2022 44th Annual International Conference of
the IEEE Engineering in Medicine & Biology Society (EMBC)
Scottish Event Campus, Glasgow, UK, July 11-15, 2022

2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 978-1-7281-2782-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/EMBC48229.2022.9871301

Towards Naturalistic Speech Decoding from Intracranial Brain Data*

Julia Berezutskaya[1], Luca Ambrogioni[1], Nick F. Ramsey[2] and Marcel A.J. van Gerven[1]

***Abstract*— Speech decoding from brain activity can enable development of brain-computer interfaces (BCIs) to restore naturalistic communication in paralyzed patients. Previous work has focused on development of decoding models from isolated speech data with a clean background and multiple repetitions of the material. In this study, we describe a novel approach to speech decoding that relies on a generative adversarial neural network (GAN) to reconstruct speech from brain data recorded during a naturalistic speech listening task (watching a movie). We compared the GAN-based approach, where reconstruction was done from the compressed latent representation of sound decoded from the brain, with several baseline models that reconstructed sound spectrogram directly. We show that the novel approach provides more accurate reconstructions compared to the baselines. These results underscore the potential of GAN models for speech decoding in naturalistic noisy environments and further advancing of BCIs for naturalistic communication.**

***Clinical relevance*— This study presents a novel speech decoding paradigm that combines advances in deep learning, speech synthesis and neural engineering, and has the potential to advance the field of BCI for severely paralyzed individuals.**

I. INTRODUCTION

The ability to speak and understand speech is remarkable and unique to the human species. Various accounts of how speech processing occurs in the brain exist, yet it remains difficult to relate continuous naturalistic speech to its underlying brain activity. Interest in speech decoding from the brain continues to grow due to the demand for brain-computer interface (BCI) technology to restore communication in severely paralyzed people. The BCI field is already making considerable progress, yet we believe it could benefit further from use of more advanced machine learning techniques to improve neural speech decoding.

In this paper we propose and validate a new speech decoding scheme based on generative adversarial neural networks (GANs). We used a publicly available dataset of spoken speech to train a GAN. Then, using an intracranial brain dataset we trained a decoder network to predict latent vectors, which were given as input to the GAN generator. The GAN generator was used to reconstruct speech spectrograms that were synthesized into speech using an external vocoder. We show that the proposed framework achieves best prediction accuracy, and the reconstructed speech is more natural-sounding than the tested baselines.

*This work was supported by the European Research Council (Advanced iConnect Project Grant ADV 320708) and the Netherlands Organisation for Scientific Research (Language in Interaction Project Gravitation Grant 024.001.006)

[1]Julia Berezutskaya, Luca Ambrogioni, Marcel A.J. van Gerven are with Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Thomas van Aquinostraat 4, 6525 GD Nijmegen, the Netherlands {yuliya.berezutskaya, luca.ambrogioni, marcel.vangerven}@donders.ru.nl

[2]Nicolas F. Ramsey is with Brain Center, Department of Neurology and Neurosurgery, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, the Netherlands {n.f.ramsey}@umcutrecht.nl

II. RELATED WORK

Speech decoding from the brain activity started with decoding of a few classes of isolated words and phonemes [1]–[3]. However, to make BCI communication more spontaneous and realistic, attempts to decode continuous speech signals have been made. Some studies used intracranial brain data to decode and concatenate individual phonemes for reconstruction of continuous speech [4], [5]. Others decoded vocoder parameters, such as pitch, envelope and aperiodicity to create intelligible speech reconstructions [6]. Another approach relied on brain decoding of speech spectrogram directly [7], [8]. Yet another used kinematic features as an intermediary between brain and speech spectrogram features [9]. Finally, many of these features have been combined to achieve high-accuracy reconstructions [10], [11].

The work done so far is remarkable, yet good performance is typically observed on data where participants spoke or listened to clean isolated sentences. It has been shown that many repetitions of the same speech content and averaging of the brain activity over fragments is necessary to achieve better decoding scores [11].

