HHS Public Access

Author manuscript

J Neural Eng. Author manuscript; available in PMC 2020 June 01.

Published in final edited form as:

J Neural Eng. 2019 June ; 16(3): 036019. doi:10.1088/1741-2552/ab0c59.

Speech Synthesis from ECoG using Densely Connected 3D Convolutional Neural Networks

Miguel Angrick[1,‡], **Christian Herff**[1,2,‡], **Emily Mugler**[3], **Matthew C. Tate**[4], **Marc W. Slutzky**[3,5], **Dean J. Krusienski**[6], **Tanja Schultz**[1]

[1]Cognitive Systems Lab, University of Bremen, Germany [2]School of Mental Health and Neuroscience, Maastricht University, The Netherlands [3]Dept. of Neurology, Northwestern University, Chicago, Illinois, USA [4]Dept. of Neurological Surgery, Northwestern University, Chicago, Illinois, USA [5]Depts. of Physiology, Biomedical Engineering and Physical Medicine & Rehabilitation, Northwestern University, Chicago, Illinois, USA [6]Dept. of Biomedical Engineering, Virginia Commonwealth University, Richmond, Virginia, USA

Abstract

Objective—Direct synthesis of speech from neural signals could provide a fast and natural way of communication to people with neurological diseases. Invasively-measured brain activity (electrocorticography; ECoG) supplies the necessary temporal and spatial resolution to decode fast and complex processes such as speech production. A number of impressive advances in speech decoding using neural signals have been achieved in recent years, but the complex dynamics are still not fully understood. However, it is unlikely that simple linear models can capture the relation between neural activity and continuous spoken speech.

Approach—Here we show that deep neural networks can be used to map ECoG from speech production areas onto an intermediate representation of speech (logMel spectrogram). The proposed method uses a densely connected convolutional neural network topology which is well-suited to work with the small amount of data available from each participant.

Main results—In a study with six participants, we achieved correlations up to $r = 0.69$ between the reconstructed and original logMel spectrograms. We transfered our prediction back into an audible waveform by applying a Wavenet vocoder. The vocoder was conditioned on logMel features that harnessed a much larger, pre-existing data corpus to provide the most natural acoustic output.

Significance—To the best of our knowledge, this is the first time that high-quality speech has been reconstructed from neural recordings during speech production using deep neural networks.

Keywords

Speech Synthesis; Neural Networks; Wavenet; Electrocorticography; BCI; Brain-Computer Interfaces

miguel.angrick@uni-bremen.de, c.herff@maastrichtuniversity.nl.

‡These authors contributed equally

1. Introduction

The ability to speak is crucial for our daily interaction. However, a number of diseases and disorders can result in a loss of this ability, for example brainstem stroke, cerebral palsy [1], amyotrophic lateral sclerosis [2, 3] and laryngeal cancer [4]. Various technologies have been proposed to restore the ability to speak or to provide a means of communication (see [5] for a review), including Brain Compuer Interfaces (BCIs, [6]). The most natural approach for BCIs would be to directly decode brain processes associated with speech [7]. Invasively-measured brain activity, such as ECoG is particularly well-suited for the decoding of speech processes [8, 9, 10, 11], as it balances a very high spatial and temporal resolution with broad coverage of the cortex.

In recent years, significant progress has been made in the decoding of speech processes from intracranial signals. The spatio-temporal dynamics of word retrieval during speech production was shown in [12]. Other approaches demonstrated that speech can be decoded from invasively measured brain activity such as words [13], phonemes [14, 15], phonetic features [16, 17], articulatory gestures [18] and continuous sentences [19, 20].

These great advances would require a two-step approach to restore a natural conversation, where in the first step neural signals are transformed into a corresponding textual representations via a speech recognition system and in the second step this text is transformed into audible speech via text-to-speech synthesis.

However, using the textual representation as pivot has several disadvantages [5]: (1) recognition / classification errors are propagated to the downstream processes, i.e. if the speech recognition system produces wrong results, the speech synthesis will be wrong as well, (2) since the recognition process needs to take place prior to resynthesis, the two-step approach is too time-consuming (>50ms [21, 22]) for real-time scenarios where immediate audio feedback is needed, and (3) speech carries information about prosody, emphasis, emotion etc. which are lost once transformed into text. For these reasons, it would be desirable for spoken communication to directly synthesize an audible speech waveform from the neural recordings [23].

In [24], Santoro et al. reconstructed the spectro-temporal modulations of real-life sounds from fMRI response patterns. Kubanek et al. [25] were able to track the speech envelope from ECoG. Neural signatures of speech prosody in receptive [26] and productive [27] speech cortices have been described. Studies were able to reconstruct perceived speech from ECoG signals [28] and to reconstruct spectral dynamics of speech from ECoG [29]. In a pilot study, we showed that it was possible to reconstruct an audio waveform from ECoG signals during speech production [30], but to the best of our knowledge, no study has reconstructed a high-quality audio waveform of produced speech from ECoG using deep neural networks.

Neural networks have shown great success in speech recognition [31] and speech synthesis [32]. However, so far they have been used only rarely for brain recordings. This is partially due to the very limited amount of speech data available for individual participants and the requirement to train speaker dependent models. In traditional speech processing, speaker

dependent systems use dozens of hours of data for training while high performing speaker independent systems accumulate thousands of hours.

More recently, initial studies have successfully applied deep learning methods to brain data [33, 34, 35, 36] and BCI applications [37, 38, 39, 40, 41, 42]. Here, we show that densely-connected convolutional neural networks can be trained on limited training data to map ECoG dynamics directly to a speech spectrogram. This densely connected network architecture is specifically tailored to cope with the small amount of data. We then use a Wavenet vocoder [43] conditioned on logMel features to transform the reconstructed logMel spectrogram to an audio waveform. In this study, we restrict the vocoder to the conversion from an encoded speech representation to an audible wavefile. The Wavenet vocoder uses a much larger training corpus to map the logMel spectrograms directly onto an acoustic speech signal.

Recent BCI studies [44, 40, 45] apply deep neural networks on raw signals to avoid their dependency on hand-crafted features by learning low-level features in the front layers and high-level patterns at later stages. However, feature extraction on raw signals has to be trained by the network and requires a sufficient amount of data to optimize the additional parameters. In order to keep the number of parameters inside the densely-connected convolutional neural networks small, we have used specific hand-crafted features to extract information in the high-gamma band. Previous studies have shown that these features are suitable for the transformation into a textual representation [19] and spectral coefficients [30].

The resulting audio is of high quality and reconstructed words are often intelligible. Thus, we present the first deep neural network reconstructing high-quality audio from neural signals during speech production.

