



EEG-based Voice Conversion : Hearing the Voice of Your Brain

Yizhong Geng, Wenxin Fu, Qihang Lu, Bingsong Bai, Cong Wang, Yingming Gao, Ya Li*

Beijing University of Posts and Telecommunications, China

yzgeng@bupt.edu.cn, fuwenxin2003@bupt.edu.cn, lqh2021213559@bupt.edu.cn,
bingsongbai@bupt.edu.cn, congwang@bupt.edu.cn, yingming.gao@bupt.edu.cn,
yli01@bupt.edu.cn

Abstract

The connection between Electroencephalography (EEG) signals and human voice has gained significant attention, with studies demonstrating the feasibility of speech synthesis from EEG data. However, EEG-based voice conversion (VC) remains largely unexplored. To address this, we present the first EEG-based zero-shot VC system that converts speech into a target speaker's voice without prior target data. Our method integrates an EEG feature extraction module with an alignment module to map EEG features to speaker-specific voice features. By leveraging an innovative three-stage training strategy and a pre-trained VC model—trained solely on speech data—we achieve zero-shot conversion. Experiments on the Single-Word-Production Dutch-iBIDS dataset confirm the system's ability to reliably convert speech to a target speaker's voice. This work highlights the potential of EEG-based VC for advancing assistive communication and brain-computer interfaces. All demos are available in <https://doi.org/10.5281/zenodo.15510829>.

Index Terms: EEG, zero-shot voice conversion, cross modal generation, speaker embedding

1. Introduction

Brain-computer interfaces (BCIs) have seen significant advancements in recent years, with their potential to revolutionize various fields, including communication, neuroprosthetics, and assistive technologies [1, 2]. One of the most promising applications of BCIs is their ability to assist individuals with disabilities, enabling them to regain lost functions and interact with the world more effectively [3, 4]. Inspired by these applications, we propose a novel EEG-based voice conversion system. To the best of our knowledge, this study is the first to successfully introduce EEG-based voice conversion, offering the potential to restore a key element of human expression—one's voice. By leveraging EEG signals to drive voice conversion, this technology can help individuals without speech abilities recover a personalized version of their original voice.

Voice, as the foundation of human expression, plays a vital role in interpersonal communication, serving not only as a means to exchange information but also as a key aspect of personal identity. The unique qualities of a person's voice are a defining aspect of their identity. Losing the ability to speak strips away a distinctive personal trait, leaving individuals without a vital element of their individuality and self-expression [5].

Voice conversion (VC) modifies a source speaker's voice to match the target speaker's characteristics while preserving linguistic content [6]. Recent zero-shot VC advances, like FreeVC

[7], enable direct extraction of speaker features from speech, offering a more flexible and scalable approach. This innovation has sparked the use of other modalities, such as face [8, 9] and EEG signals, to extract and align features for voice conversion.

EEG signals are inherently unique, with each individual exhibiting distinct patterns that can serve as reliable identifiers for personal differentiation and authentication [10]. Recent research has delved into the relationship between EEG signals and vocal, exploring how neural activity can be leveraged to recognize the heard speaker [11, 12] or reconstruct speech [13, 14]. However, most of these studies focus on extracting semantic content or basic sound features from EEG data, neglecting the timbral qualities that define the individual's voice [15].

Building on recent research linking brain activity to vocal identity [11, 12], we present a novel EEG-based zero-shot voice conversion framework. Our approach leverages solely the target speaker's EEG features, thereby eliminating the dependence on conventional vocal data and underscoring the intrinsic connection between neural patterns and voice timbre. To address the challenges of cross-modal feature mapping and data scarcity, we propose an innovative three-stage training strategy [16]. This comprehensive training process ensures that even in the absence of target speaker data, accurate voice conversion can be achieved. Our framework not only minimizes data requirements but also paves the way for advancements in personalized voice synthesis, assistive communication technologies, and brain-to-speech systems.

To contextualize our research, we delineate our contributions as follows:

- We propose a Zero-Shot EEG-based Voice Conversion framework, enables speaker-specific voice conversion through EEG signals.
- We introduce a phased training strategy, leverages pre-trained voice conversion models to achieve effective zero-shot performance even with limited data.
- We demonstrate the connection between EEG features and voice features, validating the relationship between EEG signals and speaker-specific vocal attributes.

