Speech synthesis from intracranial stereotactic Electroencephalography using a neural vocoder

Frigyes Viktor Arthur and Tamás Gábor Csapó

Abstract—**Speech is one of the most important human biosignals. However, only some speech production characteristics are fully understood, which are required for a successful speech-based Brain-Computer Interface (BCI). A proper brain-to-speech system that can generate the speech of full sentences intelligibly and naturally poses a great challenge. In our study, we used the SingleWordProduction-Dutch-iBIDS dataset, in which speech and intracranial stereotactic electroencephalography (sEEG) signals of the brain were recorded simultaneously during a single word production task. We apply deep neural networks (FC-DNN, 2D-CNN, and 3D-CNN) on the ten speakers' data for sEEG-to-Mel spectrogram prediction. Next, we synthesize speech using the WaveGlow neural vocoder. Our objective and subjective evaluations have shown that the DNN-based approaches with neural vocoder outperform the baseline linear regression model using Griffin-Lim. The synthesized samples resemble the original speech but are still not intelligible, and the results are clearly speaker dependent. In the long term, speech-based BCI applications might be useful for the speaking impaired or those having neurological disorders.**

Index Terms—**human-computer interaction, sEEG, BCI, brain-computer interface**

I. INTRODUCTION

It is expected that 0.4% of the European population suffers from a speech impairment [1], [2], [3]. Digital applications using speech technology could significantly help their everyday communication. Loss of speech can cause social isolation, and feelings of loss of identity and can lead to clinical depression [4]. Augmentative and alternative communication (AAC) technologies, such as brain-computer interfaces (BCIs) might directly read brain signals to restore lost speech capabilities [5]. In the future, the application of speech neuroprostheses have the potential to help patients with neurological disorders or speech impairment.

Brain-computer interfaces enable direct control of computers without physical activity, with potential applications as rehabilitation devices for motor-impaired persons (e.g., input system for writing, prosthetic control). Ideally, BCI applications operate in naturalistic scenarios, requiring a neural input with good temporal resolution, minimal preprocessing needs and relative ease of measurement. There are several available modalities for neuroimaging, including electroencephalography (EEG) [6], stereotactic depth electrodes [7], intracranial electrocorticography (ECoG) [8], Magnetoencephalography (MEG) [9], Local Field Potential (LFP) [8].

Budapest University of Technology and Economics Department of Telecommunications and Media Informatics (BME), Hungary

corresponding author (e-mail: arthur@tmit.bme.hu) (e-mail: csapot@tmit.bme.hu)

DOI: 10.36244/ICJ.2024.1.6

From the above, EEG has been the most widely studied one for BCI [6], [10]. EEG is a non-invasive method for measuring small electrical currents on the scalp, which reflect brain activity. It allows one to assess cortical excitability and effective connectivity in clinical and basic research without extensive invasive surgical installation. However, obtaining clean and usable EEG recordings (e.g., signals, data) is challenging due to the various bio-physiology-related artifacts that contaminate the electroencephalographic signal. In biomedical applications, such as monitoring brain activity during surgery or in sleep studies, EEG measurements typically utilize multiple electrodes, ranging from 32 to 256, with sampling rates around 256–2048 Hz. Relative to other methods recording electric potentials from the brain (ECoG, MEG, LFP), at the cost of poorer SNR and lower spatial resolution [8], EEG is non-invasive, cheap, and can be obtained even with wearable devices that allow for measurements outside the lab [11].

Csapó et al. [12] present a novel multimodal analysis method that combines EEG, articulatory movements, and speech signals for multimodal analysis, combining brain signal analysis during speech with ultrasound-based articulatory data. This study developed a fully connected deep neural network (FC-DNN) to predict articulatory movements using EEG signals. The study has demonstrated a clear relationship between EEG and articulatory movements and therefore provides valuable insights for future research in speech BCI.

Arthur and Csapó [13] discuss using deep learning to process EEG brain signals and synthesize speech. EEG signals were processed and used in this study to estimate the mel-spectral parameters of speech using deep learning models. Although not intelligible, the synthesized speech resembled the original speech signal, presenting a promising avenue for further investigation.

While initial results are encouraging, it is important to recognize the current limitations and challenges facing EEG-based BCI systems in the context of speech synthesis. Although these systems show potential, especially for aiding individuals with speech impairments, the extent of their effectiveness and practical applicability remains an area of ongoing research. The journey towards refining these technologies to reliably and effectively synthesize speech involves overcoming significant technical and scientific hurdles. Continued research and development are crucial to enhance our understanding and to push the boundaries of what is achievable with EEG-based BCIs. Ultimately, the goal is to leverage these advancements to improve the quality of life for those facing communication challenges, but it is essential to maintain a realistic perspective on the current state of the technology and the work that still lies ahead.

More invasive methods offer increased insights into brain activity compared to EEG. Still, invasive EEG-based speech-BCIs (e.g., brain-to-speech and brain-to-text) are not yet successful due to the fact that the input brain signal and the target speech signal or text are spatially, acoustically, and temporally too distant from each other. All studies related to this topic [14], [15], [16], [17], [18], except for a feasibility experiment [19], use estimated or "indirect" articulatory information, meaning that they consider the articulatory data derived from the speech signal or the textual content. Recently, a novel database featuring parallel speech and intracranial stereotactic Electroencephalography recordings has been introduced (SingleWordProduction-Dutch-iBIDS dataset, [7]). This dataset employs a baseline linear regression method for sEEG-to-speech mapping, utilizing the Griffin-Lim algorithm for speech generation. As highlighted by Verwoert et al. in [7], the application of neural vocoders in conjunction with deep neural networks for sEEG-to-speech prediction has not been previously explored.

A. Goal of the current study

Speech is one of the most important human biosignals, but not all the characteristics of speech production are fully understood, which are required for a successful speech-based BCI [20]. A proper brain-to-speech system capable of generating full sentences in an intelligible and natural manner presents significant challenges and necessitates multidisciplinary approaches. In this paper, we apply deep neural networks for sEEG-to-speech synthesis, using neural vocoders.