2. Material and Methods

2.1. Experiment and Data Recording

We recorded ECoG from six native English speaking participants while they underwent awake craniotomies for brain tumor resection. All participants gave informed consent and the study was approved by the Institutional Review Board at Northwestern University. All subjects had normal speech and language function and normal hearing. ECoG was recorded with a medium-density, 64-channel, 8 × 8 electrode grid (Integra, 4 mm spacing) placed over the ventral motor cortex (M1v), premotor cortex (PMv) and inferior frontal gyrus pars opercularis (IFG). Grid locations were determined using anatomical landmarks and direct cortical stimulation to confirm coverage of speech articulatory sites. ECoG recordings were sampled at 2 kHz with Neuroport data acquisition system (Blackrock Microsystems, Inc.).

Participants were asked to read aloud single words shown to them on a screen while we recorded ECoG signals. These words were predominantly monosyllabic and consonant-vowel-consonant, and mostly compiled from the Modified Rhyme Test [46]. Participants read between 244 and 372 words resulting in recording length between 8.5 and 12.7 minutes. Note the extremely limited amount of subject dependent data compared to

traditional speech synthesis. Acoustic speech was recorded using a unidirectional lapel microphone (Sennheiser) and sampled at 48 kHz. Figure 1 visualizes our experimental procedure. Stimulus presentation and simultaneous recording were facilitated using BCI2000 [47].

2.2. Data Processing

We first applied linear detrending to the ECoG signals and attenuated the first harmonic of the 60 Hz line noise using elliptic IIR notch filter (118–122 Hz, filter order 13). To extract meaningful information from the ECoG signals, we calculated logarithmic power in the broadband-gamma frequency range (70–170 Hz, filter order 13 and 14, respectively), which is thought to largely reflect ensemble spiking [48] and contains localized information about movement and speech processes [49]. Broadband-gamma power was extracted in 50 ms windows with an overlap of 10 ms. This time interval was chosen short enough to capture the fast processes of speech production while simultaneously being long enough to estimate broadband-gamma reliably. To estimate signal power, we calculated the mean of the squared signal in each window. We applied a logarithm transform to the extracted broadband-gamma features to make their distribution more Gaussian.

We normalized broadband-gamma activity of each electrode individually to zero mean and unit variance. To capture the long-range dependencies of the speech production process [50, 51], we included neighboring broadband-gamma activity up to 200 ms into the past and future in our study, resulting in 9 temporal contexts being stacked to form our final features. This results in a feature space of size 64 *electrodes* × 9 *temporal offsets*. To represent the spatial topography of the electrode array, we arranged our features as a 8 × 8 × 9 matrix for decoding each time window.

For processing the acoustic data, we first downsampled the speech signal to 16 kHz. We transformed the waveform data into the spectral domain using 50 ms windows with an overlap of 10 ms to maintain alignment with the neural data. We discarded the phase information and only used the magnitude of the spectrograms [29, 52, 28]. To compress the magnitude spectrograms, we used 40 logarithmic mel-scaled spectral bins which should better represent the acoustic information in the speech data [53]. The logarithmic mel-scaled spectrograms (logMels) are extracted by taking the magnitude spectrogram and mapping it onto the mel-scale using triangular filter banks. The compression is based on the usage of the mel-scale which condenses frequency bands in accordance to the human perception [53]. From now on we refer to the logMel representation as the spectrogram.

2.3. Decoding Approach

To transform the recorded neural signals into an audio waveform, we first trained densely connected convolutional neural networks (DenseNets) [54] to map the spatio-temporal broadband-gamma activity onto logarithmic mel-scaled spectrograms. This was done for each participant individually, as electrode grid placements and underlying brain topologies are vastly different between participants. The DenseNet regression model is described in Section 2.4.

We then used a Wavenet vocoder conditioned on the same spectral features of speech to recreate an audio waveform from the reconstructions of our densely connected convolutional neural network. For this Wavenet vocoder, a much larger data corpus could be used, as no user specific mapping had to be learned. Section 2.5 describes the Wavenet vocoder in more detail. Figure 2 highlights our decoding approach. The broadband-gamma activity over time (purple) is fed into the DenseNet regression model to reconstruct spectral features of speech (yellow). These are then transformed into an audio waveform using the Wavenet vocoder.

2.4. DenseNet Regression Model

The DenseNet architecture is a feed-forward multilayer network which uses additional paths between earlier and later layers in a dense structure. Each layer receives all feature maps of previous layers as an input and propagates its own feature maps to all subsequent layers. By passing the feature maps to subsequent layers, shortcut connections are established which improve the gradient and information flow in the training. This behavior is similar to the ResNet[55] architecture, where an identity function is used to circumvent the degradation problem. DenseNets were previously found to require fewer parameters to achieve a similar performance as ResNet [54]. ResNet's solution relies on a residual mapping which combines the input and the output of a layer sequence by an addition operation. In contrast, DenseNet concatenates the feature maps of preceding layers. This way, each convolution adds its local information to the collective knowledge in the network and therewith forms an implicit global state. In a classification task, the final softmax layer can be seen as a classifier which takes the output of the previous layer and all preceding feature maps under consideration for its prediction.

A few adjustments needed to be done to adapt DenseNet for our regression task. On the one hand, it is intended for the convolution operations to apply on all three dimensions, namely the x and y position within the electrode grid and *time,* to find pattern in the spatio-temporal space. We therefore used 3D convolutions as well as 3D pooling layers throughout the network instead of their 2D counterparts used in traditional image processing. On the other hand, we changed the output layer to a fully connected layer with 40 neurons and a linear activation function to create a continuous output for the spectral coefficients.

Figure 3 shows an overview of the network structure and its integration inside the synthesis pipeline. The architecture consists of three dense blocks which group together a sequence of sublayers. Each sublayer is a composition of the following consecutive functions: Batch Normalization, rectified linear unit (ReLU) and a 3×3×3 convolution. The number of feature maps is set initially to 20 and increases according to a growth rate of $k = 10$. Two sublayers were used in each dense block yielding a total amount of 80 feature maps. Dense blocks are connected through transition layers, which are used as a downsampling operation for the feature maps of preceding layers, to fit their dimensions for the input of the next block. The output layer is a regressor, which estimates the spectral coefficients based on the input data and the feature maps in the collective knowledge. Overall, our resulting network comprises around 83,000 trainable parameters. For the training procedure, we used Adam[56] to minimize the mean squared error loss and used a fixed number of 80 epochs.

2.5. Wavenet Vocoder for the Reconstruction of Audible Waveforms

In the transformation process of acoustic data into logarithmic mel-scaled coefficients, we discard phase information. Nevertheless, this information is crucial for the inversion from the frequency domain back into a temporal signal. Recent studies show that high quality speech waveforms can be synthesized by using Wavenet [57] conditioned on acoustic features estimated from a mel-cepstrum vocoder [58]. During network training, the model learns the link between speech signal and its acoustic features automatically without making any assumptions about prior knowledge of speech. The synthesized acoustic waveform generated by Wavenet recovers the phase information previously lost.