2. Related work

2.1. EEG-voice association

EEG signals are unique to individuals, exhibiting distinct patterns that make them valuable for tasks such as personal identification [10]. Beyond this, EEG signals are closely associated with acoustic features, as demonstrated by studies that reconstruct speech from neural activity, including imagined and overt speech [15]. Researchers have also utilized EEG to aid speaker recognition by combining neural and acoustic infor-

* Corresponding author

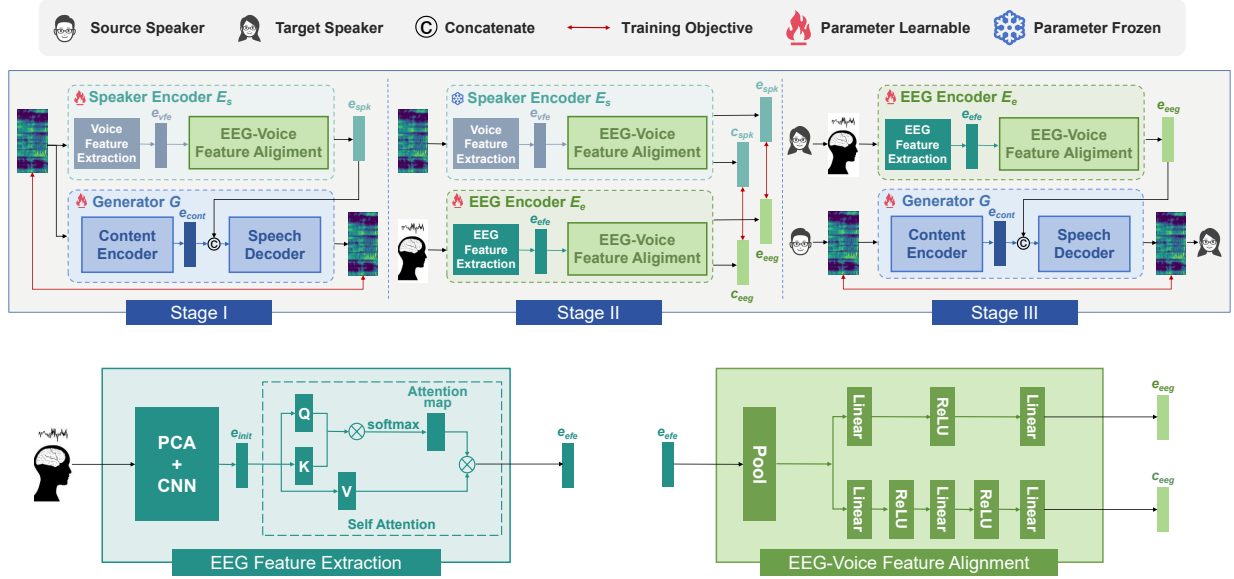


Figure 1: Overview of the proposed system architecture and training strategy. The upper section outlines the three-stage training process: I) Generator and speaker encoder pre-training with speech data; II) EEG-speaker embedding alignment; III) Generator fine-tuning for EEG-based synthesis (training reconstructs the source voice; inference converts to the target voice via EEG). The bottom section highlights the EEG encoder, which includes a feature extraction module and an EEG-voice feature alignment module.

mation [17]. Additionally, cross-modal approaches like face-driven voice conversion show that non-speech modalities can guide vocal synthesis [8, 9]. Inspired by these works, we propose a novel framework that leverages an individual’s EEG signals to drive voice conversion, linking neural activity with vocal identity and enabling EEG-based voice transformation.

2.2. Voice conversion

Voice conversion (VC) aims to modify a source speaker’s voice to match the target speaker’s characteristics while preserving linguistic content [6]. Traditional VC methods, such as GMMs [18], required parallel datasets, limiting scalability [19]. Deep learning models like StarGAN-VC [20] and CycleGAN-VC [21] removed this need by enabling non-parallel training, improving flexibility. More recently, self-supervised learning (SSL) techniques, extract robust features that allow for zero-shot conversion of unseen speakers, further advancing VC [22, 23, 24, 25]. Following this trend, non-speech modalities like facial features have been explored for guiding VC [8, 9]. Inspired by these developments, we propose using EEG signals for voice conversion, linking neural activity directly to vocal identity and enabling EEG-based transformations.

3. Method

In this work, we propose an EEG-based zero-shot voice conversion (VC) system that transforms speech based on the target speaker’s EEG signals while preserving the source audio’s semantic information. The system comprises: 1) an EEG Encoder for EEG Feature Extraction and EEG-voice Feature Alignment; 2) a Speaker Encoder for speaker-specific embeddings [26]; and 3) a Generator, built on a modified FreeVC framework [7], with a Content Encoder for semantic extraction and a Speech De-

coder for final audio generation. The detailed architecture and training strategy are illustrated in Figure 1.