More robust brain decoding that can tackle continuous speech in noisy realistic environments is needed. In the field of vision, most recent work has shown promise of GANs in photo-realistic reconstruction of images from the brain activity [12]–[14]. Instead of decoding complex high-dimensional data such as natural images directly from the brain, pretrained GANs were used to generate this output from a compressed latent vector of only $100 - 300$ values. This latent vector can be decoded from the brain activity. We rely on this previous promising work from vision as inspiration for building a GAN-based speech decoder and evaluate it against other methods.

III. METHODS

A. Intracranial Dataset

Intracranial data (9 subjects in total) were collected using grids – electrocorticography (ECoG, 7 subjects) and depth electrodes – stereoelectroencephalography (sEEG, 2 subjects) in patients with medication-resistant epilepsy while they watched a full-length Dutch feature film (Minoes, 2001) [15]. The study was approved by the Medical Ethical

Committee of the University Medical Center Utrecht in accordance with the Declaration of Helsinki (2013). All patients gave written informed consent to participate. Most ECoG subjects had left hemisphere coverage with electrodes in perisylvian regions, i.e. temporal and frontal cortices (Figure 1). Most sEEG electrodes were implanted in the left hemisphere and sampled from Heschl's gyrus, hippocampus, insula and frontal regions. In total, 687 electrodes (449 ECoG and 238 sEEG) were analyzed. Data preprocessing was done per subject and included noisy channel rejection, notch filtering of line noise (50 Hz) and its harmonics, common average referencing and high-frequency component extraction (65–200 Hz) using Gabor wavelet decomposition in 1 Hz frequency bins. The high-frequency data were averaged over extracted frequencies and downsampled to 172 Hz. Based on the previous work [16], for training decoding models data were concatenated across all subjects.

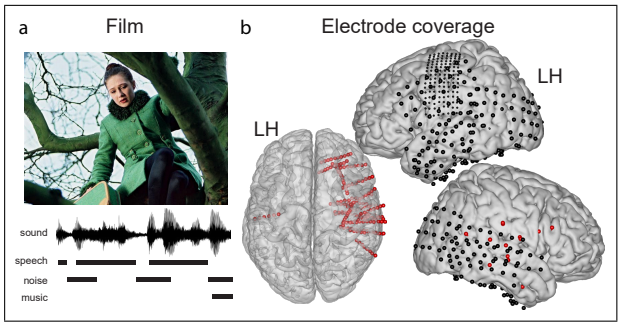

Fig. 1. a. Experimental setup. b. Total brain coverage with ECoG (in black) and sEEG (in red) electrodes.

B. Speech GAN Model

We used SpecGAN architecture from [17] and retrained it on an audio dataset that combined a spoken Dutch corpus IFA [18] (ca. 5.5 hours) and a subset of FSD50K [19] that contained broad sound categories (ca. 3 hours). The model was trained using original parameters, except that we increased the size of latent vectors from 100 to 300. From latent vectors the model generated .9-second log-mel spectrograms ($nfft = 1024$, $hop\_size = 256$) with 80 mel frequency bins. SpecGAN was trained for 3000 epochs on a single GPU (GeForce RTX 2080 Ti) using PyTorch [20].

C. Neural Decoders

All neural decoders were based on ResNet-18 [21] (Figure 2). The GAN-based neural decoder, termed GAN-Z, was trained on brain input to predict latent vectors that were then input to the generator of SpecGAN to produce a log-spectrogram for the target audio. During training a weighted combination of feature and pixel losses was minimized: $\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \sum_{l=1}^{L} \mathcal{L}_f^{(l)}$, where $\mathcal{L}_p$ and $\mathcal{L}_f$ are the mean squared error pixel and feature loss, respectively. After some experiments we set λ_p to 500 and λ_f to 1. $\mathcal{L}_p$ was computed on the generated and target spectrogram values. SpecGAN's discriminator was used as a feature extractor for both generated and target spectrograms and $l = \{1, 2, ..., L\}$ are its convolutional layers.