In this paper, we conditioned Wavenet in the same way on logarithmic mel-scaled coefficients as described in Tacotron-2 [43] for Text-to-Speech synthesis. The filterbank features tend to outperform cepstral coefficients for local conditioning [59]. We used a separate dataset for training our Wavenet model. The LJ-Speech corpus [60] contains utterances from a single speaker reading passages from seven non-fictional books. The length of an utterance varies between one and ten seconds and sums up to a total amount of around 24h of speech data. Empirical tests based on the spectral coefficients of the reference data showed that this corpus is suitable to train Wavenet and reconstruct high quality acoustic waveforms containing intelligible speech and capturing the speaker characteristics of the participants.

The internal architecture is depicted in Figure 4. The model expects two feature matrices with possibly differing dimensions as inputs: the acoustic speech waveform **x** and the spectral features **c** for the local conditioning. Due to the fact that the dimensions of both input data might not match, a transformation is needed as an adjustment. An initial 1×1 convolution is used to increase the number of channels, known as residual channels. The spectral features get upsampled by four consecutive transpose convolutions to match the dimensions with the convolved acoustic speech signal.

After adjusting both input sequences, a stack of residual blocks is used whose interior is illustrated inside Figure 4. Each block contains a gated activation function which calculates a hidden state z given the following equations:

$$F^{(j)} = W_f^{(j)} * x^{(j-1)} + V_f^{(j)} * c \tag{1}$$

$$G^{(j)} = W_g^{(j)} * x^{(j-1)} + V_g^{(j)} * c \tag{2}$$

$$z^{(j)} = \tanh\left(F^{(j)}\right) \odot \sigma\left(G^{(j)}\right), \tag{3}$$

where * implies a dilated causal convolution, $\odot$ denotes the hadamard product and $\sigma(\cdot)$ corresponds to the sigmoid function. The superscript j indicates the current residual block. W_f, W_g, V_f and V_g are trainable weights of the convolution. In the gated activation function, the equations F and G indicate the filter and gate, respectively. A residual block uses two outputs to speed up the training procedure of the network. Both outputs are based on the

intermediate result of the hidden state z. The residual blocks are connected throughout their stack by using residual connections [55] which enable training of very deep neural networks. For the prediction of the next audio sample, the network uses skip connections. Both outputs are computed in the following way:

$$x^{(j)} = R^{(j)} * z^{(j)} + x^{(j-1)} \tag{4}$$

$$s^{(j)} = S^{(j)} * z^{(j)}, \tag{5}$$

where * denotes a 1 × 1 convolution to adjust the dimensionality of the channels. R and S represent trainable weights of their convolution operation.

In the stack of residual blocks, the dilation rate of the dilated causal convolution increases exponentially and gets reset after a fixed amounts of layers to start a new cycle. Parameter choices have been made in accordance to the Tacotron-2 System [43] which results in a stack of 30 residual blocks and three dilation cycles.

Wavenet is an autoregressive model which predicts the next audio sample based on all previously seen samples inside its receptive field and its conditioning features:

$$P\left(y \mid c,\ x_{i-1},\ \ldots,\ x_{i-R}\right) \tag{6}$$

For the prediction of the next audio sample, the model considers all skip connections from the residual blocks by summation and processes the results through a sequence of rectified linear units and 1 × 1 convolutions as shown in Figure 4.

Recent improvements suggest to model the generated speech with a discretized mixture of logistic distributions [61, 62]. We follow the proposed approach from Tacotron 2 and use a linear projection layer to predict the parameters for each mixture component [43]. The Wavenet vocoder uses a 10-component mixture of logistic distributions to model the reconstructed speech signal.

In network training, we use Adam as our optimization method with an initial leaning rate of 0.001. We trained our Wavenet vocoder for a fixed amount of 600,000 update steps. All hyperparameter choices are summarized in Table 1. We based our Wavenet vocoder on an open source implementation available on GitHub [63].

3. Results

For each participant, we partitioned the recorded data into disjoint sets for training and testing in a 5-fold cross validation. This approach ensured that the complete spectrogram could be reconstructed without operating on data that has been seen before. Twenty percent of the training data are reserved as a separate validation set to analyze the optimization process of the network training. The reconstruction from DenseNet is frame based and yields a spectrogram of the same length as the original one.

We used the Pearson product-moment correlation for the evaluation of our synthesized speech. The correlations are computed for each frequency bin individually between the reconstructed spectrogram and its original counterpart. For comparison, we computed a chance level by splitting the data at a random point in time and swapping both partitions. This ensures that the structure of the neural signals is preserved but the alignment to the spectral representation of speech is shifted. To reconstruct a spectrogram based on broken alignment, we performed additional network training in a 5-fold cross validation with the randomized dataset. For each participant, we repeated the estimation of a random chance level 60 times by using unique splits along the temporal dimension to approximate chance level.

Figure 5 (a) summarizes our results for all participants showing average correlations and their corresponding standard deviation. We achieve correlations significantly better than chance level ($p<0.001$) for all six participants with scores (mean ± standard deviation) of $r_1 = 0.19 \pm 0.12$, $r_2 = 0.29 \pm 0.06$, $r_3 = 0.56 \pm 0.18$, $r_4 = 0.34 \pm 0.12$, $r_5 = 0.69 \pm 0.10$ and $r_6 = 0.41 \pm 0.07$, respectively. Participant 5 clearly outperforms the other participants. We hypothesize that this might be due to having electrodes placed in a better position to capture articulator-related information. Across all participants, correlations of reconstructed spectrograms are considerably above chance level.

In Figure 5 (b) we investigate the distribution of Pearson correlation coefficients for Participant 5 in more detail. For each spectral bin, the chance level is given by the mean correlation coefficient under consideration of all 60 spectrograms from the baseline approximation. Our decoding approach achieves consistently high correlations above chance level across all spectral bins.

In order to evaluate the intelligibility of our reconstructed speech waveform, we employed the Short-Time Objective Intelligibilty measure (STOI) [64]. We used the original audio from the recording of the experiment as our reference. Figure 5 (c) reports the STOI scores for each participant. For the chance level, we took the randomized spectrogram that was better than 95% of the randomization to estimate a waveform using the Wavenet vocoder.

An example reconstruction of an excerpt from the experiment session of participant 5 is shown in Figure 6 (a) for visual inspection. The top row corresponds to the spectrogram of the reference data while the bottom row contains the time aligned spectral coefficients estimated by the DenseNet model. It is evident that the model has learned a distinguishable representation between silence and acoustic speech and captures many of the intricate dynamics of human speech. Furthermore, early characteristics of resonance frequencies are present in the spectral coefficients of the predicted word articulation.

