Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Neural Eng. Author manuscript; available in PMC 2022 October 24.

Published in final edited form as:

J Neural Eng. ; 17(6): . doi:10.1088/1741-2552/abc742.

Brain2Char: a deep architecture for decoding text from brain recordings

Pengfei Sun[1,2], **Gopala K Anumanchipalli**[1,2], **Edward F Chang**[1,3]

[1]Center of Integrative Neurosciences, University of California, San Francisco, CA, United States of America

[2]These authors contributed equally to this work.

Abstract

Objective.—Decoding language representations directly from the brain can enable new brain–computer interfaces (BCIs) for high bandwidth human–human and human–machine communication. Clinically, such technologies can restore communication in people with neurological conditions affecting their ability to speak.

Approach.—In this study, we propose a novel deep network architecture Brain2Char, for directly decoding text (specifically character sequences) from direct brain recordings (called electrocorticography, ECoG). Brain2Char framework combines state-of-the-art deep learning modules—3D Inception layers for multiband spatiotemporal feature extraction from neural data and bidirectional recurrent layers, dilated convolution layers followed by language model weighted beam search to decode character sequences, and optimizing a connectionist temporal classification loss. Additionally, given the highly non-linear transformations that underlie the conversion of cortical function to character sequences, we perform regularizations on the network's latent representations motivated by insights into cortical encoding of speech production and artifactual aspects specific to ECoG data acquisition. To do this, we impose auxiliary losses on latent representations for articulatory movements, speech acoustics and session specific non-linearities.

Main results.—In three (out of four) participants reported here, Brain2Char achieves 10.6%, 8.5%, and 7.0% word error rates respectively on vocabulary sizes ranging from 1200 to 1900 words.

Significance.—These results establish a new *end-to-end approach* on decoding text from *brain signals* and demonstrate the potential of Brain2Char as a high-performance communication BCI.

Keywords

ECoG; convolutional neural network; regularization; BCI

[3]Author to whom any correspondence should be addressed. Edward.Chang@ucsf.edu.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

1. Introduction

Several demonstrations in recent years have shown that it is possible to decode cognitive, linguistic and speech representations directly from the brain through machine learning on neurophysiological imaging datasets. Attempts have been successful in decoding word classes or semantic representations from fMRI data (Mitchell *et al* 2008, Wehbe *et al* 2014, Huth *et al* 2016, Pereira *et al* 2018). However, since speech communication happens at a much faster rate, accurately decoding cortical function at the rate of fluent speech requires neurophysiological imaging at higher spatial resolution (in the order of millimeters) and temporal resolution (in the order of milliseconds). Among the modalities that offer the best resolution for assaying neural function is electrocorticography (ECoG, Sejnowski *et al* 2014, Chang, *et al* 2015). ECoG is an invasive neuroimaging technique where a flexible array of electrodes (3 in × 3 in) is placed directly on the surface of the cortex (or as depth electrodes to reach deeper structures) as part of clinical treatment for intractable epilepsy. Each electrode records the raw voltage potentials at the cortical surface, an aggregate electrical activity of thousands of neurons underneath the contact. The analytic amplitude of the High Gamma frequency band (70–150 Hz) of this aggregate activity has been shown to be a robust correlate of multi-unit spiking (Edwards *et al* 2005, Crone *et al* 2011). This setting provides a unique opportunity to create datasets of parallel neural and behavioral data as participants perform tasks such as listening or speaking naturally. Indeed, several key results have been published in recent literature on inferring speech representations directly from associated ECoG activity, broadly referred to here as neural speech recognition (NSR). We refer to NSR as decoding any aspect of spoken language from the brain—including produced or perceived speech (waveform) or words (lexical items), etc.

Prior works on speech/language decoding from ECoG used data either from the auditory or the speech motor cortices. Pasley *et al* (2012) and Akbari *et al* (2019) report methods for direct reconstruction of external speech stimuli from auditory cortex activations. Martin *et al* (2016) use non-linear SVM to classify auditory stimuli and Moses *et al* (2016) use Hidden Markov Modeling to infer continuous phoneme sequences from neural data during listening. Similarly, for neural decoding during speech production, Mugler *et al* (2018) have used linear models to classify phonemes and Herff *et al* (2015) use Hidden Markov Models (HMM)/Gaussian Mixture Models (GMM) based decoding to achieve a word error rate (WER) of 60% on a 50-word vocabulary task. Makin *et al* (2020), use a Machine-translation inspired sequence-to-sequence LSTM model to convert ECoG activity to complete sentences. Recent attempts have also focused on decoding audible speech directly from the sensorimotor cortex, including Angrick *et al* (2019), using deep convolutional architectures (Wavenet), Anumanchipalli *et al* (2019), using recurrent architectures.

In speech processing applications like automatic speech recognition (ASR) and text-to-speech synthesis, much progress has been made to achieve near-human performance on standard benchmarks. These include the use of recurrent (Hannun *et al* 2014) and convolutional architectures (Collobert *et al* 2016) towards end-to-end speech recognition minimizing connectionist temporal classification (CTC) loss, which remain unexplored for decoding text from brain data. Given the demonstration of naturalistic and continuous

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

speech synthesis directly from the brain, one way to approach the Brain-to-text problem is a two-stage model where neural data is converted to speech, which can then be converted to text using a state-of-the-art ASR system. Indeed, we propose a strong baseline (section 2.2) for the Brain-to-text problem along these lines, that to our knowledge, already achieves the best reported performance on this task. To improve this baseline further, we propose Brain2Char (section 2.3), a deep network architecture that borrows ideas from speech processing, and develop optimizations appropriate for NSR. Brain2Char implements an end-to-end network that jointly optimizes various sub-problems, like neural feature extraction, optimizing latent representations and session calibration through regularization via auxiliary loss functions. Performance of Brain2Char is quantified on four volunteer participants who spoke overtly or silently, and various aspects of decoder design are objectively evaluated (section 3).

1.1. Main challenges

1.1.1. Neural basis of speech production—Neural mechanisms for linguistic planning and execution occur at diverse timescales, and at diverse alignment offsets with respect to the speech signal. Also, different cortical regions encode distinct aspects of speaking. For example, the inferior frontal gyrus (IFG) is linked to motor sequence planning (Flinker *et al* 2015) and the articulator kinematic trajectories (AKTs) is linked to the articulatory aspects in producing speech (Mugler *et al* 2014, Chartier *et al* 2018). Each electrode location in the AKT codes a unique kinematic plan, spanning multiple vocal tract articulators and timescales to orchestrate continuous speech articulation. Electrodes in the superior temporal gyrus (STG) encode spectrotemporal aspects of speech signal (Mesgarani *et al* 2014). This diversity of cortical function and tuning properties contributes to most of the speech related variability in the ECoG signal. It is therefore important for an NSR system to model these representations, as appropriate for a given electrode location and participant behavior (e.g. whether it is a speaking task or a listening task).

