Proceedings of the
9th Graz Brain-Computer Interface Conference 2024

10.3217/978-3-99161-014-4-084

USING TRANSFORMER NETWORKS FOR STREAMING SPEECH SYNTHESIS FROM INTRACRANIAL EEG

Joaquín Amigó-Vega[1, 2 †], Maxime Verwoert[2], Maarten C. Ottenhoff[2], Pieter L. Kubben[2], Christian Herff[2]

[1]Gran Sasso Science Institute, Computer Science Department, L'Aquila, Italy
[2]Neural Interfacing Lab, Department of Neurosurgery, Mental Health and Neuroscience Research Institute, Maastricht University, Maastricht, the Netherlands

† joaquin.amigo@gssi.it

ABSTRACT: Speech Neuroprostheses have the potential to enable users to communicate without the need for overt muscle movement. Several recent approaches have demonstrated the feasibility of decoding textual and acoustic representations of speech from invasively measured neural activity. However, most approaches decode or synthesize speech after several seconds or complete utterances. While this provides tremendous communicative ability to patients, it lacks the full expressive power of natural conversations. Ideally, a speech neuroprosthesis would synthesize speech without a noticeable delay.

Here, we present a real-time speech decoding pipeline that generates speech output in a streaming fashion, i.e., with delays of less than 40 ms. Intracranial EEG data is measured, processed, decoded, and synthesized into an audio waveform using our fast and modular framework. Notably, we employ a Transformer architecture for the decoding step from neural features to a spectral representation of speech.

INTRODUCTION

Speech plays an important role in human interaction, serving as a primary means of conveying thoughts and emotions. It is integral to the fabric of our social existence and personal identity. However, various conditions, such as Amyotrophic Lateral Sclerosis (ALS) and locked-in syndrome, can impair one's ability to speak, significantly impacting the quality of life. These diseases may leave cognitive functions intact, while debilitating the muscular activity required for speech production.

Speech Brain-Computer Interfaces (BCIs), also called speech neuroprostheses, are a groundbreaking technology designed to help people in need. By harnessing neural signals through invasive or noninvasive methods, these BCIs decode speech-associated brain activity. This process involves extracting and translating neural patterns related to speech formation into actionable outputs, thereby enabling communication or device control.

The ultimate goal of a speech BCI is to facilitate seamless, naturalistic conversation, akin to normal speech. Achieving this requires real-time processing of neural signals, a technical challenge that remains at the forefront of current research. Despite ongoing advancements, many existing speech BCI systems rely on offline evaluations, where signal analysis and method validation occur after data collection [1–5]. While recent studies have made strides toward closed-loop systems capable of generating textual representations [6–8] or synthesized sentences [9], these technologies typically operate with delays, processing complete sentences or phrases [10] before producing output. For completely natural communication, the patient needs to produce speech output immediately to ensure natural flow, e.g. to interrupt the conversational partner or to modulate their own speech.

This paper introduces a novel real-time streaming synthesis pipeline for speech BCIs, distinguished by its low latency and modular framework. Developed in Python and based on the framework Timeflux [11], our pipeline processes and decodes neural data into a speech waveform with less than 40 ms of delay. Longer delays have been found to severely impair speech production [12]. Notably, our system employs a transformer encoder to translate sequences of neural data into speech spectral sequences. The attention mechanism [13] in transformers is particularly well suited for learning the temporal dynamics in the neural and speech data and has successfully been used on offline data before [14].

To validate our streaming speech BCI, we conducted simulated online studies using a previously recorded dataset of intracranial EEG during speech production [15].

MATERIALS AND METHODS

Participants:

Our closed-loop experiments are conducted with voluntary participants implanted with sEEG electrodes as part of the clinical therapy for their pharmaco-resistant epilepsy. All participants gave written informed consent before joining the study, and the electrode locations were purely determined based on clinical necessity. All participants were Dutch native speakers and had normal speech, hearing, and language functions.

For this simulated online evaluation, we employ our pre-

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-084

viously published open-access Single Word Production Dutch-iBIDS (SWPD) dataset [15], consisting of 10 participants speaking 100 words each.

