and imagined words using electrocorticographic signals in humans. *J. Neural Eng.* **8,** 46028 (2011).

16. Martin, S. *et al.* Decoding spectrotemporal features of overt and covert speech from the human cortex. *Front. Neuroeng.* **7,** 14 (2014).
17. Mesgarani, N. & Chang, E. F. E. F. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* **485,** 233–236 (2012).
18. O'Sullivan, J. A. *et al.* Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cereb. Cortex* bht355 (2014).
19. Ding, N. & Simon, J. Z. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci.* **109,** 11854–11859 (2012).
20. Santoro, R. *et al.* Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. *Proc. Natl. Acad. Sci.* **114,** 4799–4804 (2017).
21. Moses, D. A. D. A., Mesgarani, N., Leonard, M. K. M. K. & Chang, E. F. E. F. Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity. *J. Neural Eng.* **13,** 56004 (2016).
22. Khalighinejad, B., da Silva, G. C. & Mesgarani, N. Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *J. Neurosci.* **37,** 2176–2185 (2017).
23. Di Liberto, G. M., O'Sullivan, J. A. & Lalor, E. C. Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Curr. Biol.* **25,** 2457–2465 (2015).
24. Kellis, S. *et al.* Decoding spoken words using local field potentials recorded from the cortical surface. *J. Neural Eng.* **7,** 56007 (2010).
25. Herff, C. *et al.* Brain-to-text: decoding spoken phrases from phone representations in the brain. *Front. Neurosci.* **9,** 217 (2015).
26. Vidaurre, C., Sannelli, C., Müller, K.-R. & Blankertz, B. Machine-learning-based coadaptive calibration for brain-computer interfaces. *Neural Comput.* **23,** 791–816 (2011).
27. McFarland, D. J., Sarnacki, W. A. & Wolpaw, J. R. Should the parameters of a BCI translation algorithm be continually adapted? *J. Neurosci. Methods* **199,** 103–107 (2011).
28. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436 (2015).
29. Hinton, G. *et al.* Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Process. Mag. IEEE* **29,** 82–97 (2012).
30. Luo, Y. Y., Chen, Z. & Mesgarani, N. Speaker-Independent Speech Separation With Deep Attractor Network. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **26,** 787–796 (2018).

31. Chen, Z., Luo, Y. Y. & Mesgarani, N. Deep attractor network for single-microphone speaker separation. in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* 246–250 (IEEE, 2017). doi:10.1109/ICASSP.2017.7952155
32. O'Sullivan, J. *et al.* Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *J. Neural Eng.* **14,** 56001 (2017).
33. David, S. V. S. V., Mesgarani, N., Fritz, J. B. J. B. & Shamma, S. A. S. A. Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. *J Neurosci* **29,** 3374–3386 (2009).
34. Mesgarani, N., David, S. V. S. V., Fritz, J. B. J. B. & Shamma, S. A. S. A. Mechanisms of noise robust representation of speech in primary auditory cortex. *Proc. Natl. Acad. Sci.* **111,** 6792–6797 (2014).
35. Liberman, Mark, et al. TI 46-Word LDC93S9. *Linguistic Data Consortium, Philadelphia* (1993).
36. Crone, N. E., Boatman, D., Gordon, B. & Hao, L. Induced electrocorticographic gamma activity during auditory perception. *Clin. Neurophysiol.* **112,** 565–582 (2001).
37. Edwards, E. *et al.* Comparison of time–frequency responses and the event-related potential to auditory speech stimuli in human cortex. *J. Neurophysiol.* **102,** 377–386 (2009).
38. LeCun, Y. *et al.* Handwritten digit recognition with a back-propagation network. in *Advances in neural information processing systems* 396–404 (1990).
39. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. in *Advances in neural information processing systems* 1097–1105 (2012).
40. Pinto, N., Doukhan, D., DiCarlo, J. J. & Cox, D. D. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. *PLoS Comput. Biol.* **5,** e1000579 (2009).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15,** 1929–1958 (2014).
42. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv Prepr. arXiv1502.03167* (2015).
43. Chi, T., Ru, P. & Shamma, S. A. Multiresolution spectrotemporal analysis of complex sounds. *J Acoust Soc Am* **118,** 887–906 (2005).
44. Mesgarani, N., Slaney, M. & Shamma, S. A. S. A. Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. *IEEE Trans. Audio. Speech. Lang. Processing* **14,** 920–930 (2006).