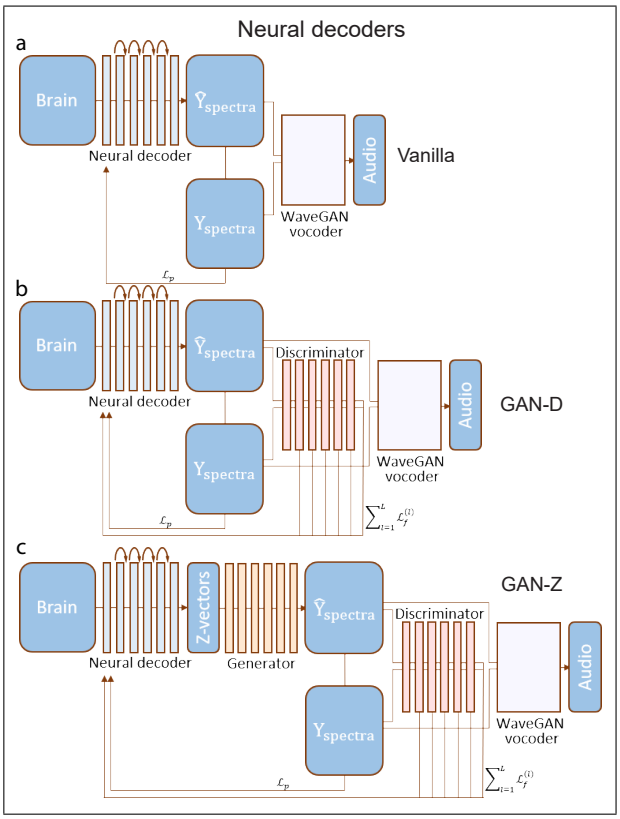

Fig. 2. Architectures of the three neural decoders. a. Vanilla decoder. b. GAN-D decoder. c. GAN-Z decoder.

We considered two baseline neural decoding models. The first model, termed Vanilla, was trained on input brain data to predict log-mel spectrograms of the corresponding audio directly by minimizing $\mathcal{L}_p$. Second model, GAN-D, was its modification, trained to minimized the combined pixel-feature loss $\mathcal{L}$. All models were trained using Adam optimizer ($\alpha = 5 \times 10^{-5}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and early stopping on 80% of data; remaining data were split between validation and test sets. All results are reported for test data. No cross-validation has been performed.

D. Speech synthesis

In order to synthesize sound from the reconstructed log-spectrograms we used a pretrained Parallel WaveGAN model (`https://github.com/kan-bayashi/ParallelWaveGAN`) [22].

E. Evaluation

Several metrics were used to evaluate the performance of the neural decoders: Pearson correlation, voice activity detection (VAD) match and pitch match between reconstructions and target audio fragments. Webrtcvad library (`https://github.com/wiseman/py-webrtcvad` was used to calculate VAD per each 30-second audio window. Parcelmouth [23] and Praat [24] were used to extract pitch. Per model, p-values for median values of each metric were calculated based on 1000 shuffles of reconstructed-target pairs. Pairwise model comparisons for each metric were made using Wilcoxon signed-rank tests.

F. Analysis of the GAN Latent Space

To explore the GAN latent space, a linear support vector machine (SVM) classifier with default parameters (as implemented in scikit-learn [25]) was used to fit a hyperplane separating the latent space into subspaces corresponding to speech and non-speech sounds [26]. For this, 2000 audio fragments were created using the GAN generator and passed through the VAD algorithm. All fragments with >80% speech frames were labeled 'voice', all fragments with <20% speech frames were labeled 'non-voice'.