Data recording:

Patients were implanted with platinum-iridium sEEG electrode shafts (Microdeep intracerebral electrodes; Dixi Medical, Beçanson, France) with a diameter of 0.8 mm, a contact length of 2 mm and an inter-contact distance of 1.5 mm with each shaft containing between 5 and 18 contacts. Neural data was recorded using two or more Micromed SD LTM amplifier(s) (Micromed S.p.A., Treviso, Italy) with 64 channels each. Electrode contacts were referenced to a common white matter contact. Data was recorded at either 1024 Hz or 2048 Hz.

Simulated Online Experiment:

To assess the real-time capabilities of our closed-loop speech decoding pipeline, we conducted simulated online experiments using the SWPD dataset [15]. The dataset's recording environment mirrors the anticipated operational scenario for our pipeline, making it an ideal choice for our evaluation process. As part of the assessment, we divided the data from each participant into training and testing sets, allocating 75% for model training and the remaining 25% for testing. After training the pipeline with the designated data, we streamed the testing dataset through LabStreamingLayer (LSL), emulating the amplifier characteristics used in our real setup. This approach was designed to closely replicate the dynamics of real neural signal acquisition and processing, thereby providing a realistic approximation of how the pipeline would perform in live application scenarios.

Closed-loop pipeline:

Pipeline Design and Requirements

In the initial phase of constructing our speech Brain-Computer Interface (BCI) pipeline, we focused on identifying key requirements to ensure its effectiveness for real-time communication. Among our primary objectives were ensuring *real-time decoding*, *rapid model training*, and a high degree of *modularity and configurability*. Real-time decoding is crucial as the pipeline must process neural signal samples swiftly to minimize latency, thereby enabling near-instantaneous speech synthesis. Given the constraints of on-site training, it was imperative that the machine learning models employed could be trained quickly to avoid reducing valuable data collection time with participants. Additionally, to facilitate rapid experimentation and adaptation of new approaches, the system architecture needed to be both modular and easily configurable.

Framework and Technology Selection

We used Python and the framework Timeflux [11] for building the pipeline. Timeflux facilitates the creation of applications as directed acyclic graphs (DAGs), where processing nodes are interconnected through YAML syntax, enabling efficient data flow and simultaneous processing. For communication between graphs we used ZeroMQ, an asynchronous messaging library, ensuring robust data exchange without interrupting the execution.

Pipeline Architecture

The pipeline involves two main stages: *Initialization* and *Real-time decoding*. Each stage consists of a series of graphs and nodes executed concurrently, optimizing data processing speed.

Initialization Stage: This stage prepares the system for the online decoding. It uses the open-loop recorded experiment data, which includes synchronized neural signals, audio, and markers, to extract the relevant parameters and train the machine learning models.

First, we segregate the data into distinct datasets labeled "neural" and "audio" and adjust their format, length, and type.

Irrelevant channels, such as clinical markers and heart-rate, are eliminated from the "neural" dataset, and the power line noise and its first harmonic are filtered out, using causal IIR bandstop-filters. Afterwards, the signal is extracted in a broadband high-frequency range ($70-170$ Hz) and windowed, subsequently calculating the log power for each window. At the same time, another graph extracts the "audio" features by decimating the audio signal to lower its sampling rate, then using a sliding window with the same window size and frameshift as the neural data to extract a mel-scaled spectrogram, aligned to the neural features.

The size of the window used for the audio and neural data can be different. However, the shift needs to always be the same, allowing seamless alignment between the features.

Both feature sets are aligned and scaled before being used to train the Machine Learning model. When the training is completed, we save the model parameters and additional helpful metadata for the decoding stage.

Real-Time Decoding Stage: This stage, presented in Fig. 1, is responsible for the on-the-fly decoding of neural signals into audible speech. It encompasses system initialization, data intake, feature extraction, neural decoding, audio reconstruction, and finally, data management and preservation.

