ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096082

RELATING EEG RECORDINGS TO SPEECH USING ENVELOPE TRACKING AND THE SPEECH-FFR

Mike Thornton[1], Danilo Mandic[1], Tobias Reichenbach[2]

[1]Imperial College London, [2]Friedrich-Alexander-Universität Erlangen-Nürnberg

ABSTRACT

During speech perception, a listener's electroencephalogram (EEG) reflects acoustic-level processing as well as higher-level cognitive factors such as speech comprehension and attention. However, decoding speech from EEG recordings is challenging due to the low signal-to-noise ratios of EEG signals. We report on an approach developed for the ICASSP 2023 'Auditory EEG Decoding' Signal Processing Grand Challenge. A simple ensembling method is shown to considerably improve upon the baseline decoder performance. Even higher classification rates are achieved by jointly decoding the speech-evoked frequency-following response and responses to the temporal envelope of speech, as well as by fine-tuning the decoders to individual subjects. Our results could have applications in the diagnosis of hearing disorders or in cognitively steered hearing aids.

Index Terms— EEG decoding, deep learning, speech

1. INTRODUCTION

When people listen to speech, their brainwaves synchronise with acoustic features such as the speech envelope. The degree of this neural tracking reflects cognitive factors such as attention to speech, speech comprehension, and intelligibility [1, 2]. Accurately decoding speech from electroencephalography (EEG) is, however, a challenging task owing to the low signal-to-noise ratios of EEG recordings and the limited availability of EEG data recorded during speech perception.

Neural speech tracking in a particular subject is best assessed when EEG responses to speech from that particular subject are available and can be used to train a subject-specific decoder. Decoding EEG responses to speech from unseen subjects is a harder task, since EEG signals vary greatly between individuals. The ICASSP 2023 'Auditory EEG Decoding' Signal Processing Grand Challenge involves both types of decoding. Its task is to develop population match-mismatch decoders: given a temporal segment of EEG data and two candidate speech segments, the decoder should predict which of the speech segments corresponds to the EEG signal. This should be done both for EEG recordings from subjects that have been included in the training set, and for others whose EEG data has not been seen in the training stage.

Here we describe the development of our decoders which placed first in this match-mismatch task. Besides the speech envelope, we relate a second acoustic feature to the EEG recordings. This feature is the temporal fine structure of the voiced parts of speech (which consist of a fundamental frequency and many higher harmonics). The electroencephalogram displays a strong response at the fundamental frequency of speech, termed the speech-frequency-following response or speech-FFR [3]. The speech-FFR is driven both by the fundamental frequency itself as well as by higher harmonics, and can be decoded from EEG responses to continuous speech [3, 4, 5, 6].

Mike Thornton is supported by UK Research and Innovation. [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1].

2. MATERIALS AND METHODS

2.1. Dataset. A large training dataset was provided by the ICASSP competition organisers, which consisted of EEG recordings from 71 subjects who listened to speech material [7]. Decoders were evaluated against a heldout dataset comprising EEG from 70 subjects included in the training dataset, and 15 new unseen subjects.

We used two pre-processed versions of the dataset that contained the two speech features of interest together with the corresponding EEG signals. The first dataset consisted of speech envelopes and EEG recordings sampled at $64\,\mathrm{Hz}$. The second dataset contained the envelope modulations of the higher harmonics of the fundamental frequency of the voiced parts of speech, together with EEG sampled at $512\,\mathrm{Hz}$ [6]. Pre-processing and methodological details will be described in a forthcoming publication [8].

2.2. Baseline decoder. The baseline decoder is a deep neural network (DNN) proposed by Accou *et al.* [2]. This DNN brings the speech envelopes and EEG recordings into a space where the matched envelope and EEG segments are maximally similar. The output layer consists of a single Sigmoid neuron. The matched and mismatched envelopes are presented as an ordered pair, and the Sigmoid neuron predicts the probability that the first envelope is matched to the EEG. The decoder is trained with the binary crossentropy loss function and Adam optimizer without regularisation.

2.3. Baseline + speech-FFR decoder. The baseline decoder relates the EEG recordings to the temporal envelope of speech only. We were also interested in relating the EEG signals to the temporal fine structure of speech, since this can improve classification performance [9]. We retained the architecture of the baseline decoder, but swapped the speech envelopes for the high-frequency envelope-modulations feature [6]. The Sigmoid outputs of this speech-FFR decoder and the baseline decoder were combined via linear discriminant analysis (LDA) to produce the final predicted label.

2.4. Decoder training and fine-tuning. Training examples were presented to the decoders as temporal segments of $3\,\mathrm{s}$ in duration (the same duration was used for evaluation). Sources of randomness in the training procedure include the decoder initialisation, and the order in which training examples were presented. The effects of these were marginalised by averaging the Sigmoid outputs of several trained instances of the decoders. For the population decoders, the hop length between the onsets of the training examples was $1\,\mathrm{s}$. When fine-tuning the population decoder to individual subjects, this was reduced to $0.125\,\mathrm{s}$, and regularisation (batch normalisation and spatial dropout) was applied to the input layers.

3. RESULTS

3.1. Averaging of decoder outputs. We trained 100 instances of the baseline decoder. By averaging the Sigmoid outputs of the instances, the classification accuracy was improved (Figure 1). Therefore, we formed two averaged population decoders: these used 50 instances of the baseline decoder, and 30 instances of the speech-FFR decoder, respectively. The averaged decoders were used in Section 3.2.