Figure 6 (b) shows the resynthesized acoustic waveforms of the same excerpt from the reconstruction example by using the described Wavenet vocoder. Additionally, some listening examples of original and reconstructed audio can be found in the supplementary material. To compensate for artifacts of this conversion, we applied the transformation on the original and reconstructed spectrograms to isolate the synthesis quality of the trained network.

4. Discussion

We have shown that speech audio can be decoded from ECoG signals recorded from brain areas associated with speech production. We achieve this by combining two deep neural network topologies specifically designed for very different goals. In the first step, we employ a densely connected convolutional neural network which is well-suited to be trained on the extremely limited datasets. This network transforms the measured brain activity to spectral features of speech. Correlations of up to $r = 0.69$ across all frequency bands were achieved by this network. Subsequently, a Wavenet vocoder is utilized to map these spectral features of speech back onto an audio waveform. As this model does not have to be trained for each participant individually, it is trained on a much larger speech dataset and the topology is tailored to maximize speech output quality. Selected examples of reconstructed audio can be found in the supplementary material in which the original articulation of participant 5 and the corresponding reconstruction is presented in pairs.

To the best of our knowledge, this is the first time that high quality audio of speech has been reconstructed from neural recordings of speech production using deep neural networks. This is especially impressive considering the very small amount of training data available. In general, traditional speaker dependent speech processing models are trained with dozens of hours of data. This is an important step towards neural speech prostheses for speech impaired users. However, in its current state, the proposed method is based on an offline analysis. Due to its computational complexity, it is not yet suitable for a real-time application. The integration of our approach into a closed-loop system is limited by two drawbacks. First, the optimization of the DenseNet regression model described in Section 2.4 is time-consuming and exceeds the recording time of a typical experimental session. Second, the transformation from the logMel spectrogram into audible audio using the Wavenet vocoder is in its current implementation not able to generate an acoustic signal in real-time.

In the future, further investigation is needed to improve training and decoding time to make it feasable to implement in a real-time scenario. Such a real-time performance is necessary for the embedding in a closed-loop system. Various strategies such as parallel training and model averaging have demonstrated reduced training time [65] and could be applied to DenseNet model training. Considering the decoding phase, recent studies using the Wavenet architecture have shown promising results that a parallel feed-forward network [62] or a recurrent neural network [66] can be trained for high-fidelity audio generation.

While our study uses overtly-produced speech in motor-intact participants, there are recent advances in decoding motor function [67, 68, 69, 70] in paralyzed patients that might be extended to the decoding of speech from motor areas in the future. Initial studies in the decoding of attempted [71] and imagined speech [72] emphasize this point.

Our approach is a first step towards communication for speech impaired users, but more research is needed to understand speech representations in paralyzed patients. In summary, we show for the first time how two specialized deep learning approaches can be used to reconstruct high-quality speech from intracranial recordings during speech production.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

MA, CH, DK and TS acknowledge funding by BMBF (01GQ1602) and NSF (1608140) as part of the NSF/NIH/BMBF Collaborative Research in Computational Neuroscience Program. MS acknowledges funding by the Doris Duke Charitable Foundation (Clinical Scientist Development Award, grant 2011039), a Northwestern Memorial Foundation Dixon Translational Research Award (including partial funding from NIH National Center for Advancing Translational Sciences, UL1TR000150 and UL1TR001422), NIH grants F32DC015708 and R01NS094748.

References

[1]. Pirila S, van der Meere J, Pentikainen T, Ruusu-Niemi P, Korpela R, Kilpinen J, and Nieminen P, "Language and motor speech skills in children with cerebral palsy," Journal of communication disorders, vol. 40, no. 2, pp. 116–128, 2007. [PubMed: 16860820]

[2]. Turner GS, Tjaden K, and Weismer G, "The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis," Journal of Speech, Language, and Hearing Research, vol. 38, no. 5, pp. 1001–1013, 1995.

[3]. Kent RD, Kent JF, Weismer G, Sufit RL, Rosenbek JC, Martin RE, and Brooks BR, "Impairment of speech intelligibility in men with amyotrophic lateral sclerosis," Journal of Speech and Hearing Disorders, vol. 55, no. 4, pp. 721–728, 1990. [PubMed: 2232752]

[4]. Starmer HM, Tippett DC, and Webster KT, "Effects of laryngeal cancer on voice and swallowing," Otolaryngologic Clinics of North America, vol. 41, no. 4, pp. 793–818, 2008. [PubMed: 18570960]

[5]. Schultz T, Wand M, Hueber T, Krusienski DJ, Herff C, and Brumberg JS, "Biosignal-based spoken communication: A survey," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 25, no. 12, pp. 2257–2271, 11 2017.

[6]. Wolpaw J, Birbaumer N, McFarland D, Pfurtscheller G, and Vaughan T, "Brain-computer interfaces for communication and control," Clinical neurophysiology, vol. 113, no. 6, pp. 767–791, 2002. [PubMed: 12048038]

[7]. Bocquelet F, Hueber T, Girin L, Chabardes S, and Yvert B, "Key considerations in designing a speech brain-computer interface," Journal of Physiology-Paris, vol. 110, no. 4, pp. 392–401, 2016.

[8]. Herff C and Schultz T, "Automatic speech recognition from neural signals: a focused review," Frontiers in neuroscience, vol. 10, 2016.

[9]. Slutzky MW and Flint RD, "Physiological properties of brain-machine interface input signals," Journal of neurophysiology, vol. 118, no. 2, pp. 1329–1343, 2017. [PubMed: 28615329]

[10]. Chakrabarti S, Sandberg HM, Brumberg JS, and Krusienski DJ, "Progress in speech decoding from the electrocorticogram," Biomedical Engineering Letters, vol. 5, no. 1, pp. 10–21, 2015.

[11]. Rabbani Q, Milsap G, and Crone NE, "The potential for a speech brain–computer interface using chronic electrocorticography," Neurotherapeutics, pp. 1–22, 2019. [PubMed: 30652252]

[12]. Ries SK, Dhillon RK, Clarke A, King-Stephens D, Laxer KD, Weber PB, Kuperman RA, Auguste KI, Brunner P, Schalk G et al., "Spatiotemporal dynamics of word retrieval in speech production revealed by cortical high-frequency band activity," Proceedings of the National Academy of Sciences, vol. 114, no. 23, pp. E4530–E4538, 2017.

[13]. Kellis S, Miller K, Thomson K, Brown R, House P, and Greger B, "Decoding spoken words using local field potentials recorded from the cortical surface," Journal of neural engineering, vol. 7, no. 5, p. 056007, 2010. [PubMed: 20811093]

[14]. Ramsey N, Salari E, Aarnoutse E, Vansteensel M, Bleichner M, and Freudenburg Z, "Decoding spoken phonemes from sensorimotor cortex with high-density ecog grids," NeuroImage, 2017.