Initially, all nodes remain inactive and await the initialization parameters saved during the first stage. After reading the parameters, they are broadcast to all the pipeline's nodes while the LSL flow of incoming "neural" data is paused. The pause lasts 10 s and ensures that all the nodes are ready to promptly process the data once the flow resumes.

After the flow of LSL-streamed "neural" data is resumed, the features are extracted. The process is similar to the one described in the *Initialization Stage* where the data is filtered, windowed, and log power is extracted.

The features are scaled and fed into the Transformer model to obtain audible acoustic representations. We use 1.3 s of features to produce a representation of that same size, however, only the last 34.69 ms are passed to the following synthesis stage.

Subsequently, the Griffin-Lim algorithm transforms these spectral representations into an audio waveform that is

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-084

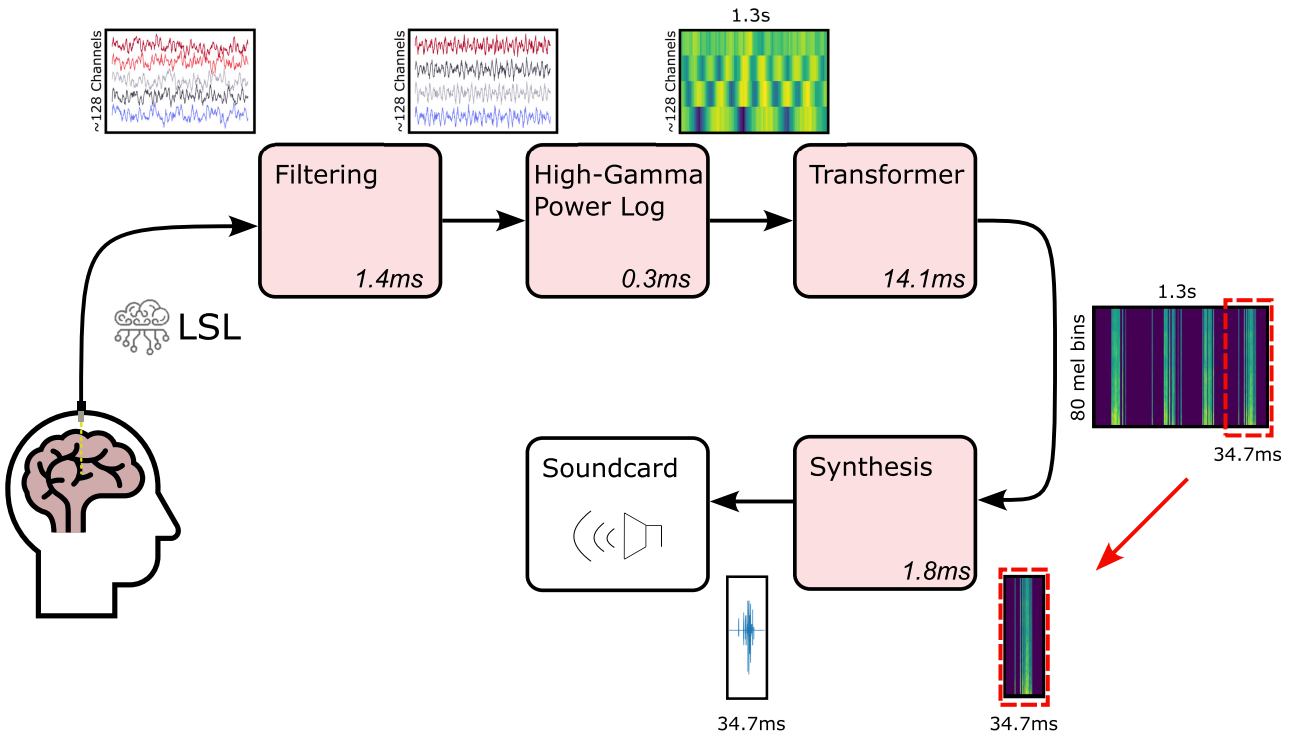

Figure 1: Illustration of the process of converting neural signals into audible speech. The system initializes by setting up nodes and pausing data intake. Neural data is then resumed and processed—filtered, feature-extracted, and transformed into a 1.3-second audio representation by a Transformer model, with only the last 34.7 ms used for sound synthesis via the Griffin-Lim algorithm. The generated waveform is sent to a sound card for playback. Concurrently, a data management system ensures the integrity and continuity of the operation without data loss.

send to the sound card for immediate auditory feedback.