3.1. EEG encoder

The EEG encoder is designed to extract robust, discriminative features from noisy, high-dimensional, and time-series EEG signals. To achieve zero-shot voice conversion, the extracted features must not only uniquely represent each speaker but also encode timbre-related information. Therefore, the EEG encoder consists of two components: EEG feature extraction and EEG-voice feature alignment.

3.1.1. EEG feature extraction

The EEG feature extraction module processes high-dimensional EEG data, $X \in \mathbb{R}^{T \times N}$, with T time steps and N electrodes, into compact representations for voice conversion. PCA reduces the dimensionality, preserving the most significant variance:

$$X_{\text{PCA}} = \text{PCA}(X) \in \mathbb{R}^{T \times M} \quad (1)$$

where M is the reduced number of principal components, minimizing redundancy and complexity.

Building on EEG Conformer [27], we employ a hybrid model combining CNNs and Transformer-based self-attention. A shallow CNN processes X_{PCA} to capture local spatiotemporal patterns, producing an intermediate representation $e_{\text{init}} = \text{CNN}(X_{\text{PCA}})$. Then, the Transformer encoder models global dependencies via multi-head self-attention, computing queries Q , keys K , and values V :

$$Q = e_{\text{init}}W_Q, \quad K = e_{\text{init}}W_K, \quad V = e_{\text{init}}W_V \quad (2)$$

Table 1: Evaluation results of different voice conversion methods across various metrics. The definitions of all metrics are provided in Section 4.3.

Method	Homogeneity (\uparrow)	Consistency (obj) (\uparrow)	Consistency (sub) (\uparrow)	Naturalness (\uparrow)
Ground Truth (GT)	0.9001	-	-	-
FreeVC(unseen speakers)	0.9371	0.8216	3.59	3.95
EEG-based VC (seen speakers)	0.9465	0.8391	3.68	4.15
EEG-based VC (unseen speakers)	0.9437	0.8026	3.45	4.00

The attention mechanism computes:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (3)$$

where d_k is the dimension of keys, and $\sqrt{d_k}$ stabilizes gradients. The multi-head attention mechanism enhances pattern capture. The final feature representation, $e_{efe} = \text{Transformer}(e_{\text{init}})$, integrates both local and global EEG patterns, encoding speaker-specific and timbre-related information.

3.1.2. EEG-voice feature alignment

The EEG-voice feature alignment module bridges neural representations with speaker-specific voice characteristics by mapping EEG features to a shared embedding space. Given the EEG feature representation $e_{efe} \in \mathbb{R}^d$ (extracted from 3.1.1) and the speaker embedding $e_{spk} \in \mathbb{R}^s$ (obtained from the Speaker Encoder), this module ensures alignment between EEG-derived features and the target speaker’s vocal identity while predicting speaker categories for auxiliary supervision. The architecture comprises two parallel branches:

Feature projection: This branch compresses the EEG features temporally and projects them into the speaker embedding space using adaptive pooling followed by a two-layer MLP. The final projection is computed as:

$$e_{\text{hidden}} = \text{ReLU}(W_1 \cdot \text{AdaptiveAvgPool1D}(e_{vfe}) + b_1) \quad (4)$$

$$e_{eeg} = W_2 \cdot e_{\text{hidden}} + b_2 \quad (5)$$

Speaker classification: This branch predicts the speaker category by first compressing EEG features temporally and then passing them through a multi-layer network. The process is summarized as follows:

Temporal Pooling and Initial Projection:The EEG features are compressed via adaptive average pooling and projected into a higher-dimensional space:

$$h_1 = \text{ReLU}(W_1 \cdot \text{AdaptiveAvgPool1D}(e_{vfe}) + b_1) \quad (6)$$

Regularization and Final Classification:The features undergo dropout regularization, followed by further projection and softmax normalization:

$$h_2 = \text{ReLU}(W_2 \cdot \text{Dropout}(h_1) + b_2) \quad (7)$$

$$c_{eeg} = \text{Softmax}(W_3 h_2 + b_3) \quad (8)$$

3.2. Training strategy

To enhance the system’s performance, we employ a three-stage training strategy. The training process is illustrated in Figure 1. This strategy leverages a pre-trained voice conversion model that has been trained exclusively on speech data, allowing our EEG-based VC system to benefit from the rich representations

learned from large-scale speech datasets. By integrating these components, our method enables the conversion of speech to a target speaker’s voice without requiring prior exposure to their EEG or voice data.