[15]. Mugler E, Patton J, Flint R, Wright Z, Schuele S, Rosenow J, Shih J, Krusienski D, and Slutzky M, “Direct classification of all American English phonemes using signals from functional speech motor cortex,” Journal of Neural Engineering, vol. 11, no. 3, p. 035015, 2014. [PubMed: 24836588]

[16]. Lotte F, Brumberg JS, Brunner P, Gunduz A, Ritaccio AL, Guan C, and Schalk G, “Electrocorticographic representations of segmental features in continuous speech,” Frontiers in human neuroscience, vol. 9, 2015.

[17]. Mesgarani N, Cheung C, Johnson K, and Chang EF, “Phonetic feature encoding in human superior temporal gyrus,” Science, p. 1245994, 2014.

[18]. Mugler EM, Tate MC, Livescu K, Templer JW, Goldrick MA, and Slutzky MW, “Differential representation of articulatory gestures and phonemes in precentral and inferior frontal gyri,” Journal of Neuroscience, pp. 1206–18, 2018. [PubMed: 30541908]

[19]. Herff C, Heger D, de Pesters A, Telaar D, Brunner P, Schalk G, and Schultz T, “Brain-to-text: decoding spoken phrases from phone representations in the brain,” Frontiers in neuroscience, vol. 9, 2015.

[20]. Moses DA, Mesgarani N, Leonard MK, and Chang EF, “Neural speech recognition: continuous phoneme decoding using spatiotemporal representations of human cortical activity,” Journal of neural engineering, vol. 13, no. 5, p. 056004, 2016. [PubMed: 27484713]

[21]. Stuart A, Kalinowski J, Rastatter MP, and Lynch K, “Effect of delayed auditory feedback on normal speakers at two speech rates,” The Journal of the Acoustical Society of America, vol. 111, no. 5, pp. 2237–2241, 2002. [PubMed: 12051443]

[22]. Stuart A and Kalinowski J, “Effect of delayed auditory feedback, speech rate, and sex on speech production,” Perceptual and motor skills, vol. 120, no. 3, pp. 747–765, 2015. [PubMed: 26029968]

[23]. Iljina O, Derix J, Schirrmeister RT, Schulze-Bonhage A, Auer P, Aertsen A, and Ball T, “Neurolinguistic and machine-learning perspectives on direct speech bcis for restoration of naturalistic communication,” Brain-Computer Interfaces, vol. 4, no. 3, pp. 186–199, 2017 [Online]. Available: 10.1080/2326263X.2017.1330611

[24]. Santoro R, Moerel M, De Martino F, Valente G, Ugurbil K, Yacoub E, and Formisano E, “Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns,” Proceedings of the National Academy of Sciences, p. 201617622, 2017.

[25]. Kubanek J, Brunner P, Gunduz A, Poeppel D, and Schalk G, “The tracking of speech envelope in the human cortex,” PloS one, vol. 8, no. 1, p. e53398, 2013. [PubMed: 23408924]

[26]. Tang C, Hamilton L, and Chang E, “Intonational speech prosody encoding in the human auditory cortex,” Science, vol. 357, no. 6353, pp. 797–801, 2017. [PubMed: 28839071]

[27]. Dichter BK, Breshears JD, Leonard MK, and Chang EF, “The control of vocal pitch in human laryngeal motor cortex,” Cell, vol. 174, pp. 21–31, 2018. [PubMed: 29958109]

[28]. Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, and Chang EF, “Reconstructing speech from human auditory cortex,” PLoS biology, vol. 10, no. 1, p. e1001251, 2012. [PubMed: 22303281]

[29]. Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone N, Rieger J, Schalk G, Knight R, and Pasley B, “Decoding spectrotemporal features of overt and covert speech from the human cortex,” Frontiers in Neuroengineering, vol. 7, no. 14, 2014.

[30]. Herff C, Johnson G, Diener L, Shih J, Krusienski D, and Schultz T, “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 IEEE, 2016, pp. 1540–1543.

[31]. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A.-r., Jaitly N, Senior A, Vanhoucke V, Nguyen P, Sainath TN et al., “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[32]. Ze H, Senior A, and Schuster M, “Statistical parametric speech synthesis using deep neural networks,” in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on IEEE, 2013, pp. 7962–7966.

[33]. Pandarinath C, O'Shea DJ, Collins J, Jozefowicz R, Stavisky SD, Kao JC, Trautmann EM, Kaufman MT, Ryu SI, Hochberg LR et al., "Inferring single-trial neural population dynamics using sequential auto-encoders," Nature methods, p. 1, 2018.

[34]. Seeliger K, Fritsche M, Güglü U, Schoenmakers S, Schoffelen J-M, Bosch S, and van Gerven M, "Convolutional neural network-based encoding and decoding of visual object recognition in space and time," NeuroImage, 2017.

[35]. Berezutskaya J, Freudenburg Z, Ramsey N, Güçlü U, van Gerven M, Duivesteijn W, Pechenizkiy M, Fletcher G, Menkovski V, Postma E et al., "Modeling brain responses to perceived speech with lstm networks," in Duivesteijn W; Pechenizkiy M; Fletcher GHL (ed.), Benelearn 2017: Proceedings of the Twenty-Sixth Benelux Conference on Machine Learning, Technische Universiteit Eindhoven, 9–10 June 2017 [Sl: sn], 2017, pp. 149–153.

[36]. Güçlü U and van Gerven MA, "Modeling the dynamics of human brain activity with recurrent neural networks," Frontiers in computational neuroscience, vol. 11, 2017.

[37]. Sturm I, Lapuschkin S, Samek W, and Müller K-R, "Interpretable deep neural networks for single-trial EEG classification," Journal of neuroscience methods, vol. 274, pp. 141–145, 2016. [PubMed: 27746229]

[38]. Sussillo D, Nuyujukian P, Fan JM, Kao JC, Stavisky SD, Ryu S, and Shenoy K, "A recurrent neural network for closed-loop intracortical brain-machine interface decoders," Journal of neural engineering, vol. 9, no. 2, p. 026027, 2012. [PubMed: 22427488]

[39]. Rezazadeh Sereshkeh A, Trott R, Bricout A, and Chau T, "Eeg classification of covert speech using regularized neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP), vol. 25, no. 12, pp. 2292–2300, 2017.

[40]. Schirrmeister R, Springenberg J, Fiederer L, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W, and Ball T, "Deep learning with convolutional neural networks for EEG decoding and visualization," Human brain mapping, 2017.

[41]. Hennrich J, Herff C, Heger D, and Schultz T, "Investigating deep learning for fnirs based bci," in Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, 8 2015.