IV. RESULTS

A. SpecGAN Generated High-quality Naturalistic Speech

SpecGAN performance has been previously validated [17], however, since we retrained the model on a new dataset, we repeated basic checks to evaluate the generated speech. Some examples are shown in Figure 3. After synthesis we observed that several generated fragments contained intelligible speech, and many sounded like speech but were not intelligible. For quantitative evaluation we used pretrained DeepSpeech2 [27] LSTM layer features to compute Fréchet Inception Distance (FID) between GAN-generated and real audio distributions: $FID_1 = 1.23$ for LSTM hidden states and $FID_2 = 30.22$ for LSTM cell states (25k samples). For reference, we computed the same scores between real audio fragments and real audio fragments, preprocessed and resynthesized with Parallel WaveGAN: $FID_1 = 1.4$ and $FID_2 = 57.23$, respectively, and the FID scores using SpecGAN-generated audio only after 10 epochs of training: $FID_1 = 10.28$ and $FID_2 = 264.19$, respectively.

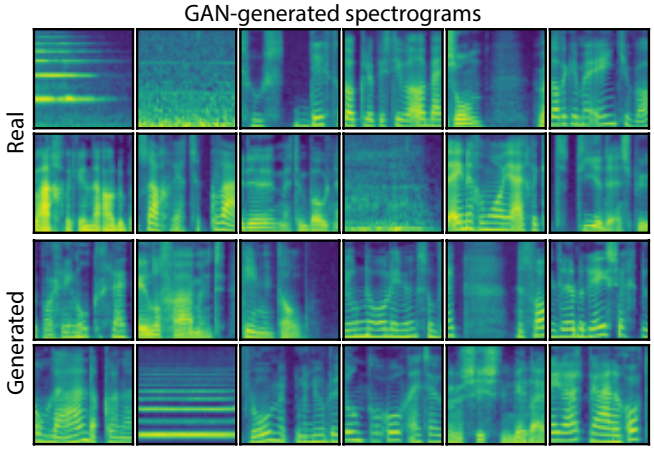

Fig. 3. Examples of real and GAN-generated spectrograms

B. Decoding GAN Latent Vectors Produced Best Speech Reconstructions

Next, we compared sound reconstructions obtained with Vanilla, GAN-D and GAN-Z neural decoders. We aimed to evaluate model predictions not only based on low-level but also higher-level sound properties, and paid particular attention to the perceptual quality of the reconstructed speech.

First, we compared sound reconstructions based on low-level sound features by computing correlations between predicted and target mel-frequency values. All models showed significantly high correlations compared to baseline surrogate distributions: $\overline{r}_{vanilla} = .64 \pm .07$, $\overline{r}_{GAND} = .62 \pm .07$, $\overline{r}_{GANZ} = .49 \pm .09$ (Figure 4). The Vanilla model showed significantly better correlations compared to other models: $Z_{Vanilla-GANZ} = 10.01$, $p = 6.84 \times 10^{-24}$ and $Z_{Vanilla-GAND} = 3.28$, $p = 5.18 \times 10^{-4}$ as assessed with Wilcoxon signed-rank tests. In general, decoders that were trained to predict mel-frequency values directly (Vanilla and GAN-D) achieved higher Pearson correlation between reconstructed and target spectrograms.

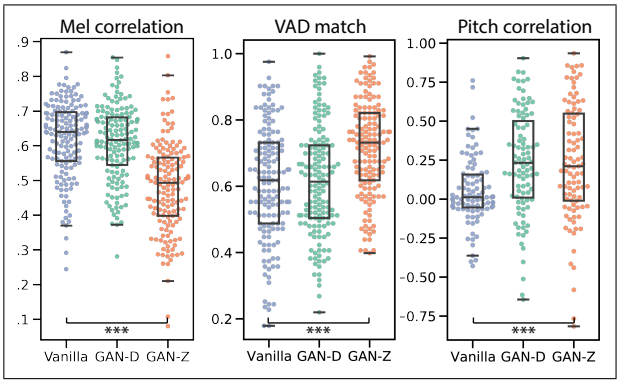

Fig. 4. Evaluation of neural decoder performance with log-mel feature correlation, VAD match and pitch match between reconstructed and target spectrograms. Boxes outline the 25th and 75th percentiles, caps show 5th and 95th percentiles, individual points show 150 test audio fragments. A solid line within each box is the median.