In our study, we employed the Griffin-Lim algorithm as a baseline method for speech generation and used linear regression for mapping brain signals to speech features, following the methodology of Verwoert et al. [7]. This choice maintains consistency with existing literature and enables direct comparison of our results. The simplicity and ease of implementation of both techniques provide easily replicable and interpretable baselines, highlighting the improvements offered by advanced methods, such as deep learning-based solutions compared to traditional techniques.

II. Related work

A. Brain-to-speech synthesis

There has been some research on non-invasive EEG-to-speech synthesis [21], [22]. As EEG provides information only from the surface of the scalp, this process is extremely difficult, and until now there has been no successful approach to predict fully intelligible synthesized speech. On the other hand, typically more invasive methods have been tested for speech BCI [20]. With participants implanted using sEEG, audible speech could be reliably generated in real-time [23].

With intracranial electrocorticography (ECoG), another highly invasive procedure, continuous speech decoding could be solved [15]. Verwoert et al. [7] applied the Griffin-Lim algorithm in combination with linear regression to show that sEEG-to-speech mapping is feasible. According to the correlations that they received during cross-validation and comparison of 10 speakers, the results are highly dependent on the speaker, most probably because of the location of the sEEG electrodes in the individual subjects.

Another recent article, Lesaja et al. [24] presents brain2vec, a self-supervised model for learning speech-related hidden unit representations from unlabeled intracranial EEG data. Brain2vec's performance rivals that of competitive supervised learning methods on speech activity detection and word classification tasks, indicating potential practical applications in speech decoding using intracranial EEG data.

The BrainBERT model, introduced as a transformer-based model, marks a significant advancement in analyzing neural signals recorded from the human brain for natural language decoding [25]. This model, an adaptation of the well-established BERT (Bidirectional Encoder Representations from Transformers) in Natural Language Processing, is specifically designed to translate brain signals into natural language. Unlike traditional methods that predominantly rely on labeled data, BrainBERT employs self-supervised learning from extensive unlabeled data, potentially enhancing its performance. As per the original BERT model, BrainBERT records context from both directions of the input data (in this case, brain signals), which allows it to understand the temporal dependency between signals [26]. Recent studies have examined BrainBERT using sEEG data, with promising results [25].

B. Neural vocoders in speech synthesis

Since the introduction of WaveNet in 2016 [27], neural vocoders have been instrumental in generating highly natural raw samples of speech. These vocoders, including recent variants like WaveGlow [28], synthesize high-quality speech by transforming mel-spectrograms or other spectral feature inputs into audio waveforms. WaveGlow, in particular, stands out as a flow-based network capable of real-time, high-quality speech synthesis from mel-spectrograms. Its simplicity and efficiency in speech generation offer considerable advantages. This approach has been effectively utilized in various applications, such as in the work of Csapó et al. [29], who integrated WaveGlow into an ultrasound-based articulatory-to-acoustic conversion process. Similarly, Cao and colleagues demonstrated the successful use of WaveGlow for synthesizing speech from Electromagnetic Articulography (EMA) data of tongue movements [30].

C. Speaker adaptation in Text-To-Speech synthesis

A significant area of research in this field has focused on the development of natural-sounding speech synthesis. Csapó et al. have extensively explored the role of prosodic variability methods in a corpus-based unit selection text-to-speech system [31], and have worked on enhancing the naturalness of synthesized speech [32]. More recently, Mandeel et al. [33] demonstrate successful speaker adaptation experiments using Tacotron2, a state-of-the-art text-to-speech synthesis system.

These advances together show rapid progress in brain-to-speech synthesis, neural vocoders, and text-to-speech synthesis. It is anticipated that the integration of cutting-edge methods and innovative approaches will provide significant