1.1.2. Neural variability—Since individual brains differ anatomically, neural data from multiple participants cannot be pooled across speakers. This limits the amount of data available to train traditional machine learning models. Additionally, there are other sources of variability in neural recordings (Nuyujukian *et al* 2014, Zhang *et al* 2018). ECoG measures the local field potentials at the cortical surface. Since each electrode contact records from thousands of underlying neurons, the signal at each may not be entirely specific to speech production. Some cortical regions responsible for certain articulatory phonetic sequences may not even be sampled, given a particular electrode coverage, or there may be redundancy across neighboring electrodes. Furthermore, the intrinsic neural dynamics (Churchland *et al* 2010, Sun *et al* 2019) mask the true signal causal to producing speech. All of these aspects contribute to a poor signal-to-noise ratio in the neural data. It is important for NSR systems to extract meaningful features from across the diverse spatial and temporal scales of the ECoG data.

1.1.3. Cross-session variability—Typical ECoG data is collected in multiple short sessions spread over a week or so, where electrode leads from the brain are manually connected to a preamplifier before recording neural data for each session (Perge *et al*

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2013). An artifactual aspect of neural recordings is the non-stationarity in the signal across recording sessions. This may be due to different 'baseline' brain states in each session, or some electrode channels not recording (e.g. sensor losing contact with the cortical surface), or (more rarely) the entire grid slightly shifted on the cortical surface, causing differences across sessions that are unrelated to speech production process itself. An ideal NSR approach subsumes session specific corrections to improve cross session compatibility in neural data.

Since most of these issues cannot be deterministically modeled, it is necessary for NSR to be realized within a statistical framework, jointly optimizing neural feature extraction, latent representation learning, session calibration and text prediction, all within the same model. We now describe Brain2Char, a deep learning architecture that implements such a framework.

2. Methods

2.1. Data

In this study, we use data from four participants P1, P2, P3, and P4. These volunteer participants read prompted sentences on a screen while their speech and ECoG data were synchronously recorded. For participants P1 and P2, the sentences were derived from MOCHA-TIMIT corpus (a 1900-word vocabulary task of 460 independent sentences) and several short stories (a combined 1500-word vocabulary). For participants P3 and P4, a limited domain dataset of verbal descriptions of three pictures consists of a 400 words vocabulary on picture description and 1200 vocabulary on free-style interview tasks. The total data collected across participants varied between 120 min and 200 min. To evaluate and validate the neural variance of across trials, a subset of the sentences (i.e. MOCHA-TIMIT and picture description) was repeated across different task sessions. Other sentences (i.e. the interview, and stories) were read only once during the task. The recordings were made in several 1 h long sessions over a week or more, while participants were implanted with grids for clinical monitoring for seizure localization. Specifically, the neural data of participants P1, P2 and P4 are recorded with 16 × 16 electrode grid covering the ventral sensorimotor cortex (vSMC), IFG and STG, only participant P3 was recorded with 16 × 8 electrode grid covering only the dorsal half of the vSMC. All participants gave their informed consent to be a participant for this research prior to surgery. The research protocol was approved by the UCSF Committee on Human Research (RB# 10-03842). All neural data was pre-processed to reject artifacts and extract the analytic amplitude in the High Gamma frequency band (70–150 Hz) and low frequency component (0–40 Hz) z-scored appropriately. For the speech data collected, acoustic-to-articulatory inversion is performed to estimate the AKT, and in the meantime, Mel Frequency Cepstral Coefficients (MFCC) are extracted from acoustic speech signals. All data were synchronously sampled at 200 Hz.

2.2. Baseline NSR systems

We employed two baseline ECoG-to-text systems inspired by previous demonstrations of speech synthesis from the ECoG data (Angrick *et al*, 2019, Anumanchipalli *et al* 2019), and off-the-shelf ASR systems. The baselines first convert the ECoG activity into speech

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

acoustic features that are then converted to text using a state-of-the-art speech recognition system (*DeepSpeech*). The first baseline model *DS_0* consists of independently pre-trained two-BiLSTM layers ϕ_e that encode ECoG signals as acoustic features (i.e. MFCC), and pretrained DeepSpeech that translates these acoustic features to text. In the second baseline model *DS_1*, the two-layer BiLSTM encoder ϕ_e is jointly trained with a pre-trained *DeepSpeech* network. The training uses the available ECoG, MFCC, and text (i.e. character labels), and the joint network is updated to optimize a CTC loss on character sequence, and a weighted auxiliary mean-squared error loss on the MFCCs. We found it beneficial to 'freeze' the pre-trained layers of *DeepSpeech*, and only allow the two-BiLSTM neural encoder layers to be learned, while still optimizing the joint loss. The performance of the baseline systems is described in section 3.1, along with the proposed Brain2Char model.

2.3. Brain2Char architecture

To further improve upon the baseline architectures, we propose Brain2Char, an NSR framework with a modular architecture comprising three parts: the neural feature encoder, the text decoder, and the latent representation regularizer. Compared with previous work, the proposed Brain2Char translates neural inputs as character sequences instead of word sequences. The modular structure is convenient for network optimization, and each submodule can be independently improved based on the general design considerations of NSR systems mentioned in the earlier section. The inference model consists of the encoder and the decoder, and the regularization networks are only used at training time. Figure 1 illustrates the architecture of Brain2Char.

In comparison to Makin *et al* (2020) who use an end-to-end approach based on sequence-to-sequence decoding to translate neural signals to sentences, our proposed network employs CTC to map the abstracted neural features to phoneme level sequences, which on the one hand avoid the ambiguity of onset-offset neural speech alignment issue, on the other hand, provide the ability to track the neural dynamics with high temporal resolution (i.e. at the timescales of precise articulatory movements or phonemes instead of word sequences). Additionally, the regularization networks directly modulate the latent feature layer, which allows effectively introducing other features (e.g. AKT, time label, and MFCC) to reduce the search space drastically especially with limited and large variance training datasets.