Next, we examined higher-level sound features: VAD and pitch, and evaluated the match between predictions and targets for both. Accurate detection of both features is directly related to correct onsets and offsets of speech, and correct reconstruction of periodic signal reflects the decoder's ability to generate realistic sounding speech. We found that contrary to the low-level features, the GAN-Z model was most accurate here, and the Vanilla model was significantly less accurate (Figure 4): $Z_{GANZ-Vanilla} = 5.49$, $p = 1.99 \times 10^{-8}$, as assessed with a Wilcoxon signed-rank test. Pitch match between reconstructions and target audio was strong for GAN-D and GAN-Z models (compared to the surrogate baseline). For Vanilla reconstructions pitch was hard to detect (Figure 5), which lead to zero values in many cases and significantly worse pitch match compared to GAN-D and GAN-Z: $Z_{GANZ-Vanilla} = 4.01$, $p = 3.1 \times 10^{-5}$ and $Z_{GAND-Vanilla} = 4.37$, $p = 6.2 \times 10^{-6}$ as assessed with Wilcoxon signed-rank tests.

Apart from the quantitative measures discussed above, it also became clear from visual inspection of the predicted and target spectrograms that the GAN-Z model better captured the high-level structure of speech, including formants (Figure 5). The difference between the models became even more apparent when we listened to the sound reconstructions. The first two models did not generate naturalistic speech but only unintelligible noise. Our main goal here was to

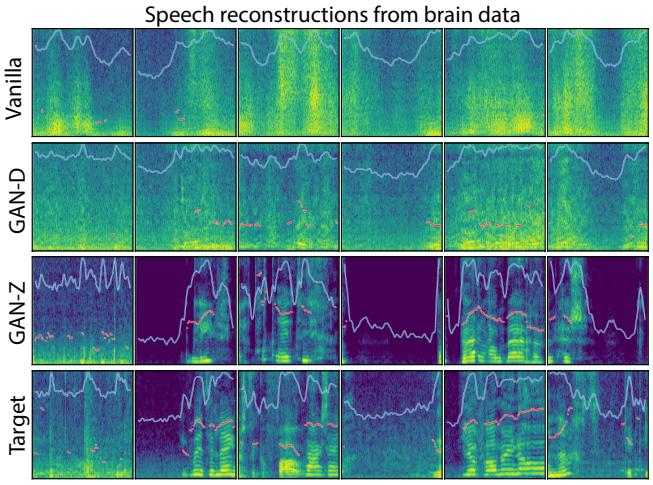

Fig. 5. Examples of sound reconstructions. Sound intensity is shown in white and pitch is shown in red.

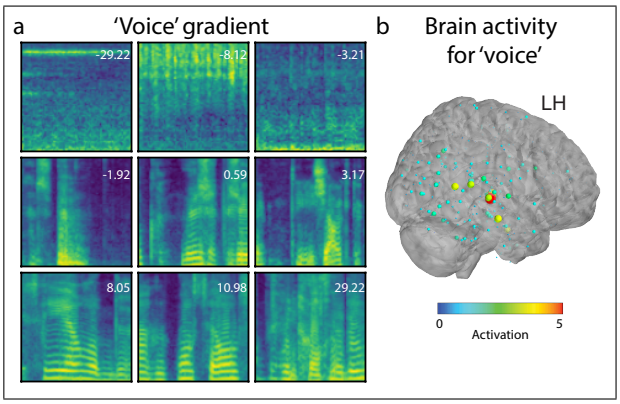

Fig. 6. a. 'Voice' gradient in the latent space, 'voice' score is reported per image on top right. b. Brain activity (pooled across all subjects) associated with high 'voice' scores, averaged over a batch of 128 examples.

reconstruct realistic, speech-like sound, and only the GAN-Z model achieved notable results. It successfully reconstructed formant structure of target sounds and its reconstructions sounded most like real speech. These results highlight the difference in sound quality between models trained to fit low-level features and models trained to approximate compressed latent vectors for non-trivial speech generation. This claim can be strengthened further by comparing judgements of reconstruction quality in human volunteers, which constitutes one of the main directions of our future work with these data.