We include an efficient data management paradigm that prevents data loss and does not disrupt the pipeline operation or hinder its speed. The paradigm includes a resource-friendly data-saving routine executed concurrently to maintain operational speed and integrity.

Transformer Architecture:

As mentioned, the pipeline can be easily configured to extract different neural and audio features and use different Machine Learning models. Here, we present initial results with a real-time-ready transformer architecture.

A Transformer model is an advanced neural network architecture that excels in processing data sequences using self-attention mechanisms [13]. These mechanisms permit the efficient extraction of hidden context and relationships within data. Transformers are highly efficient, scalable, and flexible, making them superior for tasks requiring a deep understanding of complex relationships. This is why they have become the foundation for many state-of-the-art solutions in natural language processing and beyond. Transformers have also been used in decoding speech from offline data successfully [14].

In our context of having a small amount of time-series data with limited time to train the model, it is challenging to use a Transformer because they typically require large amounts of data and significant computational resources to effectively learn the complex patterns and relationships in time-series. Their architecture, designed for capturing long-range dependencies, struggles to generalize from small datasets without overfitting and may not achieve optimal performance within a short training time-frame.

Despite these hurdles, we used a Transformer model to reconstruct auditory data from neural signals. Using only the self-attention mechanism and the encoder block, the model focuses on efficiently extracting and analyzing temporal features [16]. This approach reduces computational demands and training time, while still capturing complex patterns with less risk of overfitting. Focusing on prediction rather than sequence generation aligns the model's strengths directly with the requirements of time-series analysis, making it better suited for our tasks.

A challenge in real-time decoding with a sequential model lies in balancing the need to analyze significant temporal contexts to accurately decode complex patterns against the constraints of immediate processing. Recently published BCI works address this challenge by recording a large enough sequence of neural data and then producing the mapping to reproducible audio or text [10, 17].

Processing extensive historical data introduces latency for real-time applications like audio synthesis from neural signals, which conflicts with our real-time requirements. Our proposed solution, which exploits an idea presented by Shigemi *et al.* [18] involves using a predefined large-enough context size for analysis but synthesizing only the latest segment of the sequence. This approach allows the

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-084

Transformer to leverage enough historical data for accurate predictions, while maintaining the ability to produce outputs in real-time. It effectively addresses the challenge of adapting sequence-to-sequence mapping for real-time decoding, ensuring accuracy and immediacy in applications such as closed-loop neural interfaces.

Fig. 2 presents our conceived model architecture with the most relevant parameters. Our model first maps all the F channels of the input sequence into a 125-dimensional space through a linear transformation. Then, we use six standard encoder layers [13], each containing two main components: a Multi-Head Attention and a Feed Forward neural network, followed by an Add & Norm step to facilitate layer normalization. The Multi-Head Attention mechanism has five attention heads, and the Feed Forward neural network has a dimensionality of 2048 on the inner layer. The output of the last encoder layer undergoes another linear transformation to match the desired output dimension. A dropout rate of 0.25 is applied throughout the network to prevent overfitting. We use Mean Squared Error (MSE) as the loss criterion, and the learning rate is set to a modest 5e-4, which balances the speed of convergence with the stability of the learning process.