[42]. Angrick M, Herff C, Johnson G, Shih J, Krusienski D, and Schultz T, "Interpretation of Convolutional Neural Networks for Speech Regression from Electrocorticography," in ESANN 2018 – 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Brugge, Belgium, 2018, pp. 7–12.

[43]. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerry-Ryan R et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," arXiv preprint arXiv:171205884, 2017.

[44]. Antoniades A, Spyrou L, Took CC, and Sanei S, "Deep learning for epileptic intracranial eeg data," in Machine Learning for Signal Processing (MLSP), 2016 IEEE 26th International Workshop on IEEE, 2016, pp. 1–6.

[45]. Chambon S, Thorey V, Arnal PJ, Mignot E, and Gramfort A, "A deep learning architecture to detect events in eeg signals during sleep," in 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP) IEEE, 2018, pp. 1–6.

[46]. House AS, Williams C, Hecker MH, and Kryter KD, "Psychoacoustic speech tests: A modified rhyme test," The Journal of the Acoustical Society of America, vol. 35, no. 11, pp. 1899–1899, 1963.

[47]. Schalk G, McFarland DJ, Hinterberger T, Birbaumer N, and Wolpaw JR, "Bci2000: a general-purpose brain-computer interface (bci) system," IEEE Transactions on biomedical engineering, vol. 51, no. 6, pp. 1034–1043, 2004. [PubMed: 15188875]

[48]. Ray S, Crone NE, Niebur E, Franaszczuk PJ, and Hsiao SS, "Neural correlates of high-gamma oscillations (60–200 hz) in macaque local field potentials and their potential implications in electrocorticography," Journal of Neuroscience, vol. 28, no. 45, pp. 11526–11536, 2008. [PubMed: 18987189]

[49]. Livezey JA, Bouchard KE, and Chang EF, "Deep learning as a tool for neural data analysis: speech classification and cross-frequency coupling in human sensorimotor cortex," ArXiv e-prints, 3 2018.

[50]. Hickok G, “Computational neuroanatomy of speech production,” Nature Reviews Neuroscience, vol. 13, no. 2, p. 135, 2012. [PubMed: 22218206]

[51]. Brumberg J, Krusienski D, Chakrabarti S, Gunduz A, Brunner P, Ritaccio A, and Schalk G, “Spatio-Temporal Progression of Cortical Activity Related to Continuous Overt and Covert Speech Production in a Reading Task,” PloS one, vol. 11, no. 11, p. e0166872, 2016. [PubMed: 27875590]

[52]. Martin S, Millan J. d. R., Knight RT, and Pasley BN, “The use of intracranial recordings to decode human language: challenges and opportunities,” Brain and language, 2016.

[53]. Stevens SS, Volkmann J, and Newman EB, “A scale for the measurement of the psychological magnitude pitch,” The Journal of the Acoustical Society of America, vol. 8, no. 3, pp. 185–190, 1937.

[54]. Huang G, Liu Z, Weinberger KQ, and van der Maaten L, “Densely connected convolutional networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, vol. 1, no. 2, 2017, p. 3.

[55]. He K, Zhang X, Ren S, and Sun J, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[56]. Kingma D and Ba J, “Adam: A method for stochastic optimization,” arXiv preprint arXiv: 14126980, 2014.

[57]. Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, and Kavukcuoglu K, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:160903499, 2016.

[58]. Tamamori A, Hayashi T, Kobayashi K, Takeda K, and Toda T, “Speaker-dependent wavenet vocoder,” in Proceedings of Interspeech, 2017, pp. 1118–1122.

[59]. Nagaraj Adiga VT, “On the use of wavenet as a statistical vocoder,” 2018 [Online]. Available: http://sigport.org/2931

[60]. Ito K, “The lj speech dataset,” https://keithito.com/LJ-Speech-Dataset/, 2017.

[61]. Salimans T, Karpathy A, Chen X, and Kingma DP, “Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications,” arXiv preprint arXiv:170105517, 2017.

[62]. Oord A. v. d., Li Y, Babuschkin I, Simonyan K, Vinyals O, Kavukcuoglu K, Driessche G. v. d., Lockhart E, Cobo LC, i F et al., “Parallel wavenet: Fast high-fidelity speech synthesis,” arXiv preprint arXiv:171110433, 2017.

[63]. Yamamoto R, “Wavenet vocoder,” https://github.com/r9y9/wavenet_vocoder, 2018.

[64]. Taal CH, Hendriks RC, Heusdens R, and 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 IEEE, 2010, pp. 4214–4217.

[65]. Su H and Chen H, “Experiments on parallel training of deep neural network using model averaging,” arXiv preprint arXiv:150701239, 2015.

[66]. Kalchbrenner N, Elsen E, Simonyan K, Noury S, Casagrande N, Lockhart E, Stimberg F, Oord A. v. d., Dieleman S, and Kavukcuoglu K, “Efficient neural audio synthesis,” arXiv preprint arXiv: 180208435, 2018.

[67]. Hochberg LR, Serruya MD, Friehs GM, Mukand JA, Saleh M, Caplan AH, Branner A, Chen D, Penn RD, and Donoghue JP, “Neuronal ensemble control of prosthetic devices by a human with tetraplegia,” Nature, vol. 442, no. 7099, p. 164, 2006. [PubMed: 16838014]

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

[69]. Downey JE, Brane L, Gaunt RA, Tyler-Kabara EC, Boninger ML, and Collinger JL, “Motor cortical activity changes during neuroprosthetic-controlled object interaction,” Scientific reports, vol. 7, no. 1, p. 16947, 2017. [PubMed: 29209023]

[70]. Ajiboye AB, Willett FR, Young DR, Memberg WD, Murphy BA, Miller JP, Walter BL, Sweet JA, Hoyen HA, Keith MW et al., “Restoration of reaching and grasping movements through

brain-controlled muscle stimulation in a person with tetraplegia: a proof-of-concept demonstration," The Lancet, vol. 389, no. 10081, pp. 1821–1830, 2017.

[71]. Guenther FH, Brumberg JS, Wright EJ, Nieto-Castanon A, Tourville JA, Panko M, Law R, Siebert SA, Bartels JL, Andreasen DS et al., "A wireless brain-machine interface for real-time speech synthesis," PloS one, vol. 4, no. 12, p. e8218, 2009. [PubMed: 20011034]

[72]. Martin S, Brunner P, Iturrate I, Millan J, Schalk G, Knight R, and Pasley B, "Word pair classification during imagined speech using direct brain recordings," Scientific reports, vol. 6, p. 25803–2016.