2.3.1. Notation—Brain2Char optimizes a transformation ϕ to map the recorded neural signals $X = \{x_1, x_2, \ldots, x_n | x_i \in R^{t \times w \times h}\}$ to character sequences $Z = \{z_1, z_2, \ldots, z_n | z_i \in R^v\}$. Here the index t, w, h refer to time dimension, the anterior-posterior (width) and dorsal-ventral (height) axes of the ECoG grid, and v refers to the embedding dimension of the text vector. In the encoding phase, encoder ϕ_e projects neural inputs X to latent feature space F which in turn will be translated as the outputs Z by decoder ϕ_d in the decoding phase. Here F with the upper index (e.g. F^h) represents the basis of the latent feature space, whereas with the foot index (e.g. F_h) is a vector in the feature space.

2.3.2. Neural feature encoder network—The goal of the encoder is to extract the speech-specific part of the neural signal, robustly accounting for the spatial, temporal and spectral variations within the neural signals. In Brain2Char, the 3D ECoG signals X are

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fed into the encoder stacked by network modules similar to inception nets (Szegedy *et al* 2017). A single inception layer parallelly employs several sub-networks to extract features at different resolutions by choosing a set of hyperparameters specifying the number of channels, kernel size, and stacked layers. This design is capable of extracting features at various temporal-spatial scales within a single inception module and aggregate these multi-resolution features. Therefore, the next stage can access robust features at different resolutions simultaneously.

The neural feature encoder ϕ_e consists of several layers of grouped convolutional neural networks (CNN). For each single layer, a set of convolutional filters $\{W_1, W_2 \ldots\}$ are used in multiple sub-networks, where $W_i \in R^{c_{\text{in}} \times k_t \times k_w \times k_h \times c_{\text{out}}}$. For the weight tensor W_i, the spatial support of each kernel filter κ_i is $k_t \times k_w \times k_h$. There are c_{in} input channels and c_{out} output feature maps. Additionally, each sub-networks p in jth layer use stacked layers that employ a series of kernels κ with different sizes, that is $F_{\{\kappa\}} = W_{j,p,1}(W_{j,p,2}(\ldots(\Lambda)))$, where $F_{\{\kappa\}}$ refers to the features obtained at the pth sub-networks in jth layer of encoder by using kernel set $\{\kappa\}$. After a multi-scale feature extractor, two Bi-LSTM layers are stacked. In the proposed framework, the output of the Bi-LSTM layers are referred as the latent feature representation F_h.

2.3.3. Latent feature regularization network—To implicitly enforce a meaningful latent representation in the neural encoder, the regularization branch performs simple feed-forward transformations of the latent features of the encoder to account for known aspects of neural signal variance discussed earlier.

2.3.3.1. Session calibration.: To reduce the variability of neural signals X due to session specific artifacts, calibration across sessions is done to pool sessions in terms of their relative similarity to the earlier sessions. Since changes across sessions cannot be modeled systematically, in this study we introduce an implicit calibration on the encoded feature F to reduce intrinsic session-dependent variation. In other words, each latent feature $F_{t,t \in T_i}$ within the time duration T_i can be assigned session labels, which are indexed by continuous values. Instead of using one-hot vectors, we learned the time embedding vector Q_{T_i} in a *skip-gram* fashion (Mikolov *et al* 2013). Similar to the word embedding, Q_{T_i} describes the temporal correlation. A regression layer $M(\cdot)$ would map the dynamic latent sequence F_h to logits, and $L = -\sum Q * \log(M(F_h))$ is used as the cost function of time regularization.

2.3.3.2. Speech-specific latent representation.: The basis space F^x representing neural signals X exists in a higher dimensional space compared to basis space F^z with vocabulary Z. In Brain2Char architecture, the encoder projects the neural data X into the latent space F^h and the decoder expands the latent feature F_h to span the basis of targets Z. If trained without regularization on the output of ϕ_e, the encoder ϕ_e explores quite a large space searching for a manifold where $F^h \supset F^z$. To reduce the search space F^h, the speech/language basis F^z can be directly used as the regression target of encoder ϕ_e. However, limited by the data, complete speech/language basis F^z is not obtainable. As the baseline we proposed, other accessible feature representation $F^{z'}$ (e.g. MFCC), can be alternatively used as the regression targets

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of ϕ_e. In general, the regularization features $F^{z'}$ can be any type of features correlated with or generated from speech or text Z associated with productions. For instance, by applying autoencoder on speech acoustics, Akbari *et al* (2019) derive a low dimensional basis utilized as the regularization component in auditory speech reconstruction.

Assuming a set of feature basis $\{F^{z1}, F^{z2}, \ldots\}$, a joint feature basis is learned through the regularization networks as described in figure 1. Here the index z_i refers to the ith category features (e.g. MFCCs). Based on the obtained feature vectors F_{z_i}, the regularizations are illustrated as:

$$L(F_h, F_{z_1}, F_{z_2}, \ldots) = \sum_i \alpha_i \left\| \Omega(F_h) - F_{z_i} \right\|_2 \quad (1)$$

(where Ω represents projection operation (i.e. regularization layer), and L is cost function and α_i is the weight coefficient of the ith regularizer. By using Ω, a new ensemble feature basis is incorporated to modulate the latent representation. Physiologically generative representations or close features derived from speech acoustics make better targets for regularization.

2.3.4. Text decoder network—At the core, translating the latent feature F_h to character sequences is a sequence decoding task. In other words, any state-of-the-art sequence translation system can be adapted to the text decoder network. Brain2Char model employs three layers of dilated CNNs to process the long-short term correlations, which could be resistant to noise components. In each dilated layer, five sub-layers as shown in figure 1 are applied to learn the sequence correlations at various scales. The layer-wise residual connections ensure the features at different scales are processed simultaneously. Since the exact alignment of latent representations is hard to obtain (i.e. onset and offset of neural data corresponding to characters), on top of the dilated CNN, CTC is incorporated with a 4 gram language model as the text decoder network. Together with feature regularization and the explicit language model used for beam search, the cost function for the overall Brain2Char decoder can be illustrated as $L = \alpha L_1 + \sum_i \alpha_i L_i$, where the first component L_I refers to the loss of the inference networks (i.e. CTC loss), and the second component is the summation of the loss of regularization networks as described in equation (1).