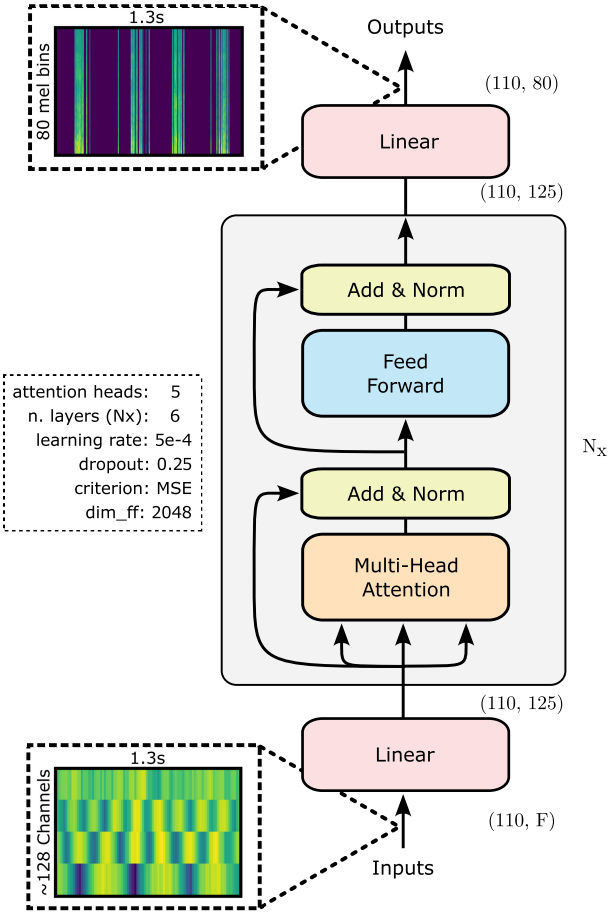

Figure 2: Transformer Model Architecture and Parameters. The model inputs are linearly transformed from F channels to a 125-dimensional space, followed by six encoder layers. Each layer consists of a Multi-Head Attention with five heads and a Feed Forward network with an inner-layer dimension of 2048, followed by an Add & Norm step. The final encoder output is linearly transformed to the desired output size. The model employs a dropout of 0.25, uses MSE as the loss function, and has a learning rate of 5e-4.

RESULTS

Given the constrained interaction duration with participants, it was critical to minimize model training times. The pipeline averaged approximately 133.11 s for model training, with a standard deviation of 12.00 s. This duration aligns well with our experimental requirements, offering a balanced compromise between training efficiency and subsequent decoding performance.

Processing latency per sample was another critical metric. Notably, each decoding operation by the transformer yields a 1.27-second audio window, from which only the latest 34.69 ms are utilized for audio reconstruction. To ensure near-real-time functionality, processing for each sample must therefore be completed in under 34.69 ms. Our performance results indicate an average processing time of approximately 17.62 ms per sample (standard deviation 1.01 ms), significantly below the 34.69 ms threshold. This efficiency meets our near-real-time criteria and provides flexibility for exploring other, more complex decoding approaches or even switching to higher-quality vocoders, such as HIFIGan [19] or VocGAN [20].

The qualitative aspect of our results involves the reconstruction of speech from neural signals. Fig. 3 aggregates the correlation outcomes across all participants, with data points indicating the correlation coefficient between the spectrograms of the original recorded and the synthesized audio for each individual, providing a visual representation of the decoding accuracy and variability among participants. Correlation are stable across the entire frequency spectrum of the mel-scale (Fig. 3 b), but vary dramatically between participants. Best results exceed average correlation coefficients of 0.66 (sub-06, Fig. 3 a).

DISCUSSION

The presented results, particularly concerning processing speeds and model training efficiency, precisely align with our pipeline's rapid training and real-time decoding objectives. This achievement highlights our pipeline's effectiveness in enabling real-time communication for individuals with speech impairments and its proficiency in decoding speech from new, unseen words. This latter capability underscores the system's robust generalization, a critical feature for practical Brain-Computer Interface (BCI) applications where pre-defining a comprehensive vocabulary is impractical.