45. Morise, M., Yokomori, F. & Ozawa, K. WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **99,** 1877–1884 (2016).
46. Kawahara, H., Masuda-Katsuse, I. & De Cheveigne, A. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds1. *Speech Commun.* **27,** 187–207 (1999).
47. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science (80-.).* **313,** 504–507 (2006).
48. Paul, D. B. & Baker, J. M. The design for the Wall Street Journal-based CSR corpus. in *Proceedings of the workshop on Speech and Natural Language* 357–362 (Association for Computational Linguistics, 1992).
49. Salza, P. L., Foti, E., Nebbia, L. & Oreglia, M. MOS and pair comparison combined methods for quality evaluation of text-to-speech systems. *Acta Acust. united with Acust.* **82,** 650–656 (1996).
50. Reconstruction audio samples: naplab.columbia.edu/Reconstruction.
51. Taal, C. H., Hendriks, R. C., Heusdens, R. & Jensen, J. A short-time objective intelligibility measure for time-frequency weighted noisy speech. in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on* 4214–4217 (IEEE, 2010).
52. Buzsáki, G., Anastassiou, C. A. & Koch, C. The origin of extracellular fields and currents—EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* **13,** 407–420 (2012).
53. Canolty, R. T. & Knight, R. T. The functional role of cross-frequency coupling. *Trends Cogn. Sci.* **14,** 506–515 (2010).
54. Paninski, L., Pillow, J. & Lewi, J. Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* **165,** 493–507 (2007).
55. Yang, M. *et al.* Speech reconstruction from human auditory cortex with deep neural networks. in *Sixteenth Annual Conference of the International Speech Communication Association* **2015–Janua,** (2015).
56. Blakely, T., Miller, K. J., Rao, R. P. N., Holmes, M. D. & Ojemann, J. G. Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2008,** 4964–7 (2008).
57. Mugler, E. M. *et al.* Direct classification of all American English phonemes using signals from functional speech motor cortex. *J. Neural Eng.* **11,** 35015 (2014).
58. Lotte, F. *et al.* Electrocorticographic representations of segmental features in continuous speech.

Front. Hum. Neurosci. **9,** 97 (2015).

59. Herff, C. *et al.* Towards direct speech synthesis from ECoG: A pilot study. in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the* 1540–1543 (IEEE, 2016).
60. Hinton, G. E. *et al.* A fast learning algorithm for deep belief nets. *Neural Comput.* **18,** 1527–1554 (2006).
61. Hajinoroozi, M., Mao, Z., Jung, T.-P., Lin, C.-T. & Huang, Y. EEG-based prediction of driver's cognitive performance by deep convolutional neural network. *Signal Process. Image Commun.* **47,** 549–555 (2016).
62. Morrell, M. Brain stimulation for epilepsy: can scheduled or responsive neurostimulation stop seizures? *Curr. Opin. Neurol.* **19,** 164–168 (2006).
63. Efron, B. The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Stat. Assoc.* **70,** 892–898 (1975).
64. Koyama, S. *et al.* Comparison of brain–computer interface decoding algorithms in open-loop and closed-loop control. *J. Comput. Neurosci.* **29,** 73–87 (2010).
65. Luo, H. & Poeppel, D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* **54,** 1001–1010 (2007).
66. Ray, S. & Maunsell, J. H. R. Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biol.* **9,** (2011).
67. Miller, K. J., Sorensen, L. B., Ojemann, J. G. & Den Nijs, M. Power-law scaling in the brain surface electric potential. *PLoS Comput. Biol.* **5,** e1000609 (2009).
68. Khodagholy, D. *et al.* NeuroGrid: recording action potentials from the surface of the brain. *Nat. Neurosci.* **18,** 310 (2015).
69. Shenoy, P., Krauledat, M., Blankertz, B., Rao, R. P. N. & Müller, K.-R. Towards adaptive classification for BCI. *J. Neural Eng.* **3,** R13 (2006).
70. Groppe, D. M. *et al.* iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. **281,**
71. Papademetris, X. *et al.* BioImage Suite: An integrated medical image analysis suite: An update. *Insight J.* **2006,** 209 (2006).
72. Fischl, B. *et al.* Automatically parcellating the human cerebral cortex. *Cereb. cortex* **14,** 11–22 (2004).
73. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and

sulci using standard anatomical nomenclature. *Neuroimage* **53,** 1–15 (2010).

74. Yang X. Shamma S. A., W. K. Auditory representations of acoustic signals. *IEEE Trans. Inf. Theory* **38,** 824–839 (1992).

75. Coates, A. & Ng, A. Y. Selecting receptive fields in deep networks. in *Advances in Neural Information Processing Systems* 2528–2536 (2011).

76. LeCun, Y. & Bengio, Y. Convolutional networks for images, speech, and time series. *Handb. brain theory neural networks* **3361,** (1995).

77. Abadi, M. *et al.* TensorFlow: A System for Large-Scale Machine Learning. in *OSDI* **16,** 265–283 (2016).

78. Maas, A. L., Hannun, A. Y. & Ng, A. Y. Rectifier nonlinearities improve neural network acoustic models. in *Proc. icml* **30,** 3 (2013).

79. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv Prepr. arXiv1511.07289* (2015).

80. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv Prepr. arXiv1412.6980* (2014).

81. Khalighinejad, B., Nagamine, T., Mehta, A. & Mesgarani, N. NAPLib: An open source toolbox for real-time and offline Neural Acoustic Processing. in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on* 846–850 (IEEE, 2017). doi:10.1109/ICASSP.2017.7952275