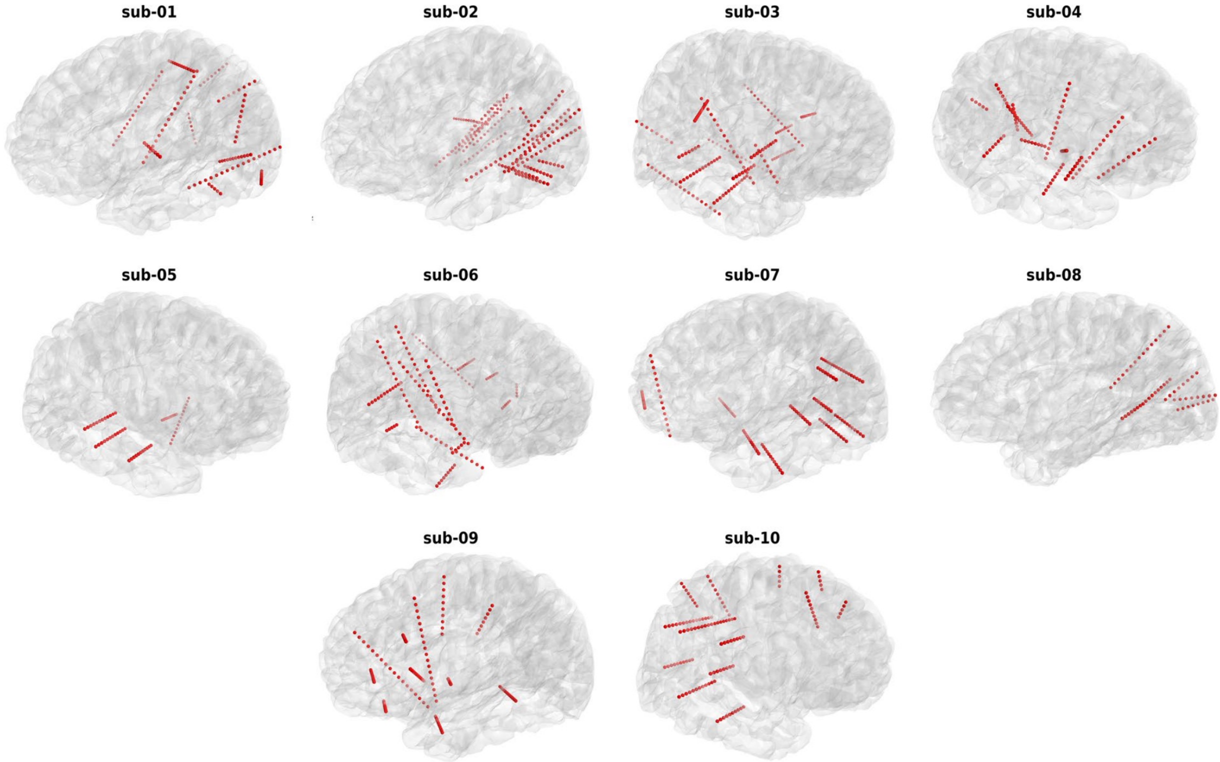

Fig. 1. The electrode locations of each participant were visualized on the surface reconstruction of their native anatomical MRI, as sourced from the SingleWordProduction-Dutch-iBIDS [7] dataset. Each red sphere in the figure represents an implanted electrode channel. This visualization is pivotal to our study as it illustrates the diverse and individualized placement of sEEG electrodes across participants, all of whom are part of the dataset used in this research. The variation in electrode placement is dictated by clinical requirements for treating epilepsy.

advancements in communications technology in the future, particularly for individuals with speech impairments.

III. METHODS

A. Data

We used the SingleWordProduction-Dutch-iBIDS dataset ([7], https://osf.io/nrgx6/) that contains in total 10 speakers with drug-resistant epilepsy (mean age 32.4 +/- 12.6 years; 5 male, 5 female). sEEG electrodes (Fig. 1.) were implanted as part of the clinical management of their epilepsy. The location of the electrodes was determined solely on the basis of clinical need. All participants were native Dutch speakers. Participants' voices were pitch-shifted to ensure anonymity. A total of 100 words were recorded for each participant, resulting in a total recording time of 300 seconds. Participants were implanted with platinum-iridium sEEG electrode arrays. Neural data were recorded using one or two Micromed SD LTM amplifier(s) with 128 channels each. Electrode connections were mapped to a common white matter contact. Data were recorded at 1024 Hz or 2048 Hz and downsampled to 1024 Hz. The audio was recorded at 48 kHz.

Recording of brain and speech signals using separate but time-aligned devices was already provided with the dataset. Synchronization is essential to ensure that each segment of EEG data corresponds to the specific speech output. This is achieved through a precise time-stamping process during recording, which aligns the EEG signals with the respective speech segments.

B. Preprocessing the brain and speech signals

On the sEEG brain signal, we followed a detailed preprocessing protocol as described in the publication we acquired the data set from [7], using the tools at https://github.com/neuralinterfacinglab/SingleWordProductionDutch/. Specifically, we executed several steps to refine the EEG data:

Extraction of the Hilbert Envelope: We targeted the high-frequency activity (70–170Hz) for each electrode contact using a bandpass filter (4th order IIR). This step was crucial for isolating significant neural activity relevant to speech processes. Hilbert transform provides several advantages for sEEG signal analysis, including the construction of analytic signals, extraction of instantaneous amplitude and phase information, improved time-frequency analysis, envelope detection, cross-frequency coupling analysis, and applicability to non-linear and non-stationary signals. These advantages can help better understand the underlying brain activity.

Attenuation of Line Noise: To minimize electrical interference, particularly the harmonics of 50Hz line noise, we employed two bandstop filters (4th order IIR).

Temporal Windowing and Stacking: We averaged the filtered signal over 50ms windows with a 10ms frameshift. To incorporate temporal context, which is vital for understanding the dynamics of brain activity, we stack features from multiple time windows. Specifically, for each time window of interest, we include features from the 4 preceding and 4 succeeding windows alongside the current window, totaling nine windows per feature set.

Normalization: For each feature, we normalized the data to zero mean and unit variance using the statistics from the training data. This normalization was then consistently applied to the evaluation data to maintain data integrity across different sets.

After preprocessing the sEEG signal, we calculate 80-dimensional mel-spectrogram of the speech using the 'librosa' library. During synthesis, we obtain the estimated speech using the WaveGlow model with inverse STFT transform [28], using a pre-trained model provided by NVIDIA, https://drive.google.com/file/d/1cjKPHbtAMh_4HTHmuIGNkbOkPBD9qwhj/view?usp=sharing.

Regarding the database split, we used a standard approach where the dataset was divided into training and testing subsets. Specifically, 80% of the data was used for training, and the remaining 20% for testing. This split was performed on a per-speaker basis, ensuring that the model's performance could be evaluated on unseen data from each subject.

C. Linear regression (baseline)

The baseline study [7] reconstructed the log-mel spectrogram from the high-gamma features using linear regression models. In these models, the high-gamma feature vector is multiplied with a weight matrix to reconstruct the log-mel spectrogram. The weights are determined using a least-squares approach. For the waveform reconstruction, they utilized the Griffin-Lim method.

D. Deep learning architectures

Next, we train the deep learning algorithms, which receive windowed sEEG Hilbert transformed components as input and produce 80-dimensional mel-spectral coefficients as output.

As for the hyperparameters, the learning rate, number of epochs, and other training parameters were selected through a series of preliminary experiments aimed at optimizing model performance. The number of epochs was set to 100, with early stopping using a patience of three, to prevent overfitting. The learning rate was initially set to a standard value of 0.001 and was adjusted based on the model's performance during the validation phase. Regarding learning rate scheduling, we used a dynamic approach where the learning rate was halved if there was no improvement in model performance on the validation set for two epochs.

Our method is illustrated in Figure 2, which shows the general flow from the raw sEEG input to the final synthetic speech. In order to obtain an analytical signal from the sEEG data, the Hilbert transform is used to acquire both amplitude and phase information (as detailed in Sec. III-B). We then apply the transformed signal as the input of our neural network models, including FC-DNN, 2D-CNN, and 3D-CNN. Based on the sEEG input, these models are trained to predict the mel-spectrograms of speech, thereby creating a mapping between brain activity and acoustic representations of speech. WaveGlow neural vocoder is used to convert the predicted mel-spectrogram into audible speech.