Variations in decoding results across participants likely mirror the differential placement of sEEG electrodes. This suggests that proximity to speech-related brain areas might significantly influence both neural signal decoding quality and model training success. Notably, these

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-084

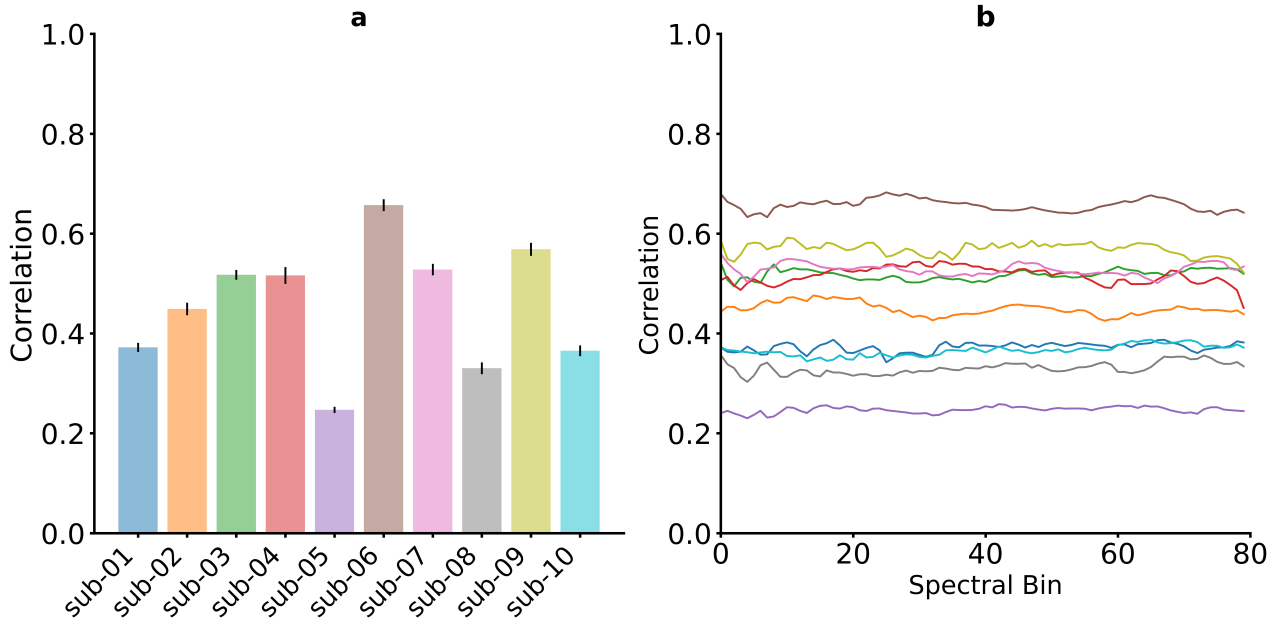

Figure 3: Correlation between original and reconstructed spectrograms. **a)** Mean correlation coefficients across all spectral bins for each participant, with error bars indicating the standard deviation. **b)** Mean correlation coefficients for each spectral bin.

achievements come from training on relatively limited data, approximately 225 s per participant. This is in stark contrast to the vast datasets employed in training the current state-of-the-art speech BCIs, which often utilize data ranging from dozens of minutes to several hours [6, 7, 10, 21], highlighting the efficiency and potential of our approach even with constrained datasets.

Currently, real-time reconstructed speech results are not intelligible, leaving room for further improvements in decoding approach and vocoder.

While the current results stem from simulations using the SWPD dataset, delineating the pipeline's capability for real-time speech decoding from neural signals, presenting online results falls outside this paper's scope. Nevertheless, addressing this gap is a priority in our ongoing research.

The promising outcomes achieved with the Transformer model open exciting future research directions, such as refining model architectures and devising new strategies to minimize further decoding latency. The 17.62 ms extra in processing time also permits using a more complex synthesizer that better reconstructs the audible speech from the spectrogram.

CONCLUSION

This paper introduced a closed-loop speech decoding pipeline designed for real-time operation. Our system is distinctively characterized by its low latency and modular framework, facilitating seamless, near-instantaneous communication. Utilizing Python and the Timeflux framework, we developed a modular pipeline that allows for swift prototyping and testing, catering to the dynamic needs of BCI research.