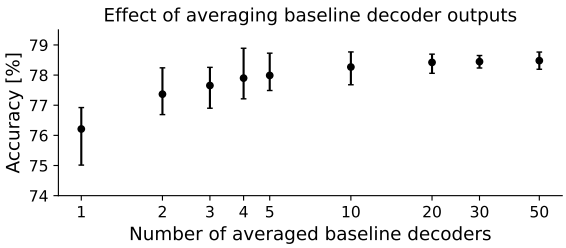

Fig. 1. Classification accuracies against number of averaged decoders (log scale). The bootstrapped mean and range of the accuracies are denoted by the dots and errorbars respectively.

3.2. Combining averaged baseline and speech-FFR decoders. The baseline decoder generally achieves higher accuracies than the speech-FFR decoder (Figure 2, left). The LDA classifier was fitted on an unseen portion of the training dataset, for which the correlation between the Sigmoid outputs of the decoders was moderate ($R = 0.229$, Figure 2, right). For the heldout dataset the correlation was similar ($R = 0.206$), confirming that there was no severe overfitting or distributional shift. This composite decoder achieved an accuracy of 81.18% on the heldout dataset for unseen subjects.

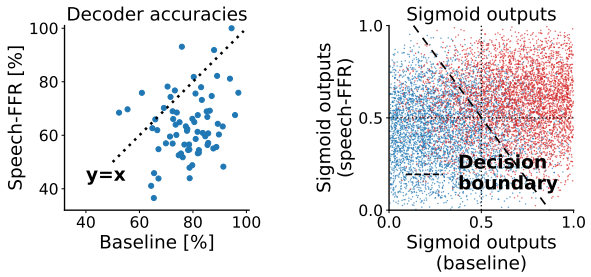

Fig. 2. Comparison between the averaged baseline and speech-FFR decoders. (Left) The accuracies of the two decoders are shown for the 71 subjects. (Right) The Sigmoid outputs of the averaged decoders are shown. A red (blue) point indicates an EEG segment that matches the first (second) speech segment (recall that these are grouped as an ordered pair). The LDA decision boundary is shown.

3.3. Decoder fine-tuning. Figure 3 shows the effect of fine-tuning the decoders to individual subjects. For both the baseline decoders and the speech-FFR decoders, fine-tuning significantly improved the classification accuracy ($p < 0.001$, signed-rank tests). The fine-tuned baseline decoders achieved an accuracy of 82.71% on the held-out dataset for seen subjects.

4. CONCLUSIONS

The performance of the baseline decoder could be improved by averaging the outputs of several trained decoder instances. Combining

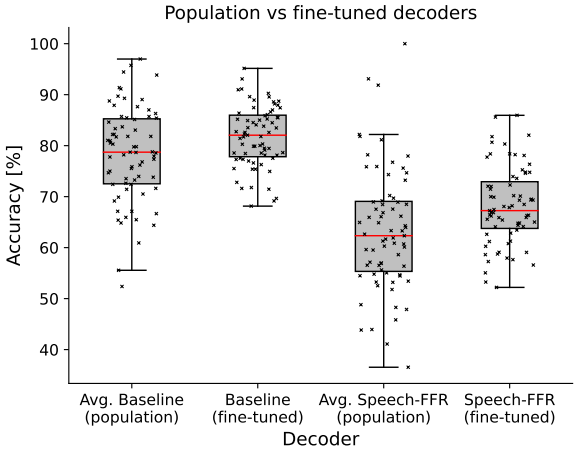

Fig. 3. Effect of fine-tuning the decoders to individual subjects. Each datapoint represents the classification accuracy for a single subject.

the averaged baseline and speech-FFR decoders enhanced the decoding accuracies further. The best results were obtained by using versions of the baseline decoder which were fine-tuned to individual subjects where possible, and by using the composite decoder for unseen subjects. This approach won the match-mismatch task of the ICASSP 2023 Signal Processing Grand Challenge 'Auditory EEG Decoding'. Future work will establish which aspects of the fine-tuning procedure led to such high classification accuracies.

5. REFERENCES

[1] O. Etard and T. Reichenbach, "Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise," *J. Neurosci.*, vol. 39, no. 29, pp. 5750–5759, 2019.

[2] B. Accou *et al.*, "Predicting speech intelligibility from EEG in a non-linear classification paradigm," *J. Neural Eng.*, vol. 18, no. 6, pp. 066008, 2021.

[3] N. Kraus *et al.*, "The frequency-following response: a window into human communication," in *The Frequency-Following Response*, pp. 1–15. Springer, 2017.

[4] A. E. Forte *et al.*, "The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention," *eLife*, vol. 6, pp. e27203, 2017.

[5] O. Etard *et al.*, "Decoding of selective attention to continuous speech from the human auditory brainstem response," *NeuroImage*, vol. 200, pp. 1–11, 2019.

[6] P. Kulasingham *et al.*, "High gamma cortical processing of continuous speech in younger and older listeners," *NeuroImage*, vol. 222, pp. 117291, 2020.

[7] L. Bollens *et al.*, "A Large Auditory EEG decoding dataset,", doi: 10.48804/K3VSND 2023.

[8] M. Thornton *et al.*, "Relating EEG with speech using envelope tracking and the speech-FFR," *in preparation*.

[9] C. Puffay *et al.*, "Relating the fundamental frequency of speech with EEG using a dilated convolutional network," in *Proc. Interspeech 2022*, pp. 4038–4042.