2.4. Implementation

We implemented our method using Tensorflow. In terms of model training, we use a cyclic learning rate with maximal value 0.005 and minimal value 0.0001. Linear decay coefficients are applied to the weight coefficients of regularization components. The batch size is 50 at the sentence level for training (the time durations for each sentence range from 2 s to 5 s, and tail padded with 0), the feature dimension of MFCC and AKT are 26 and 33, respectively. For 3D inception module, the kernel sizes k_t, k_w, k_h along each axle are selected with different combinations of 3, 5, 7, 9. Bottleneck layer with kernel size 1 is also used. The dimension reduction is achieved by using two step strides in inception modules. For dilated CNN, the kernel size is fixed as 11, and the dilated ratios for five sub-layers are [1, 2, 4, 8, 16]. The convolutional layer after the dilated CNN uses kernel size 1, and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the output channel is the dimension of character vocabulary. Two BiLSTM layers use 0.5 dropout rate, and the output of dilated CNN is set to 0.15 dropout rate. To ensure robust sequence learning, the onset and offset of neural features are randomly jittered about a time window aligned to acoustic speech boundaries. A 4 gram language model incorporated with beam search is applied to the outputs of CTC. KenLM (Heafield *et al* 2013) is used to train the word language models of the speech corpus that is used for speaking tasks. Additionally, the pre-trained *LibriSpeech* language model used in DeepSpeech (Hannun *et al* 2014) is employed as a baseline language model. The weighting coefficient of the language model is set to 1.5.

Brain2Char was designed as a proof-of-principle demonstration on previously collected ECoG data, without explicit considerations for real-time latency. The choice of using Bidirectional RNNs and dilated CNNs require analysis windows that look-ahead (window centered around current frame) which introduce latency. In the Brain2Char model, the RNN expects 20 context frames by 5 ms per frame. For the dilated CNN, we use dilation ratio of 16 with kernel size 11, where the context size is 150 frames. In all, the context latency of Brain2Char model is 850 ms. The runtime complexity of the inference is 600 MMAC s^{-1} with 10 M parameters. On a NVIDIA T4 GPU server, for the off-line inference model implemented on an 8-core E5-2650 V2 server, the computational latency is less than 150 ms (averaged across ten trials).

3. Results

3.1. Quantitative results

We conducted a quantitative evaluation of the baselines and the proposed Brain2Char architecture. Figure 2(a) compares the performance of various systems with increasing amounts of training data. Systems indexed *DS_0* and *DS_1* are independently pretrained and jointly optimized baseline models described in section 2.2. Across three patients, the joint optimized *DS_1* systems show consistent gains of around 30% in WER over *DS_0*. This suggests that current neural speech synthesis methods do not give a sufficient quality of speech that can be reliably decoded by off-the-shelf speech recognition systems. Hence, customizing the intermediate representations in ϕ_e is critical, as is done in *DS_1* by optimizing the neural feature encoder with a CTC loss within the joint network. Figure 2(a) also shows the performance of the proposed Brain2Char networks (indexed B2C) against these baselines. In all cases, we see significant gains of an additional 30% in WER. This suggests that Brain2Char's modular architecture of neural feature encoder, feature space regularizer and text decoder is well suited for high performance NSR. The performance trends with increasing training data size also suggest that the architecture makes optimal utilization of the available data compared to the baseline (larger slope in WER gain with more data, whereas the baselines seem to plateau).

Figure 2(b) quantifies the contribution of language models towards Brain2Char Performance. For different amounts of training data and for three participants, three different language modeling conditions at inference-decoding with no language model (indexed *_NL*), the default language model in Deep-Speech (trained on *librispeech* corpus, a general purpose character-level language model of English, *_L1*), and a domain specific language

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

model (*_L2*) created using all training data specific to the task. It can be seen that language models generally improve decoding performance, and task dependent language models help further. In all, Brain2Char achieves roughly 10% to 25% WER improvement across three participants, and the performance improvements against *librispeech* based language model range from 3% up to 15%. It is to be noted that the performance is still respectable in the *_NL* conditions, possibly due to implicit language model in the text decoder. This may sometimes lead to overfitting if the decoder is simply memorizing the task language independently of the neural data. To explicitly test this, we ran inference on trials that were randomly cut-off, either at the start or the end. Figure 2(c) shows the error rates as a function of amount of signal cut-off (in seconds, either at the start, or at the end of a trial, indexed *_onset* and *_offset* in the legend). The results confirm that Brain2Char is not merely performing a classification task by memorizing sentences, but is sensitive to the length of the trial in time. It also seems that onset cut-off is worse than offset cut-off.

Additionally, the offset cut-off condition in figure 2(c) also shows that Brain2Char is capable of synchronous, incremental decoding (instead of waiting for whole sentence length neural data inputs that cause latency), which is a critical desirable of a real time communication brain–computer interface (BCI). Table 1 shows an example performance of Brain2Char on two sentences as data are provided in increments of 0.2 s at inference time. Note that the decoded sentences are shorter in length, as would be expected for shorter time windows of neural inputs, and the errors are typically in the last word(s) that maybe cut off mid-word.

3.2. Importance of regularization

One of the salient features of Brain2Char framework is the regularization branch that implicitly enforces a meaningful and robust latent representation in the neural encoder. To quantify the effect of various regularization factors used, we trained several systems each with different regularization strategies. Firstly, to study the effect of the session embedding regularization (calibration), we trained comparable systems where only in the calibrated condition, the latent representation F_h regresses to the attached time embedding. The imposed time constraints on F_h reduce cross-session neural variability, and the results in figure 3(a) confirm this trend, across increasing amounts of training data. In general, the session calibration enhanced the performance by about 4% against the non-calibrated approach.

The second regularization we evaluated was that of the latent speech representation in F_h. We built variants of Brain2Char systems, where F_h was unconstrained (no regularization), and added regularization branches from F_h to either (i) acoustic features (MFCC), (ii) articulatory kinematic trajectories (AKT) and (ii) MFCC + AKT. We observed improvements in all these cases compared to the case where no regularization was performed. Figure 3(b) summarizes these effects in terms of WER improvement from an unregularized Brain2Char system. While all speech representations result in positive gains, articulatory representations are significantly better regularization factors than the spectral MFCC representations. The best improvements were obtained using both representations (MFCC + AKT achieving a 15% absolute improvement in P2), as neural signals may explain some acoustic variations, complementary to the articulatory features. These results indicate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that implicitly enforcing physiological aspects in latent representations heavily contribute to explaining the neural speech variance, that cannot otherwise be learnt in an unsupervised fashion, given these smaller scale datasets. The benefits of the articulatory representations are also consistent with earlier studies about neural encoding in the vSMC (Chartier *et al* 2018).