We demonstrated the feasibility of real-time streaming speech synthesis from neural signals through rigorous offline validations using aligned neural and audio recordings from the SWPD dataset. Our pipeline employs a Transformer model optimized for time-series data, achieving fast decoding speed and reasonable results. Despite the challenges associated with limited data availability and the constraints of working within clinical settings, our system managed to train models efficiently, with an average model training time of approximately 133.11 s and a decoding processing time of about 17.62 ms per sample, well below the threshold required for real-time functionality [12]. Notably, our streaming approach could allow for natural conversation, as sound is produced almost immediately, as opposed to other approaches which produce chunks of audio corresponding to whole sentences.

The qualitative results further underscore the efficacy of our pipeline in reconstructing audible speech. The reconstructed spectrograms and the correlation coefficients across participants highlight the potential of our technology to provide a voice for those who have lost their natural ability to speak due to neurological conditions.

Our work showcases an improvement in speech BCIs and opens new avenues for research and development toward more intuitive and accessible communication solutions. Future work will focus on enhancing the decoding results, reducing latency further, and expanding the system's adaptability. By continuing to refine and validate our pipeline, we aim to bring this technology closer to widespread clinical application, offering hope for improved quality of life for individuals with severe speech impairments.

Published by
Verlag der Technischen Universität Graz

10.3217/978-3-99161-014-4-084

REFERENCES

[1] C. Herff *et al.*, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Frontiers in neuroscience*, vol. 9, p. 217, 2015.

[2] M. Angrick *et al.*, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036 019, 2019.

[3] J. Berezutskaya, Z. V. Freudenburg, M. J. Vansteensel, E. J. Aarnoutse, N. F. Ramsey, and M. A. van Gerven, "Direct speech reconstruction from sensorimotor brain activity with optimized deep learning models," *Journal of Neural Engineering*, vol. 20, no. 5, p. 056 010, 2023.

[4] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[5] J. G. Makin, D. A. Moses, and E. F. Chang, "Machine translation of cortical activity to text with an encoder–decoder framework," *Nature neuroscience*, vol. 23, no. 4, pp. 575–582, 2020.

[6] F. R. Willett *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, 2023.

[7] S. L. Metzger *et al.*, "Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis," *Nature Communications*, vol. 13, no. 1, p. 6510, 2022.

[8] D. A. Moses *et al.*, "Neuroprosthesis for decoding speech in a paralyzed person with anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217–227, 2021.

[9] S. L. Metzger *et al.*, "A high-performance neuroprosthesis for speech decoding and avatar control," *Nature*, vol. 620, no. 7976, pp. 1037–1046, 2023.

[10] M. Angrick *et al.*, "Online speech synthesis using a chronically implanted brain-computer interface in an individual with als," *medRxiv*, 2023.

[11] P. Clisson, R. Bertrand-Lalo, M. Congedo, G. Victor-Thomas, and J. Chatel-Goldman, "Timeflux: An open-source framework for the acquisition and near real-time processing of signal streams," in *BCI 2019-8th International Brain-Computer Interface Conference*, 2019.

[12] A. Stuart, J. Kalinowski, M. P. Rastatter, and K. Lynch, "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.

[13] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] J. Kohler *et al.*, "Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework," *Neurons, Behavior, Data analysis, and Theory*, 2022.

[15] M. Verwoert *et al.*, "Dataset of speech production in intracranial electroencephalography," *Scientific data*, vol. 9, no. 1, p. 434, 2022.

[16] Y.-E. Lee and S.-H. Lee, "Eeg-transformer: Self-attention from transformer architecture for decoding eeg of imagined speech," in *2022 10th International winter conference on brain-computer interface (BCI)*, IEEE, 2022, pp. 1–4.

[17] F. R. Willett, D. T. Avansino, L. R. Hochberg, J. M. Henderson, and K. V. Shenoy, "High-performance brain-to-text communication via handwriting," *Nature*, vol. 593, no. 7858, pp. 249–254, May 2021. (visited on 03/13/2024).

[18] K. Shigemi *et al.*, "Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[20] J. Yang, J. Lee, Y. Kim, H. Cho, and I. Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.

[21] N. S. Card *et al.*, "An accurate and rapidly calibrating speech neuroprosthesis," Neurology, Preprint, Dec. 2023. (visited on 12/27/2023).

Published by
Verlag der Technischen Universität Graz