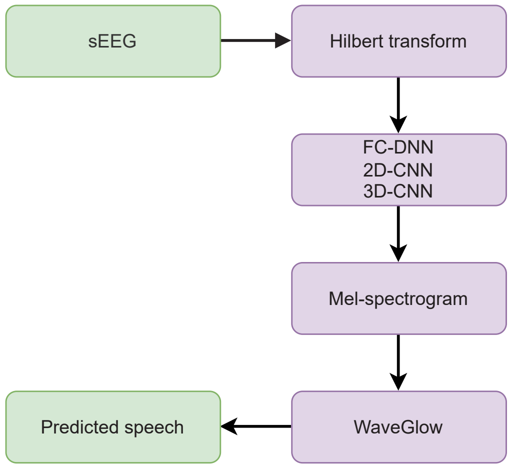

Fig. 2. General block diagram of our methods: from sEEG input, we predict mel-spectrogram of speech, which is synthesized to audio using a neural vocoder.

1) FC-DNN Architecture: We utilized a Fully Connected, Feed-Forward Deep Neural Network (FC-DNN) as our foundational model. This architecture incorporates five hidden layers, each consisting of 1000 neurons. We employed a Rectified Linear Unit (ReLU) as the activation function. The network's input layer has a dimensionality of 1143, which represents features calculated from a combination of 127 EEG channels and 9 temporal windows, as detailed in Section III-B. The output layer features 80 neurons, corresponding to the number of mel-spectral coefficients.

2) 2D-CNN: Our 2D convolutional network starts with two convolutional layers, each equipped with a 5x5 kernel size, having swish activation. The input data is formatted as 9x127 dimensions (9 temporal windows with 127 features in each). After a maxpooling layer, there is a third convolutional layer. The filter sizes are 30, 60 and 70. Dropout layers with a rate of 0.2 are used. Subsequent to the convolutional layers, the network architecture includes two fully connected layers. The first fully connected layer contains 1000 neurons. The final layer in our 2D-CNN model is the output layer, having linear activation, and designed with 80 neurons to match the number of mel-spectral bands for the waveform reconstruction.

3) 3D-CNN: Standard CNN considers 2D images to extract features by convolving 2D filters over images. Therefore, to model temporal information, a third dimension has to be considered [34], [35]. Here we use a 3D-CNN variation by adding a third dimension as (2+1)D CNN which shows good performance in video action recognition task [36]. It also shows good results when used with ultrasound images and it could be considered as a substitute of CNN+LSTM [37]. This network processed 5 frames of input that were 6 frames apart (6 is the stride parameter of the convolution along the time axis) [37]. Following the concept of (2+1)D convolution, the 5 frames were first processed only spatially, and then got combined along the time axis just below the uppermost dense

TABLE I
MCD SCORES ON THE TEST SET.

Speaker	Mel-Cepstral Distortion (dB)			
	Linear Regression with Griffin-Lim	FC-DNN with WaveGlow	3D-CNN with WaveGlow	2D-CNN with WaveGlow
sub-01	6.25	4.63	4.86	4.64
sub-02	6.41	4.95	5.19	4.98
sub-03	5.52	4.39	4.50	4.51
sub-04	5.28	4.16	4.86	4.50
sub-05	6.20	6.12	6.08	6.39
sub-06	4.36	3.63	4.16	4.10
sub-07	5.50	4.32	5.39	4.31
sub-08	5.03	5.00	5.50	5.13
sub-09	5.12	4.29	5.56	5.15
sub-10	4.26	4.01	4.34	4.13
Mean	5.39	4.55	5.04	4.78

layer.

Our 3D model begins with an input layer that handles the reshaped sEEG data, formatted into a 9x127 dimension. To accommodate the 3D processing, the data is expanded into a five-dimensional structure, ensuring compatibility with the subsequent 3D convolutional layers. The core of our 3D-CNN comprises three convolutional layers, each utilizing a kernel size of (5, 13, 13), strides set to (6, 2, 2), and having swish activation. These layers are designed to extract and analyze both spatial and temporal features from the sEEG data. There is a maxpooling layer after the second convolution. The filter sizes are 30, 60 and 70. Dropout layers with a rate of 0.2 are used. Subsequent to the convolutional layers, the network architecture includes two fully connected layers, similarly to the 2D-CNN, finally predicting the 80-dimensional mel-spectrogram.

After the trainings with the above deep neural networks, the predicted spectrograms of the test data are used to synthesize speech using the WaveGlow vocoder (Sec. III-B).

IV. Results

A. Demonstration sample

Fig. 3 a) shows the spectrogram of a natural utterance and b–e) those of synthesized speech from sEEG input with linear regression (baseline from [7]) and various DNNs. The synthesized speech has a similar envelope as the natural speech, but few of the spectral details are included. Although the speech reconstructed from the mel-spectral parameters estimated on the test pile resembles the original speech, it is noisy and difficult to understand. However, in some parts, sections of synthesized speech (e.g. vowels) are similar to the original audio. Synthesized samples are available at http://smartlab.tmit.bme.hu/icj2023_sEEG.

B. Objective evaluation

To check whether the proposed DNNs can reproduce the features of the original speech, we evaluated the spectral differences between natural speech and synthesized speech using Mel-Cepstral Distortion (MCD) [38], which is a standard metric for text-to-speech synthesis evaluation. As MCD is an error measure, lower values indicate higher similarity between the original and synthesized speech. Table I displays the MCD values calculated on the test data for each speaker. In