In figure 3(b), different feature sets are used for regularization validation. The results show that all three feature sets *MFCC, AKT, MFCC* + *AKT* demonstrate significant improvements in NSR tasks. Specifically, the combination of MFCC and *AKT* achieves the best performance that enhances up to 15% WER in absolute on participant P2. The next best improvement was achieved on both participant P1 and P2 by only using *AKT*. Based on *MFCC*, 8% WER improvement was obtained across two participants. The results indicate that all physiological features contribute to explaining the neural speech variance, and especially AKT is considerably good at representing the neural speech variances.

3.3. Unseen sentences and importance of multiple repetitions

We conducted experiments where no sentences overlapped between training and testing sets (for instance, training on MOCHA-TIMIT and testing on fair story, and vice versa). By employing non-overlapped training/testing decoding tasks, we can explore the generalization of the proposed decoder on phoneme or word level recognition. Table 2 shows the number of words in training/test data for unique sentences. The training dataset for P1 includes both MOCHA-TIMIT and storytelling. For participant P2, it includes an interview. Both participant P3 and P4 include free-style picture descriptions as the training dataset. The vocabulary of test sets is averaged across ten-fold cross validation. The total words take into account all the tokens (i.e. words) in the training/testing dataset.

Due to considerably low overall word repeat ratio (defined as Unique words/Total words, less than four across four participants and large neural variance, the WERs are expectedly high. As shown in figure 4(a) (blue bar), the WERs range from 25% to 120% with an average of 80% across four participants. However, WER metric is not sufficient to evaluate extremely low resource word recognition tasks. Scrutiny of the decoder output reveals that the phonetic features are quite well presented (for instance, among our worst examples for P1 with a 100% WER is the sentence ‘Tina Turner is a pop singer’ decoded as ‘teashureasapopcargr’). Table 2 demonstrated some decoding results varied with different WER levels. The phonetic features are well translated compared to the ground truth. While the model has done a reasonable job of approximating the character string, the high WER is largely because the final text decoding module was not trained on sizable English text datasets to learn general pronunciation rules and legitimate word boundaries. Note that in this example, none of the test words were seen by Bran2Char in training.

To investigate this systematically, we computed the phonetic error rates (PERs) (Wagner and Fischer 1974) between the original and decoded sentences. figure 4(a) shows for all four participants, the WER (blue bar) and PER (orange bar) on unseen sentences in test (each trial is also plotted as a dot). We noticed that, across participants, the PERs (i.e. 66% on P1, 74% on P2, 65% on P3, 72% on P4) were much lower, and qualitatively many sentences were decoded quite phonetically, barring typos and mistaken word boundaries (see table 3

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

for representative examples). In other words, the decoder successfully learned the feature distribution of neural signals. However, the neural variability makes the phonetic n-gram statistics (i.e. $P(p_i|p_{i-k} \dots p_{i+k})$, where P(.) is the probability distribution with phoneme component p_i) ambiguous especially with limited training data, and the proposed decoder maps the neural signals of the characters to phonetic units with similar distributions. Note that the model was never trained on any phonetic transcriptions. We observe that unseen content words are a major contributor to high WERs, that would require Brain2Char to do zero-shot learning. To understand why sentence repetitions help, we trained models on P4 with increasing repetitions of test sentences. figure 4(b) shows that WER improvement (from 84% to 10%) was proportional to the PER improvement (from 72% to 9%), gradually improving word pronunciations. A classifier grossly mistakes one sentence to another in entirety, not in piecewise compositional units like we observe here.

Based on the decoding results on unseen sentences, we can conclude that articulatory context is strongly encoded in neural data, and NSR systems should be trained with datasets that densely cover all articulatory contexts of the language. In the absence of sizable datasets, training on several repetitions of valid sentences in the target domain can achieve reasonable gains by the trading off some generalizability.

4. Discussion

BCIs for communication have recently gained attention due to the novel invasive neurophysiology techniques and availability of datasets from intracranially implanted participants. The performance of best communication BCI is at eight words min^{-1} (Pandarinath *et al* 2017). This work uses multi-electrode arrays implanted in the hand-motor region of a paralyzed participant, decoding the movement of a cursor on the screen to navigate an alphabet keyboard. While still state-of-the-art, this performance is far below natural human speaking rates. There is a need for developing BCIs decoding directly from the speech motor cortex, which has the potential for much faster communication rates (Chang and Anumanchipalli 2020).

Since several nonlinear transformations underlie the conversion of vocal intent in the brain to vocal speech production, deep learning offers a natural solution to design a decoder for BCI. However, current deep learning methods are agnostic to the underlying generative processes and are not suited to be directly employed for BCI. In this work, we objectively show the advantages of inducing physiologically appropriate biases into the latent representations of deep neural networks. Yet, several limitations exist before reliable translation of this technology into prostheses for paralyzed populations. Our best results indicate that the performance of Brain2Char relies on multiple repetitions of expected sentences, which is not scalable to open vocabularies of the real world. Further development is needed to improve generalization of BCI techniques, including transfer learning across subjects, integrated language models, and pretrained robust neural feature extractors. Another critical milestone is to optimize Brain2Char for real time decoding with minimal latency. Some directions toward this include reducing the model complexity and network prediction latency.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Another limitation of current study is that all participants in this study are fluent American English speakers, with no neurological conditions affecting their ability to speak. To translate the decoding framework to paralyzed patients would require silent speech data acquisition, human-in-the-loop training protocols and optimal integration of low-latency feedback that can engage brain plasticity to enhance user embodiment and seamless adoption of the communication prostheses.

5. Conclusion

We propose Brain2Char, a neural network architecture that converts Brain recordings to text. Brain2Char utilizes multi-scale grouped CNN filters to extract neural signals from ECoG data, employs physiological and artifactual regularization schemes on latent representations, and decodes character sequences optimizing a CTC loss. The jointly optimized Brain2Char model makes optimal utilization of available data and sets a new state-of-the-art performance on decoding text from ECoG recordings. This holds both in terms of vocabulary sizes and performance metrics compared to earlier studies. Furthermore, Brain2Char is amenable to incremental, real-time decoding. These results demonstrate that Brain2Char is a promising candidate for a communication BCI.

Acknowledgments

This project was funded by a research contract under Facebook's Sponsored Academic Research Agreement. Data were collected and pre-processed by members of Chang lab, some (MOCHA-TIMIT) under NIH grant U01 NS098971. Some neural networks were trained using GPUs generously donated by the Nvidia Corporation.