Stage I: Speech-driven voice conversion pre-training

In this stage, we pre-train the generator G and the speaker encoder E_s on speech data to synthesize speech by mapping semantic content to speaker-specific features extracted from the target speech data.

$$\mathcal{L}_{\text{Stage I}} = \|G(e_{\text{cont}}, e_{\text{spk}}) - x_{\text{target}}\|_2^2 \quad (9)$$

Stage II: EEG-voice feature alignment

In this stage, we train only the EEG encoder E_e to align EEG features with pre-trained speaker features e_{spk} using an embedding fitting loss and a classification loss between the predicted speaker class c_{eeg} and the true speaker class c_{spk} .

$$\mathcal{L}_{\text{Stage II}} = \lambda_1 \cdot \|E_e(x_{\text{EEG}}) - e_{\text{spk}}\|_2^2 - \lambda_2 \cdot \sum_i c_{\text{spk},i} \log(c_{\text{eeg},i}) \quad (10)$$

Stage III: EEG-based voice conversion

In the final stage, we fine-tune both the generator G and the EEG encoder E_e to directly generate speech from EEG-derived features.

$$\mathcal{L}_{\text{Stage III}} = \|G(e_{\text{cont}}, E_e(x_{\text{EEG}})) - x_{\text{target}}\|_2^2 \quad (11)$$

4. Experiment

4.1. Dataset

We conducted experiments using two datasets: the Single-Word-Production Dutch-iBIDS [28] and VCTK [29]. The VCTK dataset contains pure speech recordings, which are used in Stage I for pre-training the model in voice conversion without EEG inputs. In this stage, we train the generator and speaker encoder to capture voice characteristics from the speech data. The Single-Word-Production Dutch-iBIDS dataset provides paired EEG and voice recordings from 10 speakers, capturing neural activity during speech production. This dataset is used in Stage II and Stage III, where the model learns to align EEG signals with voice features. Both EEG and audio data are segmented into 10-second intervals to maintain consistency in data length and ensure comparability between the two modalities.

For the EEG feature extraction in Stage II, we apply PCA to reduce the dimensionality of the EEG signals. Specifically, the raw EEG data is reduced to 25 principal components, using PCA with $M = 25$, which minimizes redundancy while retaining the key features necessary for voice conversion. The model training in Stage II is guided by two loss terms, λ_1 and λ_2 , which balance the objectives of feature alignment and model regularization. We set $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$, ensuring that

the model focuses primarily on aligning the EEG and voice features while maintaining regularization to prevent overfitting.

The evaluation process involves several steps. For seen speakers, 80% of the EEG data is used for training and 20% for testing, resulting in about 1500 generated samples. For unseen speakers, we use k-fold cross-validation, training the model on EEG data from 9 speakers and testing on data from 1 unseen speaker, repeated for each speaker, yielding approximately 1500 samples. We then compare our method with the baseline Free-VC model, segmenting each audio sample into 10-second intervals for conversion, generating around 1500 results for comparison.

4.2. Comparison systems

Since our work is the first to explore EEG-driven voice conversion (VC), there are no directly comparable systems, so we evaluate our method against conventional speech-driven VC.

Ground Truth (GT): The original target speech recordings, serving for the evaluation.

FreeVC: A zero-shot voice conversion model that operates purely on speech data, selected as a baseline since our EEG-based VC is built upon its framework.

EEG-based VC (seen speakers): Our proposed system that leverages EEG signals for voice conversion on seen speakers, using EEG data from speakers the model has been trained on.

EEG-based VC (unseen speakers): Our proposed system that leverages EEG signals to achieve zero-shot voice conversion without requiring speech data from the target speaker.

4.3. Metrics

We adopt the evaluation methodology from Lee et al. [9], using several metrics to assess our EEG-based voice conversion (VC) system, including Homogeneity, Consistency (obj), Consistency (sub), and Naturalness. These metrics provide a comprehensive evaluation of both objective and subjective voice conversion quality. For subjective evaluation, 20 testers assessed 16 audio sample groups.

Homogeneity: Measures cosine similarity of speaker embeddings in synthesized audio from different EEG signals of the same speaker.

Consistency (obj): Compares cosine similarity between synthesized and ground-truth audio from the same speaker, assessing the consistency of the synthesized voice with the original.