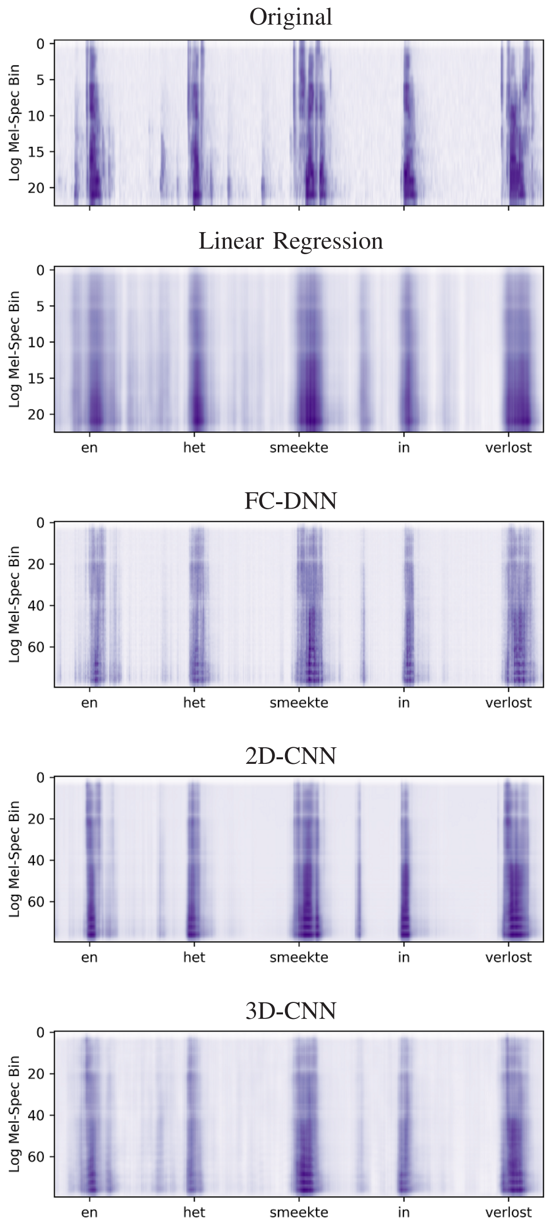

Fig. 3. Speech samples from speaker sub-06: a) original, b) synthesized using LR (baseline) c) FC-DNN, d) 2D-CNN, e) 3D-CNN.

combination with a WaveGlow vocoder, the Fully Connected Deep Neural Network (FC-DNN) model consistently produced the lowest MCD values across all speakers tested. Therefore, this combination of models and vocoders is the most effective means of reproducing speech that is resembling the original.

It is interesting to note the variation in MCD values among different speakers. For instance, speaker sub-06 consistently showed lower MCD values across all models, indicating that the acoustic features of this speaker might be easier for the models to learn and reproduce. This observation suggests that individual characteristics of each speaker's data, and most probably, the electrode positions can significantly influence the performance of speech synthesis models. The comparison with the correlations in [7] provided intriguing insights. Speakers with higher brain-speech signal correlation generally had lower MCD values, reinforcing the potential link between these two metrics. Prior studies [15] have also suggested a possible connection between neural correlates and the quality of speech synthesis.

C. Subjective evaluation

In order to determine which proposed version is closer to natural speech, we conducted an online MUSHRA-like test [39].

Our aim was to compare the natural words with the synthesized words of the baseline and the proposed approaches. In the test, the listeners had to rate the naturalness of each stimulus in a randomized order relative to the reference (which was the natural utterance), from 0 (very unnatural) to 100 (very natural). Out of the 10 speakers used in the earlier analysis, we selected four speakers for the listening test, based on the correlation analysis between brain and speech signals (Fig. 4 of [7]): 'sub-04/F', 'sub-06/M' (high correlation), and 'sub-01/F', 'sub-02/M' (low correlation). We selected four words from the test set of each speaker (altogether 16 words, each being 2 seconds long). The variants appeared in randomized order (different for each listener).

Each word was rated by non-Dutch speakers: altogether 9 listeners participated in the test; 7 males, 2 females; ages: 23-39 (avg: 32). The test took 5–28 minutes (avg: 11 minutes) to complete. Fig. 4 top shows the average naturalness scores for the tested approaches. The benchmark (Linear Regression) version achieved the lowest scores, while the natural words were rated the highest, as expected. The proposed DNN and neural vocoder based versions were performed over the baseline system for all speakers. In the overall figure, we can see a slight preference towards the FC-DNN, compared to the convolutional neural networks. To check the statistical significances, we conducted Mann-Whitney-Wilcoxon ranksum tests with a 95% confidence level. Based on this, the differences between FC-DNN, 2D-CNN, and 3D-CNN are not statistically significant.

When vizualizing the results speaker by speaker (Fig. 4 bottom), we can see the following trends: for the female speakers (sub-01 and sub-04), the 2D-CNN was preferred most, whereas this is not the case for the male speakers (sub-02 and sub-06). Based on the earlier correlation analysis on

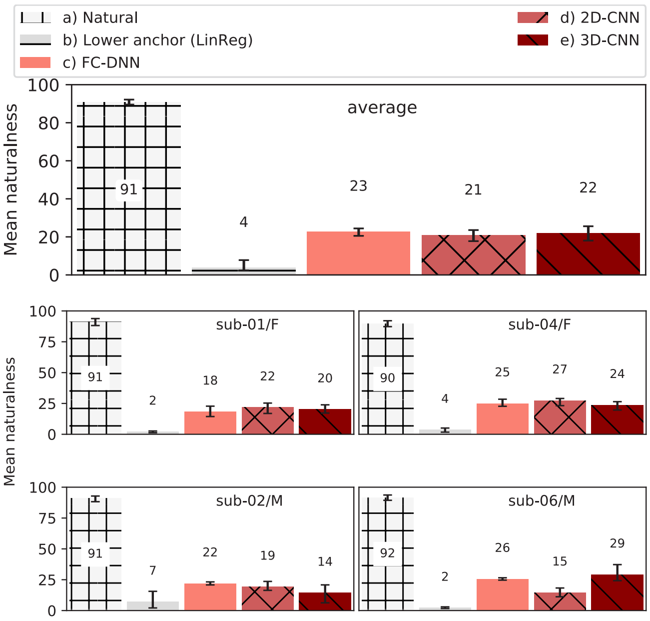

Fig. 4. *Results of the subjective evaluation with respect to naturalness, speaker by speaker (top) and average (bottom). The errorbars show the 95% confidence intervals.*

the speakers in [7] we have seen that sub-04 and sub-06 had a higher overall correlation between brain and speech signals, and this is clearly reflected in the speaker-by-speaker results of the listening test: both of them achieved reasonably higher naturalness scores compared to sub-01 and sub-02. Regarding MWM ranksum tests, the only case when the results are statistically significant, is sub-06: here, the 2D-CNN was ranked significantly lower than FC-DNN and 3D-CNN, while the difference between the latter two is not significant, but 3D-CNN is slightly preferred.