References

Akbari H, Khalighinejad B, Herrero JL, Mehta AD and Mesgarani N 2019 Towards reconstructing intelligible speech from the human auditory cortex Sci. Rep 9 874 [PubMed: 30696881]

Angrick M, Herff C, Mugler E, Tate MC, Slutzky MW, Krusienski DJ and Schultz T 2019 Speech synthesis from ECoG using densely connected 3D convolutional neural networks J. Neural. Eng 16 036019 [PubMed: 30831567]

Anumanchipalli GK, Chartier J and Chang EF 2019 Speech synthesis from neural decoding of spoken sentences Nature 568 493–8 [PubMed: 31019317]

Chang EF 2015 Towards large-scale, human-based, mesoscopic neurotechnologies Neuron 86 68–78 [PubMed: 25856487]

Chang EF and Anumanchipalli GK 2020 Toward a speech neuroprosthesis JAMA 323 413–4 [PubMed: 31880768]

Chartier J, Anumanchipalli GK, Johnson K and Chang EF 2018 Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex Neuron 98 1042–54 [PubMed: 29779940]

Churchland MM et al. 2010 Stimulus onset quenches neural variability: a widespread cortical phenomenon Nat. Neurosci 13 369 [PubMed: 20173745]

Collobert R, Puhrsch C and Synnaeve G 2016 Wav2letter: an end-to-end convnet-based speech recognition system (arXiv: 1609.03193)

Crone NE, Korzeniewska A and Franaszczuk PJ 2011 Cortical gamma responses: searching high and low Int. J. Psychophysiol 79 9–15 [PubMed: 21081143]

Edwards E, Soltani M, Deouell LY, Berger MS and Knight RT 2005 High gamma activity in response to deviant auditory stimuli recorded directly from human cortex J. Neurophysiol 94 4269–80 [PubMed: 16093343]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Flinker A, Korzeniewska A, Shestyuk AY, Franaszczuk PJ, Dronkers NF, Knight RT and Crone NE 2015 Redefining the role of Broca's area in speech Proc. Natl Acad. Sci 112 2871–5 [PubMed: 25730850]

Hannun A et al. and 2014 Deep speech: scaling up end-to-end speech recognition (arXiv: 1412.5567)

Heafield K, Pouzyrevsky I, Clark JH and Koehn P 2013 Scalable modified Kneser-Ney language model estimation Proc. of the 51st Annual Meeting of the Association for Computational Linguistics vol 2 pp 690–696

Herff C, Heger D, De Pesters A, Telaar D, Brunner P, Schalk G and Schultz T 2015 Brain-to-text: decoding spoken phrases from phone representations in the brain Frontiers Neurosci. 9 217

Huth AG, de Heer WA, Griffiths TL, Theunissen FE and Gallant JL 2016 Natural speech reveals the semantic maps that tile human cerebral cortex Nature 532 453 [PubMed: 27121839]

Makin JG, Moses DA and Chang EF 2020 Machine translation of cortical activity to text with an encoder–decoder framework Nat. Neurosci 23 575–82 [PubMed: 32231340]

Martin S, Brunner P, Iturrate I, Millán JDR, Schalk G, Knight RT and Pasley BN 2016 Word pair classification during imagined speech using direct brain recordings Sci. Rep 6 25803 [PubMed: 27165452]

Mesgarani N, Cheung C, Johnson K and Chang EF 2014 Phonetic feature encoding in human superior temporal gyrus Science 343 1006–10 [PubMed: 24482117]

Mikolov T, Sutskever I, Chen K, Corrado GS and Dean J 2013 Distributed representations of words and phrases and their compositionality Adv. Neural Proc. Syst pp 3111–9

Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA and Just MA 2008 Predicting human brain activity associated with the meanings of nouns Science 320 1191–5 [PubMed: 18511683]

Moses DA, Mesgarani N, Leonard MK and Chang EF 2016 Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity J. Neural. Eng 13 056004 [PubMed: 27484713]

Mugler EM, Patton JL, Flint RD, Wright ZA, Schuele SU, Rosenow J, Shih JJ, Krusienski DJ and Slutzky MW 2014 Direct classification of all American English phonemes using signals from functional speech motor cortex J. Neural. Eng 11 035015 [PubMed: 24836588]

Mugler EM, Tate MC, Livescu K, Templer JW, Goldrick MA and Slutzky MW (2018) Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri J. Neurosci 38 9803–13 [PubMed: 30257858]

Nuyujukian P, Kao JC, Fan JM, Stavisky SD, Ryu SI and Shenoy KV 2014 Performance sustaining intracortical neural prostheses J. Neural. Eng 11 066003 [PubMed: 25307561]

Pandarinath C, Nuyujukian P, Blabe CH, Sorice BL, Saab J, Willett FR, Hochberg LR, Shenoy KV and Henderson JM 2017 High performance communication by people with paralysis using an intracortical brain-computer interface Elife 6 e18554 [PubMed: 28220753]

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT and Chang EF 2012 Reconstructing speech from human auditory cortex PLoS Biol. 10 e1001251 [PubMed: 22303281]

Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M and Fedorenko E 2018 Toward a universal decoder of linguistic meaning from brain activation Nat. Commun 9 963 [PubMed: 29511192]

Perge JA, Homer ML, Malik WQ, Cash S, Eskandar E, Friehs G, Donoghue JP and Hochberg LR 2013 Intra-day signal instabilities affect decoding performance in an intracortical neural interface system J. Neural. Eng 10 036004 [PubMed: 23574741]

Sejnowski TJ, Churchland PS and Movshon JA 2014 Putting big data to good use in neuroscience Nat. Neurosci 17 1440 [PubMed: 25349909]

Sun P, Moses DA and Chang EF 2019 Modeling neural dynamics during speech production using a state space variational autoencoder (arXiv: 1901.04024)

Szegedy C, Ioffe S, Vanhoucke V and Alemi AA 2017 Inception-v4, inception-resnet and the impact of residual connections on learning Thirty-First AAAI Conf. on Artificial Intelligence 4278–84

Wagner RA and Fischer MJ 1974 The string-to-string correction problem J. Acm 21 168–73

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Wehbe L, Murphy B, Talukdar P, Fyshe A, Ramdas A and Mitchell T 2014 Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses PloS One 9 e112575 [PubMed: 25426840]