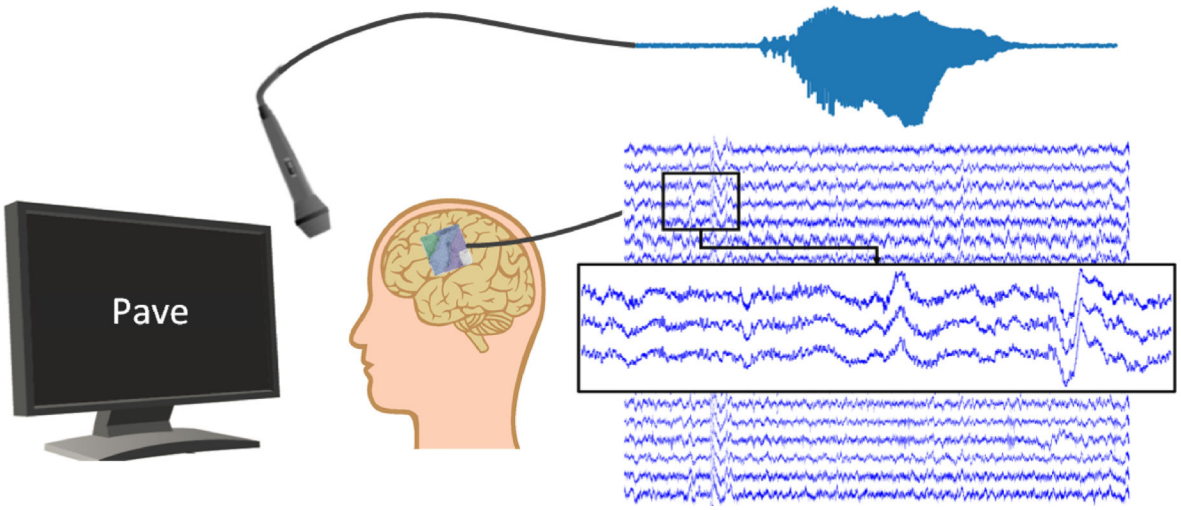

Figure 1.

Illustration of the experiment. Participants are asked to repeat words shown on a screen. During speech production, ECoG data and acoustic stream are recorded simultaneously.

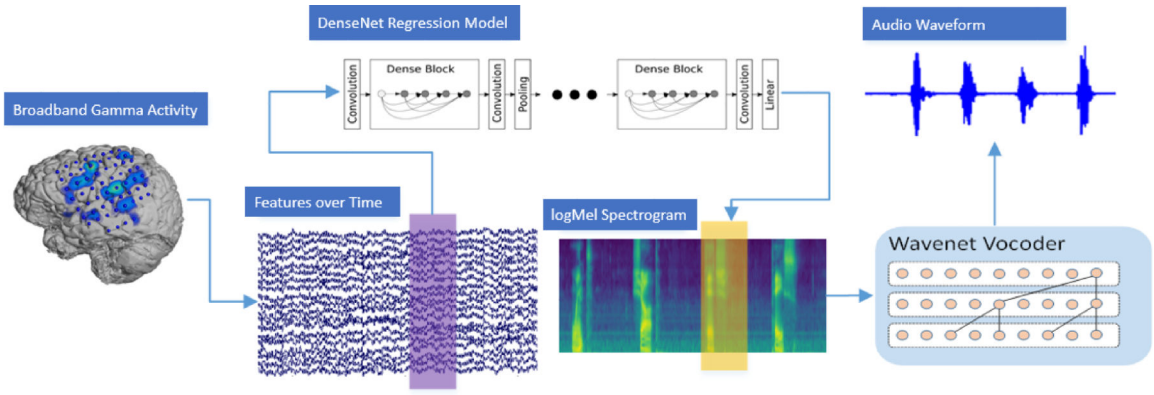

Figure 2.

Overview of the decoding approach illustrating the transformation of neural data into an audible waveform. ECoG features for each time window are fed into DenseNet regression model to reconstruct the logarithmic mel-scaled spectrogram. Wavenet is then used to reconstruct an audio waveform from the spectrogram.

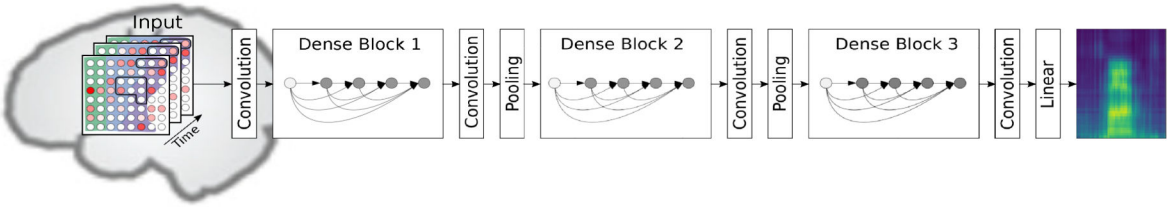

Figure 3.

Overview of the DenseNet network structure. Input samples are preprocessed features of the neural signal with the shape 8 × 8 × 9. The first two dimensions are used for the spatial alignment of the electrodes, while the third dimension comprises the temporal dynamics. The network architecture consists of three Dense Blocks to map the neural features onto the speech spectrogram.