As a summary of the evaluation, the objective MCD score was not always found to be helpful in our case (i.e, it does not highly correspond to the correlations of [7]), but clearly, the subjective listening test could show the differences between the speakers of low and high correlation. The relatively low naturalness scores (18–29) indicate that sEEG-based synthesized speech is far from being intelligible, but at least, has properties similar to the natural speech signal.

V. Discussion and Conclusions

In this paper, we applied deep neural networks (FC-DNN, 2D-CNN, and 3D-CNN) for sEEG-to-melspectrogram prediction. Next, we synthesized speech using the WaveGlow neural vocoder. Our objective evaluation (Mel-Cepstral Distortion) has shown that the DNN-based approaches with neural vocoder outperform the baseline linear regression model using Griffin-Lim for speech generation [7].

Various studies have demonstrated the feasibility of ECoG-to-text [40] and ECoG-to-speech [15] conversion using different methodological approaches, such as linear regression and deep neural networks. However, their applicability in sEEG-to-speech conversion remained largely unexplored. Our work, therefore, complements these efforts and provides an alternative approach to sEEG-to-speech synthesis. Compared to traditional methods such as Griffin-Lim, neural vocoders represent an advance in generating more natural-sounding speech than traditional methods. While the complexity of sEEG data presented significant challenges, our approach utilizing deep neural networks and a neural vocoder showed promising results in comparison to the baseline linear regression model.

However, we acknowledge that the quality of synthesized speech remains an area for improvement. Our models produced speech that has distinct speech-like characteristics but was not yet fully understandable. This is a common issue encountered in the field of brain-to-speech synthesis, including studies utilizing EEG and ECoG data.

The reason why the 2D-CNNs and 3D-CNNs produced samples with larger errors in the current study might be that the amount of training data is extremely limited (i.e., only 100 words / 300 seconds), and more complex networks cannot learn the necessary mapping. Another explanation for the low 2D-CNN and 3D-CNN results might be that as our sEEG input data is put together in a specific way (i.e, brain signal is windowed, and Hilbert-transformed values are stacked together), this type of image is difficult to process for a convolutional neural network. On the other hand, the differences are highly dependent on the speaker (and thus, most probably on the electrode positioning) : with sub-06, who had the highest correlations in [7], the 3D-CNN performed best, indicating that there is potential in applying convolutional neural networks for this task.

Both the subjective listening tests and objective evaluations show that the neural network-based approaches outperformed the linear regression baseline. The relatively low naturalness scores (18–29) indicate that sEEG-based synthesized speech is far from being intelligible, but clearly, has properties similar to the natural speech signal, both visually on the spectrograms, and when listening to the samples. Therefore, we expect that our results might help future speech-based Brain-Computer Interfaces.

VI. FUTURE WORK

Deep learning is vast and ever-evolving, providing ample opportunity to refine our sEEG-to-speech prediction models. One approach to enhance the current results could involve experimenting with different architectures and types of deep learning models. For instance, Transformer models [41], known for their effectiveness in various natural language processing tasks, could be explored for sEEG-to-speech synthesis. We may be able to gain valuable insights into how different brain regions contribute to speech production through the attention mechanism in Transformers, potentially enabling us to improve our predictive abilities [41]. We acknowledge that the efficacy of complex models like Transformers is contingent on the availability of substantial training data. However, we expect that as more and more research groups are dealing with speech and brain signal recording and processing, such larger datasets might be available in the future.

Our feature extraction process currently involves windowing the raw sEEG data and applying the Hilbert transform. However, future work could involve more sophisticated feature extraction techniques like Wavelet Transform [42] or Fourier Transform [43]. These techniques could capture different aspects of the sEEG signals, leading to improved performance of the models [44].

In terms of data, our current study is based on the SingleSpeechProductionDutch dataset [7]. While this dataset has provided valuable insights, we recognize the potential benefits of using a more extensive and diverse dataset. Consequently, we intend to record our database, expanding the pool of speakers and potentially improving the generalizability and robustness of the model. Nevertheless, it is important to note that we will use EEG signals rather than sEEG for our planned dataset, which may present new challenges and opportunities.

Furthermore, it may be beneficial to explore applying more advanced post-processing techniques. The WaveGlow neural vocoder is currently employed for speech synthesis, but future work could investigate the use of more recent vocoding techniques, like AutoVocoder [45], to enhance the quality of the speech synthesised.

The positions of sEEG electrodes in the dataset were determined by clinical needs in the treatment of epilepsy, which can influence the quality of synthesized speech [7]. This is supported by existing literature, which shows that electrodes placed closer to key speech areas, particularly in the left frontal lobe, are more likely to capture neural signals that are crucial for accurate speech synthesis. This theoretical understanding, underpinned by neurophysiological insights into speech production processes, suggests that variations in electrode arrangements could result in differences in the quality of synthesized speech. However, a detailed correlation analysis between electrode positions and synthesized speech quality was beyond the scope of our current study, presenting a valuable direction for future research.

Finally, we see many potential applications for sEEG-to-speech synthesis in the future. Due to rapid advances in deep learning, we anticipate improving our models and contributing to the development of speech-based Brain-Computer Interfaces in the future, as well as improving their performance.

VII. ACKNOWLEDGEMENTS

The research was partially funded by the National Research, Development and Innovation Office of Hungary (FK 142163 grant). T.G.Cs. was supported by the Bolyai János Research Fellowship of the Hungarian Academy of Sciences and by the ÚNKP-22-5-BME-316 New National Excellence Program of the Ministry for Culture and Innovation from the source of the National Research, Development and Innovation Fund.