C. Latent Space Encoded High-level Sound Structure that Could Be Decoded from the Brain

We saw that GAN-based model resulted in best, most naturalistic speech reconstructions. It is not trivial to understand what makes this model best. We made a first attempt to explore the latent space and relate high-level sound properties of that space to the brain input. Inspired by ideas in vision, we chose a binary semantic feature 'presence / absence of voice' ('voice') and looked for its encoding in the latent space [26]. We trained a linear SVM classifier to fit a hyperplane separating the latent space into subspaces corresponding to speech and non-speech sounds (test classification accuracy reached 97%). Similar to studies in vision, we visualized a gradient of sound spectrograms generated from the latent spaces by manipulating the strength of the 'voice' feature (Figure 6a).

Next, we used the input optimization technique [28] to obtain a brain activity map that corresponded to the voiced part of the gradient. The map showed highest activity values in Heschl's gyrus and surrounding superior temporal cortex typically involved in auditory processing (Figure 6b). Additionally, we fitted a linear regression on the real brain input in the test dataset to predict the strength of the 'voice' feature associated with each reconstructed sound. The map of the regression weights over electrodes showed a similar profile as above. Altogether, these results show that high-level features, such as 'voice', which can be seen as a proxy for speech, is encoded in the latent space. The 'voice' score can be accurately decoded from the brain auditory cortex, particularly from the Heschl's gyrus.

V. DISCUSSION

In this study we employed a generative adversarial neural network (GAN) to perform reconstruction of perceived speech from intracranial human data. We compared three neural decoding architectures: a non-GAN baseline, and two GAN-based decoders: one that only used the GAN discriminator and another that used the full GAN. Our evaluation metrics showed that GAN-based decoders may have been inferior in reconstruction of low-level audio features, but came out superior at reconstructing higher-level properties, such as speech timing, speaker pitch and speech formants. Moreover, GAN latent space exhibited interpretable audio features that mapped onto the brain activity.

The results of this work have several important implications. First, they indicate that use of high-level features (such as GAN discriminator layer activations) may lead to more natural-sounding speech reconstructions. This is in agreement with previous research on vision reporting that higher-level features of object recognition models are predictive of brain activity throughout top-down processing cortical regions [16], [29]. Moreover, studies on reconstruction of visual input show that adding high-level semantic information to decoders leads to better performance [30].

Second, this work illustrates the potential of generative models to boost neural decoding, particular in noisy naturalistic contexts. The big advantage of generative models is that they can be trained on vast speech corpora and learn complex features that define the space of all possible speech sounds. This information can be leveraged to constrain audio reconstruction from the brain, filter out irrelevant noise and provide powerful priors for more natural-sounding speech.

The present work has a number of limitations. First, we used data from passive watching of a feature film, and therefore validated our neural decoders on a speech perception task. We expect, however, that the presented methodology

is task-independent and will generalize to speech production data, which is directly relevant for BCI. Another limitation is lack of hyperparameter decoder optimization due to the small size of the dataset. Moreover, the present claims about superior perceptual quality of decoded speech require support from behavioral experiments. We are currently addressing these issues in our follow-up work on reconstruction of spoken speech from brain activity.

VI. CONCLUSIONS

The present study is among the first attempts to leverage advances in automatic sound generation with GANs for reconstructing naturalistic continuous speech from brain recordings. We showed that the GAN-based model achieved the best decoding accuracy in terms of recovering high-level sound properties and perceptual quality of sound. This was in contrast to models that were trained to decode speech spectrograms directly. These results demonstrate the potential of GAN-based models to advance the BCI field and make continuous speech decoding from the brain in naturalistic environments more plausible.

ACKNOWLEDGMENT

We thank Frans Leijten, Cyrille Ferrier, Geert-Jan Huiskamp, Sandra van der Salm and Tineke Gebbink for help with collecting data; Peter Gosselaar and Peter van Rijen for implanting the electrodes; the patients for their time and effort; and the members of the UMC Utrecht BCI team for data collection.