Zhang P, Ma X, Chen L, Zhou J, Wang C, Li W and He J 2018 Decoder calibration with ultra small current sample set for intracortical brain–machine interface J. Neural. Eng 15 026019 [PubMed: 29343650]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

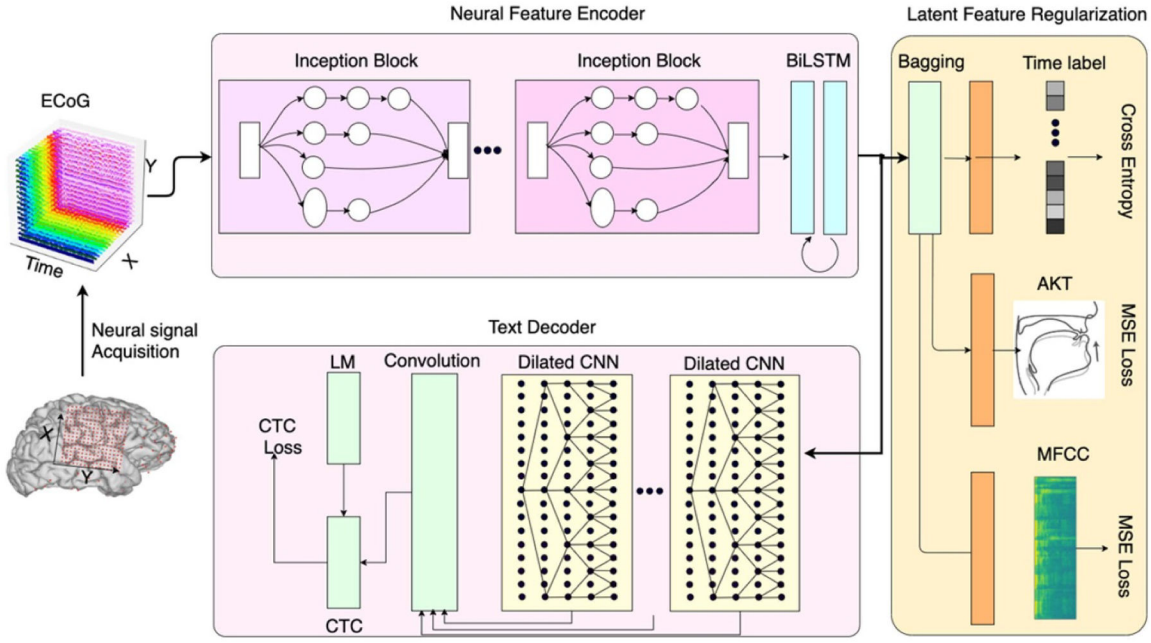

Figure 1.

Architecture of Brain2Char: neural data is recorded as participants produce speech. Different filters spanning across space, time, and frequency dimensions convert recorded potentials into appropriate feature representations. The intermediate features are fed into the regularization network and the decoder network. The regularization networks impose Mean Square Error (MSE) losses on regressed speech representations and session embedding. The decoder implements a sequence learning model that dilated CNNs convert the latent representations to character sequences, using language model weighted beam search.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

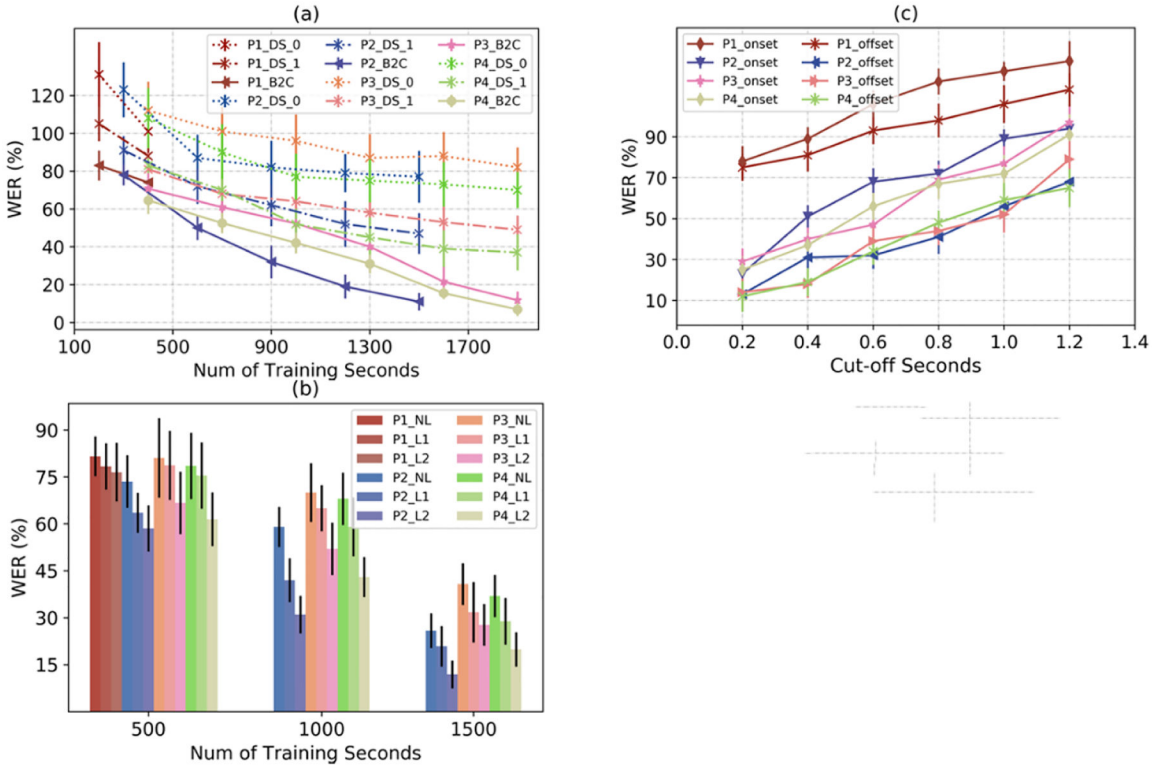

Figure 2.