REFERENCES

[1] D. Dupré and A. Karjalainen, “Employment of disabled people in Europe in 2002,” *Statistics in focus*, pp. 3–26, 2003.

[2] Hungarian Central Statistical Office, “2011. évi népszámlálás, 11. fogyatékossággal élők,” *Tech. Rep.*, 2011. [Online]. Available: http://www.ksh.hu/docs/hun/xftp/idoszaki/nepsz2011/nepsz_11_2011.pdf

[3] M. Lecerf, “Employment and disability in the European Union,” *European Parliamentary Research Service (EPRS)*, no. May, pp. 1–7, 2020.

[4] "White Rose restoring the larynx has https://www.whiterose.ac.uk/collaborationfunds/silent-speech-restoring-the-power-of-speech-to-people-whose-larynx-has-been-removed/

[5] E. F. Chang and G. K. Anumanchipalli, "Toward a speech neuroprosthesis," *JAMA*, vol. 323, no. 5, pp. 413–414, feb 2020, DOI: 10.1001/JAMA.2019.19813.

[6] D. J. McFarland and J. R. Wolpaw, "EEG-based brain–computer inter-faces," *Current Opinion in Biomedical Engineering*, vol. 4, pp. 194–200, dec 2017, DOI: 10.1016/J.COBME.2017.11.004.

[7] M. Verwoert, M. C. Ottenhoff, S. Goulis, A. J. Colon, L. Wagner, S. Tousseyn, J. P. van Dijk, P. L. Kubben, and C. Herff, "Dataset of speech production in intracranial electroencephalography," *Scientific Data 2022 9:1*, vol. 9, no. 1, pp. 1–9, jul 2022. Available: https://www.nature.com/articles/s41597-022-01542-9 [Online]. DOI: 10.1038/s41597-022-01542-9.

[8] G. Buzsáki, C. A. Anastassiou, and C. Koch, "The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes," *Nature Reviews Neuroscience*, vol. 13, no. 6, pp. 407–420, may 2012. Available: https://www.nature.com/articles/nrn3241 [Online]. DOI: 10.1038/nrn3241.

[9] D. Dash, P. Ferrari, A. Babajani-Feremi, A. Borna, P. D. Schwindt, and J. Wang, "Magnetometers vs Gradiometers for Neural Speech Decoding," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2021, pp. 6543– 6546, nov 2021, DOI: 10.1109/EMBC46164.2021.9630489.

[10] F. Lotte, L. Bougrain, and M. Clerc, "Electroencephalography (EEG)-Based Brain–Computer Interfaces," *Wiley Encyclopedia of Elec- trical and Electronics Engineering*, pp. 1–20, sep 2015, DOI: 10.1002/047134608X.W8278.

[11] A. J. Casson, "Wearable EEG and beyond," *Biomedical engineering letters*, vol. 9, no. 1, pp. 53–71, feb 2019. Available: https://pubmed.ncbi.nlm.nih.gov/30956880/ [Online]. DOI: 10.1007/S13534-018-00093-6.

[12] T. G. Csapó, F. V. Arthur, P. Nagy, and A. Boncz, "A beszéd artikulációs mozgásának predikciója agyi jel alapján – kezdeti eredmények," in *XIX. Magyar Számítógépes Nyelvészeti Konferencia, MSZNY 2023*, 2023, pp. 357–368. [Online]. Available: https://m2.mtmt.hu/api/publication/33599995

[13] F. V. Arthur and T. G. Csapó, "Deep learning alapú agyi jel feldolgozás és beszédszintézis előkészítő munkálatai," in *XVIII. Magyar Számítógépes Nyelvészeti Konferencia: MSZNY 2022,* 2022, pp. 185–198. [Online]. Available: https://m2.mtmt.hu/api/publication/32636136

[14] J. Chartier, G. K. Anumanchipalli, K. Johnson, and E. F. Chang, "Encoding of articulatory kinematic trajectories in human speech sensorimotor cortex," *Neuron,* vol. 98, pp. 1042–1054.e4, 6 2018, DOI: 10.1016/j.neuron.2018.04.031.

[15] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493– 498, 4 2019, DOI: 10.1038/s41586-019-1119-1.

[16] J. Brumberg, E. Wright, D. Andreasen, F. Guenther, and P. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex," *Frontiers in Neuroscience*, vol. 5, 2011, DOI: 10.3389/fnins.2011.00065.

[17] G. Le Godais, "Decoding speech from brain activity using linear methods," Theses, Université Grenoble Alpes [2020-....], Jun. 2022. [Online]. Available: https://theses.hal.science/tel-03852448

[18] F. H. Guenther, J. S. Brumberg, E. J. Wright, A. Nieto-Castanon, J. A. Tourville, M. Panko, R. Law, S. A. Siebert, J. L. Bartels, D. S. Andreasen, P. Ehirim, H. Mao, and P. R. Kennedy, "A wireless brain-machine interface for real-time speech synthesis," *PLOS ONE*, vol. 4, no. 12, pp. 1–11, 12 2009, DOI: 10.1371/journal.pone.0008218.

[19] S. Lesaja, C. Herff, G. D. Johnson, J. J. Shih, T. Schultz, and D. J. Krusienski, "Decoding lip movements during continuous speech using electrocorticography," in *2019 9th International IEEE/EMBS University power Consortium – Silent of speech to people Speech: whose Available: been removed." [Online]. Conference on Neural Engineering (NER)*, 2019, pp. 522–525, DOI: 10.1109/NER.2019.8716914.

[20] S. Luo, Q. Rabbani, and N. E. Crone, "Brain-computer interface: Applications to speech decoding and synthesis to augment communication," *Neurotherapeutics 2022*, vol. 1, pp. 1–11, jan 2022, DOI: 10.1007/S13311-022-01190-2.

[21] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," in *Proc. ICASSP*, online, 2020, pp. 1235–1238, DOI: 10.1109/ICASSP40776.2020.9053340.