REFERENCES

[1] T. Blakely, K. J. Miller, R. P. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ecog) grids," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 4964–4967, IEEE, 2008.

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

[3] N. F. Ramsey, E. Salari, E. J. Aarnoutse, M. J. Vansteensel, M. G. Bleichner, and Z. Freudenburg, "Decoding spoken phonemes from sensorimotor cortex with high-density ecog grids," *NeuroImage*, vol. 180, pp. 301–311, 2018.

[4] 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. 9, p. 217, 2015.

[5] C. Herff, L. Diener, M. Angrick, E. Mugler, M. C. Tate, M. A. Goldrick, D. J. Krusienski, M. W. Slutzky, and T. Schultz, "Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices," *Frontiers in neuroscience*, vol. 13, p. 1267, 2019.

[6] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, "Towards reconstructing intelligible speech from the human auditory cortex," *Scientific reports*, vol. 9, no. 1, pp. 1–12, 2019.

[7] M. Angrick, C. Herff, E. Mugler, M. C. Tate, M. W. Slutzky, D. J. Krusienski, and T. Schultz, "Speech synthesis from ecog using densely connected 3d convolutional neural networks," *Journal of neural engineering*, vol. 16, no. 3, p. 036019, 2019.

[8] M. Angrick, M. 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," *bioRxiv*, 2020.

[9] 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.

[10] 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.

[11] P. Sun, G. K. Anumanchipalli, and E. F. Chang, "Brain2char: a deep architecture for decoding text from brain recordings," *Journal of Neural Engineering*, vol. 17, no. 6, p. 066015, 2020.

[12] K. Seeliger, U. Güçlü, L. Ambrogioni, Y. Güçlütürk, and M. A. van Gerven, "Generative adversarial networks for reconstructing natural images from brain activity," *NeuroImage*, vol. 181, pp. 775–785, 2018.

[13] T. Dado, Y. Güçlütürk, L. Ambrogioni, G. Ras, S. Bosch, M. van Gerven, and U. Güçlü, "Hyperrealistic neural decoding for reconstructing faces from fmri activations via the gan latent space," *Scientific reports*, vol. 12, no. 1, pp. 1–9, 2022.

[14] L. Le, L. Ambrogioni, K. Seeliger, Y. Güçlütürk, M. van Gerven, and U. Güçlü, "Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity," *bioRxiv*, 2021.

[15] J. Berezutskaya, Z. V. Freudenburg, U. Güçlü, M. A. van Gerven, and N. F. Ramsey, "Brain-optimized extraction of complex sound features that drive continuous auditory perception," *PLoS computational biology*, vol. 16, no. 7, p. e1007992, 2020.

[16] J. Berezutskaya, Z. V. Freudenburg, L. Ambrogioni, U. Güçlü, M. A. van Gerven, and N. F. Ramsey, "Cortical network responses map onto data-driven features that capture visual semantics of movie fragments," *Scientific reports*, vol. 10, no. 1, pp. 1–21, 2020.

[17] C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," *arXiv preprint arXiv:1802.04208*, 2018.

[18] R. van Son, D. Binnenpoorte, H. v. d. Heuvel, and L. Pols, "The ifa corpus: a phonemically segmented dutch" open source" speech database," 2001.

[19] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, 2020.

[20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

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

[22] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6199–6203, IEEE, 2020.

[23] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[24] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." Version 6.1.38, retrieved 2 January 2021 `http://www.praat.org/`, 2021.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[26] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of gans for semantic face editing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020.

[27] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, pp. 173–182, PMLR, 2016.

[28] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[29] U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *Journal of Neuroscience*, vol. 35, no. 27, pp. 10005–10014, 2015.

[30] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant, "Bayesian reconstruction of natural images from human brain activity," *Neuron*, vol. 63, no. 6, pp. 902–915, 2009.