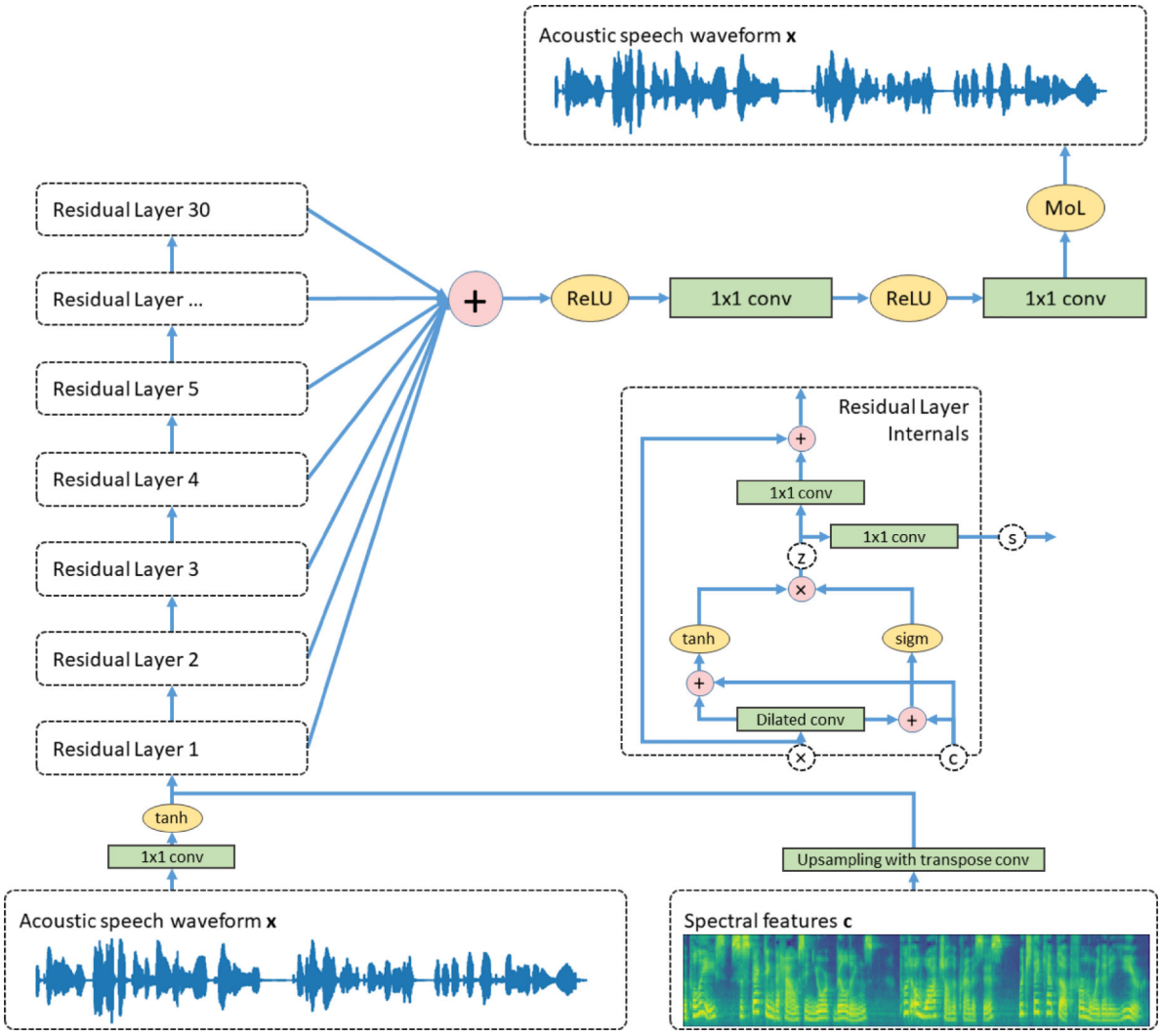

Figure 4.

Overview of the Wavenet vocoder architecture. The network comprises a stack of 30 residual blocks to find a mapping between the acoustic speech signal **x** to itself considering the extracted features **c**. Each block has a separate output which are summed in the calculation of the actual prediction. We use a 10-component mixture of logistic distributions (MoL) for the prediction of audio samples.

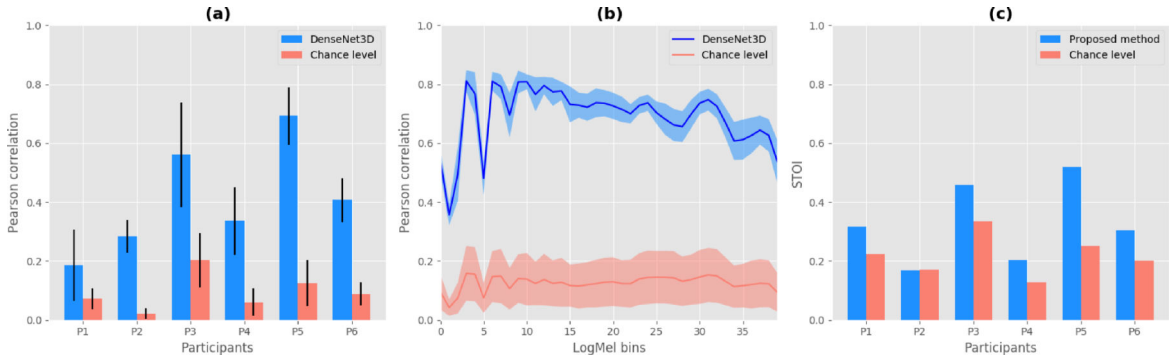

Figure 5.

Reconstruction performance of DenseNet compared to random chance. (a) Pearson correlation coefficients between original and reconstructed spectrograms for each participant. Bars indicate the mean over all logarithmic mel-scaled coefficients while whiskers denote the standard deviation. (b) Detailed performance across all spectral bins for participant 5. (c) STOI scores as an objective intelligibility measure in comparison to the chance level.

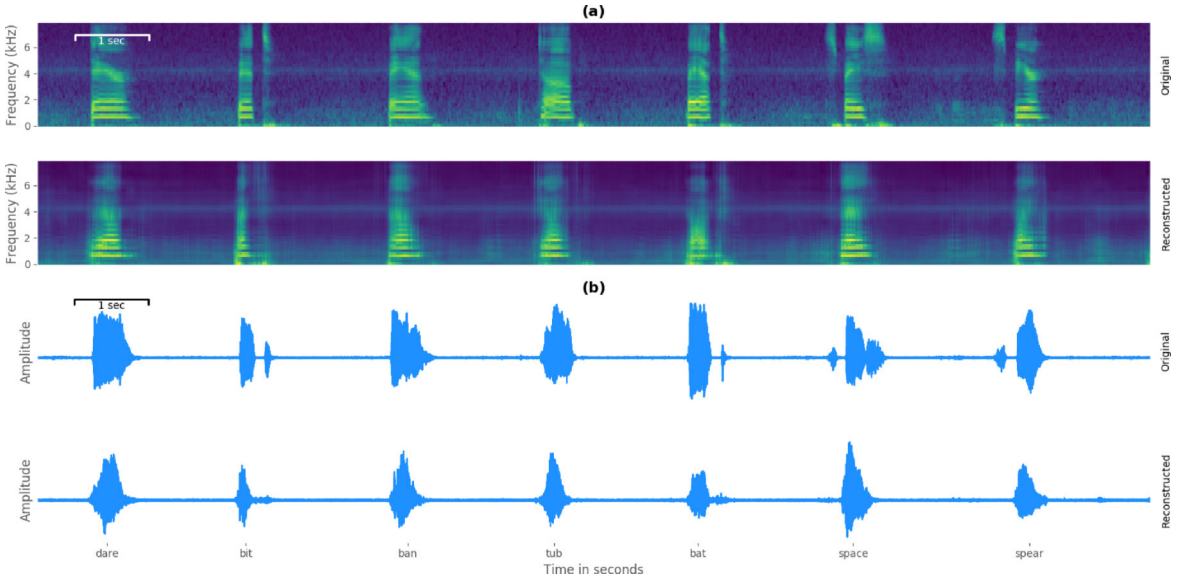

Figure 6.

Reconstruction example for visual inspection. a) compares a time-aligned excerpt in the spectral domain of participant 5 and emphasizes the quality of the reconstructed acoustic speech characteristics. b) shows the generated waveform representation of the same excerpt as in the spectrogram comparison. Spoken words are given below.

Table 1.

Overview of the hyperparameter choices made for our Wavenet vocoder.

Hyperparameter	Choice
Residual channels	512
Skip connections	256
Receptive field	~ 0.5 seconds
Residual blocks	30