[22] G. Krishna, C. Tran, M. Carnahan, and A. H. Tewfik, "Advancing speech synthesis using EEG," *International IEEE/EMBS Conference on Neural Engineering, NER*, vol. 2021-May, pp. 199–204, may 2021, DOI: 10.1109/NER49283.2021.9441306.

[23] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski, P. L. Kubben, T. Schultz, and C. Herff, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications Biology*, vol. 4, no. 1, p. 1055, 2021, DOI: 10.1038/s42003-021-02578-0.

[24] S. Lesaja, M. Stuart, J. J. Shih, P. Soroush, T. Schultz, M. Manic, and D. J. Krusienski, "Self-supervised learning of neural speech representations from unlabeled intracranial signals," *IEEE Access*, 2022, DOI: 10.1109/ACCESS.2022.3230688.

[25] C. Wang, V. Subramaniam, A. U. Yaari, G. Kreiman, B. Katz, I. Cases, and A. Barbu, "BrainBERT: Self-supervised representation learning for intracranial recordings," 2023, DOI: 10.48550/arxiv.2302.14367.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2019, DOI: 10.48550/arxiv.1810.04805.

[27] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A Generative Model for Raw Audio," *CoRR*, vol. abs/1609.0, 2016, DOI: 10.48550/arXiv.1609.03499.

[28] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 3617–3621, DOI: 10.1109/ICASSP.2019.8683143.

[29] T. G. Csapó, Zainkó, Cs., L. Tóth, G. Gosztolya, and A. Markó, "Ultrasound-based articulatory-to-acoustic mapping with WaveGlow speech synthesis," in *Proc. Interspeech, 2020*, pp. 2727–2731, DOI: 10.21437/Interspeech.2020-1031.

[30] B. Cao, A. Wisler, and J. Wang, "Speaker adaptation on articulation and acoustics for articulation-to-speech synthesis," *Sensors*, vol. 22, no. 16, p. 6056, 2022, DOI: 10.3390/S22166056.

[31] T. G. Csapó, Cs. Zainkó, and G. Németh, "A study of prosodic variability methods in a corpus-based unit selection text-to-speech system," *Infocommunications Journal*, vol. LXV, p. 2010, 01 2010.

[32] T. G. Csapó, G. Németh, and M. Fék, "Szövegfelolvasó természetességének növelése," *Híradástechnika*, vol. LXIII, p. 2008, 05 2008.

[33] A. R. Mandeel, M. S. Al-Radhi, and T. G. Csapó, "Speaker adaptation experiments with limited data for end-to-end text-to-speech synthesis using tacotron2," *Infocommunications Journal*, vol. 14, pp. 55–62, 2022, DOI: 10.36244/ICJ.2022.3.7.

[34] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012, DOI: 10.1109/TPAMI.2012.59.

[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, nov 1997, DOI: 10.1162/neco.1997.9.8.1735.

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459, DOI: 10.48550/arXiv.1711.11248.

[37] L. Tóth and A. H. Shandiz, "3D Convolutional Neural Networks for Ultrasound-Based Silent Speech Interfaces," in *Proc. ICAISC*, Zakopane, Poland, 2020, DOI: 10.48550/arXiv.2104.11532.

[38] R. F. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, Victoria, Canada, 1993, pp. 125–128, DOI: 10.1109/pacrim.1993.407206.

[39] "ITU-R Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality," 2001.

[40] C. Herff, D. Heger, A. de Pesters, D. Telaar, P. Brunner, G. Schalk, and T. Schultz, "Brain-to-text: decoding spoken phrases from phone representations in the brain," *Frontiers in Neuroscience*, vol. 8, 2015, DOI: 10.3389/fnins.2015.00217.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 2017-Decem. Neural information processing systems foundation, jun 2017, pp. 5999–6009, DOI: 10.48550/arxiv.1706.03762.

[42] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, Summer 1995, DOI: 10.1109/99.388960.

[43] R. Bracewell, *The Fourier Transform and Its Applications*, ser. Circuits and systems. McGraw-Hill, 1978.

[44] A. K. Singh and S. Krishnan, "Trends in EEG signal feature extraction applications," *Frontiers in Artificial Intelligence*, vol. 5, p. 1072801, jan 2023, DOI: 10.3389/FRAI.2022.1072801/BIBTEX.

[45] J. J. Webber, C. Valentini-Botinhao, E. Williams, G. E. Henter, and S. King, "Autovocoder: Fast Waveform Generation from a Learned Speech Representation using Differentiable Digital Signal Processing," in *Proc. ICASSP*, Rhodes, Greece, 2023, DOI: 10.48550/arxiv.2211.06989.

Frigyes Viktor Arthur is a PhD student at BME in the field of computer science. He has a background in biomedical engineering, with a master's degree, and has significant programming experience from various side-projects mainly related to medical image processing, such as Intracranial Hemorrhage Detection, and native and cross-platform Android and iOS mobile application development. His research interests include deep learning-based brain signal and speech processing, Brain-Computer Interfaces, EEG, ECG, and multimodal biological signals.

Tamás Gábor Csapó, PhD is currently a senior research fellow at the Speech Technology and Smart Interactions Laboratory (BME-SmartLab), Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics. He received his PhD at BME in 2014 about text-to-speech synthesis, with Géza Németh as the supervisor. Meanwhile, in 2014, he was a Fulbright scholar at Indiana University in Dr. Steven Lulich's lab, where he started his research on ultrasound tongue imaging and automatic contour tracking. Between 2016–2021, he was an active member of MTA–ELTE Lingual Articulation Research Group, where he started Hungarian articulatory research using ultrasound). Between 2017–2022, he had two national OTKA postdoc projects (FK-17 and PD-18) on articulatory-to-acoustic mapping using ultrasound/lip/vocal tract MRI. Currently, he is the PI of an FK-22 OTKA project on the analysis of articulation and brain signals for speech-based brain-computer interfaces, using EEG and UTI. He has 150+ publications, most of them in high-ranked conferences (e.g., Interspeech or ICASSP) and in top-ranked journals (e.g., IEEE Selected Topics in Signal Processing, Journal of the Acoustical Society of America – Express Letters).