Consistency (sub): Assesses subjective consistency between synthesized audio and the corresponding EEG signals, using a 5-point MOS scale from completely inconsistent to consistent.

Naturalness: Evaluates the perceptual quality of synthesized audio on a 5-point MOS scale, ranging from unnatural to natural, indicating how natural the speech sounds to listeners.

4.4. Results

We evaluated our EEG-based voice conversion (VC) system following the procedures in Section 4.1 and using the metrics described in Section 4.3. The evaluation results are summarized in Table 1.

4.4.1. Objective results

Our EEG-based VC system for seen speakers (EEG-based VC (seen speakers)) achieved the highest score for Homogeneity (0.9465), indicating strong preservation of speaker identity

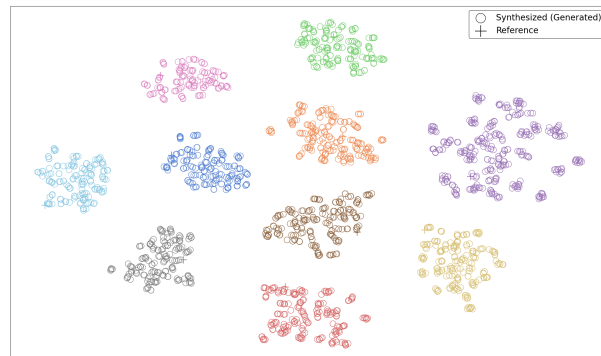


Figure 2: *t*-SNE visualization of the generated speech. The circles represent the synthesized speech, and the plus signs represent the reference audio.

across different EEG signals, followed closely by FreeVC (0.9371). For Consistency (obj), EEG-based VC (seen speakers) scored 0.8391, outperforming FreeVC (0.8216), while EEG-based VC (unseen speakers) for unseen speakers achieved a consistency score of 0.8026. These results highlight the effectiveness of our EEG-based approach, particularly for zero-shot voice conversion, where the model performs well even without prior exposure to the target speaker’s voice. Additionally, as shown in Figure 2, a *t*-SNE visualization confirms the system’s ability to maintain speaker identity, showing that the generated speech (circles) clusters closely with the reference audio (plus signs). This demonstrates that EEG-based VC successfully captures and preserves speaker characteristics, supporting the feasibility of using EEG for voice conversion.

4.4.2. Subjective results

In subjective evaluations, our EEG-based voice conversion system achieved strong perceptual performance. For seen speakers, it attained a Consistency (sub) MOS of 3.68 and a Naturalness MOS of 4.15, both slightly outperforming FreeVC. For unseen speakers, the scores were marginally lower (3.45 for consistency and 4.00 for naturalness) yet still indicate high quality. These results clearly demonstrate that our approach effectively preserves speaker characteristics from EEG signals and produces natural-sounding speech, highlighting its potential for real-world applications.

5. Conclusion

In this work, we introduce the first EEG-based voice conversion (VC) system, linking neural signals to voice characteristics. Our system preserves speaker identity and generates natural-sounding speech for both seen and unseen speakers, demonstrating EEG-based voice conversion feasibility. Unlike traditional EEG synthesis methods, which focus on reconstructing speech content but overlook voice timbre, EEG-based VC enables personalized voice restoration. This capability holds particular significance for individuals with speech impairments, offering the potential to recover their natural vocal identity. While current results are limited by data, we believe more data will improve performance and broaden applications. This work lays the foundation for advancements in personalized communication, brain-to-speech interfaces, and neuroprosthetics, enabling assistive technologies and customized speech synthesis.

6. Acknowledgements

The work was supported by the National Natural Science Foundation of China (NSFC) (No.62271083), the Key Project of the National Language Commission (No. ZDI145-81), the Fundamental Research Funds for the Central Universities (No.2023RC73), and partly supported by the Major Program of the National Social Science Fund of China (13&ZD189).