Performance evaluation of Brain2Char compared to baseline systems. (a) Word error rate as a function of increasing amount of training data for baselines and Brain2Char. (b) Comparison of performance of various language models. (c) Performance on partial neural data with cut-off at either onset or offset.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

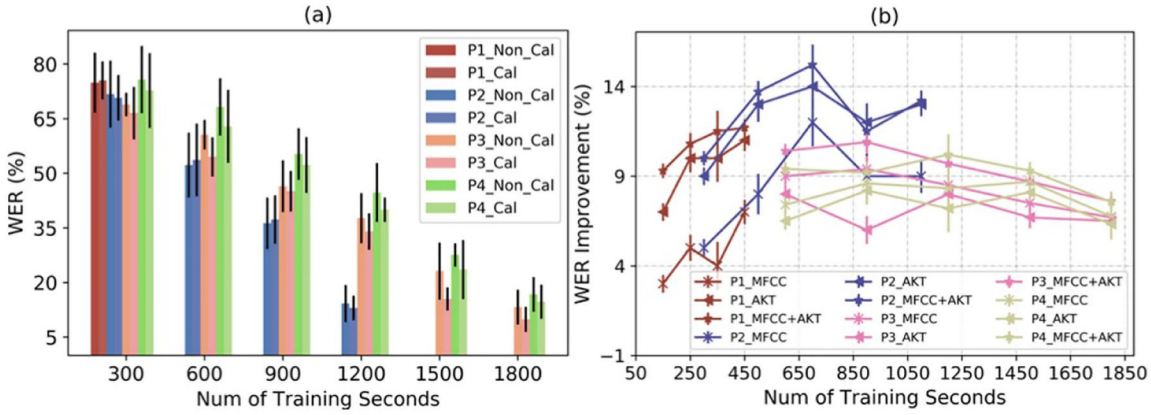

Figure 3.

Importance of regularization factors in Brain2Char: (a) Effect of session calibration on two participants. (b) Word error rate gains by imposing physiological and/or acoustic feature targets (e.g. MFCC and AKT).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

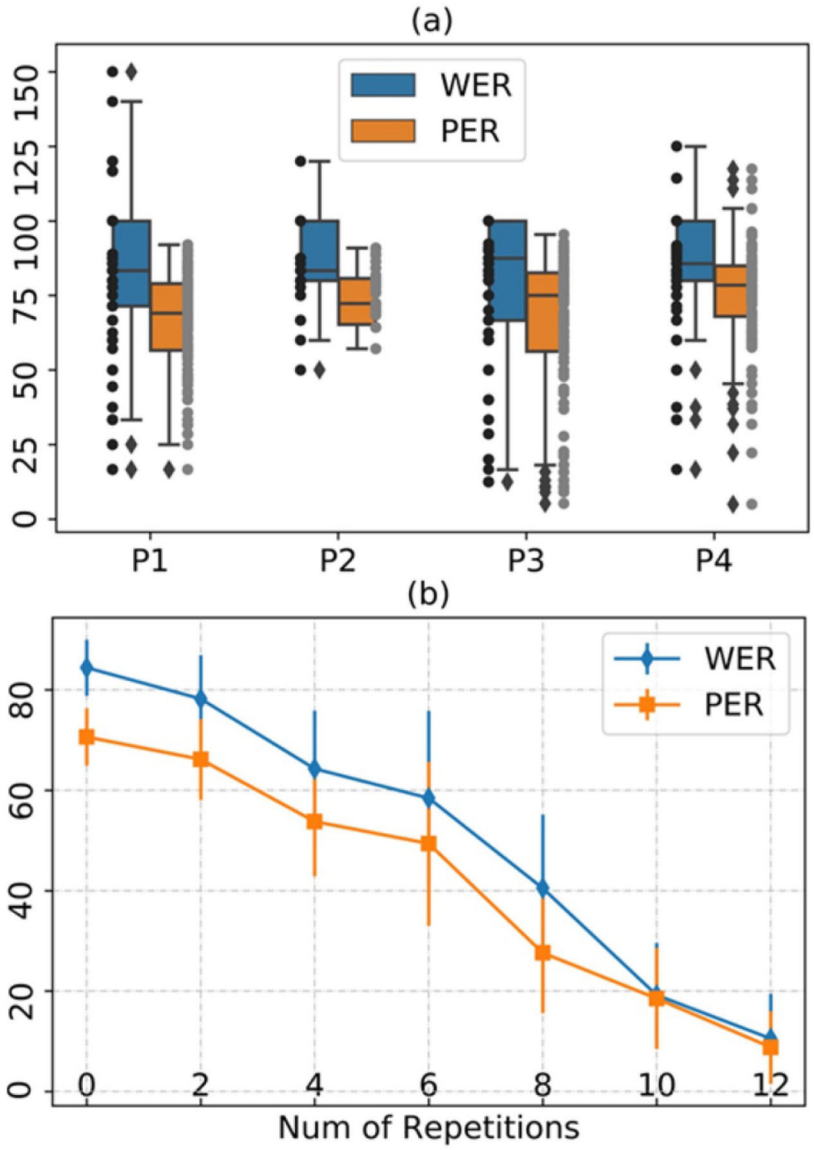

Figure 4.

(a) WER and PER across four participants (P1, P2, P3, P4). (b) WER and PER for P4 with incremental repeats.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1.

Illustration of incremental decoding by Brain2Char.

	if only the mother could pay attention to her children	**i think their water bill will be high**
0.2 s	His	in
0.4 s	if only to	i think the
0.6 s	if only the mother a	i think their water
0.8 s	if only the mother could pay atendion	i think their water bill
1.0 s	if only the mother could pay attention with	i think their water bill will be
1.2 s	if only the mother could pay attention to her children	i think their water bill will be high

J Neural Eng. Author manuscript; available in PMC 2022 October 24.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Number of words across participants.

	P1	P2	P3	P4
Unique words	2683	1806	2101	2201
Total words	14 978	6015	8520	5811
Test words	312	316	525	373

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3.

Phonetically approximate decoding of unseen sentences by Brain2Char.

Ground Truth	Decoded	WER/PER
help greg to pick a peck of potatoes	help pbrigt prinke atary oapcatiaous	87.5/64
will robin wear a yellow lily	will capuworenevaleinly	83/65
critical equipment needs proper maintenance	critical oubuemenes fpoaprinens	80/54
there is chaos in the kitchen	there is chosinteekitchen	66/33
a dog is barking at the man in the tree	dog is barking at teetree	60/42
part of the cake was eaten by the dog	part of the cake waednboythedog	55/24
while falling the boy grabs a cookie	the foiling the boy trat a cookie	43/31
the guests arrived with presents	the guesarried with presents	40/33
the little girl is giggling	the eitll girl is gigling	40/21

J Neural Eng. Author manuscript; available in PMC 2022 October 24.