7. References

- [1] N. Rahman, D. M. Khan, K. Masroor, M. Arshad, A. Rafiq, and S. M. Fahim, "Advances in brain-computer interface for decoding speech imagery from eeg signals: a systematic review," *Cognitive Neurodynamics*, pp. 1–19, 2024.
- [2] A. Pirasteh, M. Shamseini Ghiyasvand, and M. Pouladian, "Eeg-based brain-computer interface methods with the aim of rehabilitating advanced stage als patients," *Disability and Rehabilitation: Assistive Technology*, pp. 1–11, 2024.
- [3] S. Ghasemi, D. Gračanin, and M. Azab, "Empowering mobility: Brain-computer interface for enhancing wheelchair control for individuals with physical disabilities," in *International Conference on Human-Computer Interaction*. Springer, 2024, pp. 234–245.
- [4] Y. An, D. Mitchell, J. Lathrop, D. Flynn, and S.-J. Chung, "Motor imagery teleoperation of a mobile robot using a low-cost brain-computer interface for multi-day validation," in *2024 IEEE Conference on Telepresence*. IEEE, 2024, pp. 103–110.
- [5] J. Stern, C. Schild, B. C. Jones, L. M. DeBruine, A. Hahn, D. A. Puts, I. Zettler, T. L. Kordsmeyer, D. Feinberg, D. Zamfir *et al.*, "Do voices carry valid information about a speaker's personality?" *Journal of Research in Personality*, vol. 92, p. 104092, 2021.
- [6] T. Walczyna and Z. Piotrowski, "Overview of voice conversion methods based on deep learning," *Applied sciences*, vol. 13, no. 5, p. 3100, 2023.
- [7] J. Li, W. Tu, and L. Xiao, "Freevc: Towards high-quality text-free one-shot voice conversion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [8] H.-H. Lu, S.-E. Weng, Y.-F. Yen, H.-H. Shuai, and W.-H. Cheng, "Face-based voice conversion: Learning the voice behind a face," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 496–505.
- [9] J. Lee, Y. Oh, I. Hwang, and K. Lee, "Hear your face: Face-based voice conversion with f0 estimation," *arXiv preprint arXiv:2408.09802*, 2024.
- [10] W. Alsumari, M. Hussain, L. Alshehri, and H. A. Aboalsamh, "Eeg-based person identification and authentication using deep convolutional neural network," *Axioms*, vol. 12, no. 1, p. 74, 2023.
- [11] H. Zhu, S. Cai, Y. Jiang, Q. Zhang, and H. Li, "Eeg-derived voice signature for attended speaker detection," *arXiv preprint arXiv:2308.14774*, 2023.
- [12] L. Hu, L. Zhu, H. Huang, G. Lin, B. Ren, and J. Zhang, "Speaker recognition with voice evoked eeg," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 2231–2236.
- [13] X. Xu and C. Fu, "Robust imagined speech production using ai-generated content network for patients with language impairments," *IEEE Transactions on Consumer Electronics*, 2024.
- [14] D. Qi, L. Kong, L. Yang, and C. Li, "Audiodiffusion: Generating high-quality audios from eeg signals: Reconstructing audio from eeg signals," in *2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC)*. IEEE, 2023, pp. 344–348.
- [15] S.-H. Lee, Y.-E. Lee, and S.-W. Lee, "Voice of your brain: Cognitive representations of imagined speech, overt speech, and speech perception based on eeg," *arXiv preprint arXiv:2105.14787*, 2021.
- [16] B. Liu, L. Hu, Q. Dong, and Z. Hu, "An iterative co-training transductive framework for zero shot learning," *IEEE Transactions on Image Processing*, vol. 30, pp. 6943–6956, 2021.
- [17] G. Krishna, C. Tran, M. Carnahan, and A. Tewfik, "Speaker identification using eeg," *arXiv preprint arXiv:2003.04733*, 2020.
- [18] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on gmm with straight mixed excitation," 2006.
- [19] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [20] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 266–273.
- [21] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.
- [22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.
- [23] K. Qian, Y. Zhang, H. Gao, J. Ni, C.-I. Lai, D. Cox, M. Hasegawa-Johnson, and S. Chang, "Contentvec: An improved self-supervised speech representation by disentangling speakers," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18003–18017.
- [24] Z. Li, B. Tang, X. Yin, Y. Wan, L. Xu, C. Shen, and Z. Ma, "Ppg-based singing voice conversion with adversarial representation learning," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7073–7077.
- [25] H. Chung, H. Nam *et al.*, "Zero-shot voice conversion with hubert," *Phonetics and Speech Sciences*, vol. 15, no. 3, pp. 69–74, 2023.
- [26] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.
- [27] Y. Song, Q. Zheng, B. Liu, and X. Gao, "Eeg conformer: Convolutional transformer for eeg decoding and visualization," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 710–719, 2022.
- [28] 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*, vol. 9, no. 1, p. 434, 2022.
- [29] C. Veaux, J. Yamagishi, and K. MacDonald, "Dataset: Vctk corpus," <https://doi.org/10.57702/bw3hjwag>, 2024, DOI retrieved: December 2, 2024.