CerebroVoice: A Stereotactic EEG Dataset and Benchmark for Bilingual Brain-to-Speech Synthesis and Activity Detection Supplementary Material

This supplement to our main paper, "CerebroVoice: A Stereotactic EEG Dataset and Benchmark for Bilingual Brain-to-Speech Synthesis and Activity Detection," provides an in-depth explanation of the dataset collection methods and includes a comprehensive data card. It also outlines the licensing information for the dataset and includes an author statement verifying compliance with these licensing terms. Furthermore, it addresses the societal implications, providing a Preliminary Assessment and Disposal Plan of Relevant Risks as well as discussing Ethical Issues and Countermeasures. Detailed descriptions of the methods implemented on the dataset, along with the datasheets, are also included.

1 Data Collection

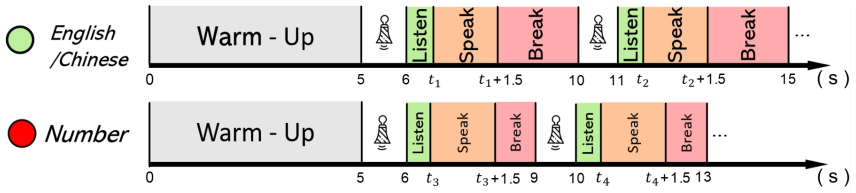

Figure 1: The timeline of experiment of each round

In our study, subjects were exposed to auditory stimuli from three different classifications: 30 categoriess of Chinese Mandarin words, 10 categoriess of Chinese Mandarin digits, and 10 categories of English words. The listening and repetition phase for both Chinese Mandarin and English words was allocated 5 seconds, whereas for Chinese Mandarin digits, this phase lasted 4 seconds. Participants underwent 8 rounds of the experiment, each round comprising 30 English words, 60 Chinese Mandarin digits, and 110 Chinese Mandarin words. At the start of each round, subjects had a 5-second preparation period, during which they were instructed through an audio prompt, "Please listen to the audio attentively and repeat loudly what you will hear," followed by a "ding" sound indicating the commencement of the speech content to be attended to. Following the playback of each word, subjects were required to repeat the speech content within 1.5 seconds and then stay relaxed until the next "ding" was heard. The data collection timeline for each round is depicted in Figure. 1.

1.1 Preliminary Assessment and Disposal Plan of Relevant Risks

To ensure the scientific property of the trial and the safety of the participants, we conducted a comprehensive assessment of the trial participants. Eligible trial participants were required to sign an informed consent form to understand the purpose, process, possible adverse reactions of the trial in detail, and clarify the relevant safety measures.

During the experiment, doctors and research teams worked together to ensure the safety and comfort of patients. If the patient felt tired during the trial, we would suspend the trial at any time to provide

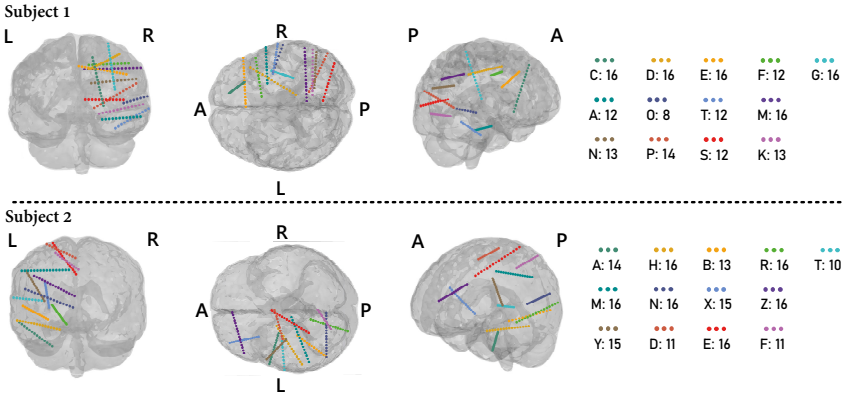

Figure 2: sEEG electrode contact locations for each subject. Dots of the same color represent electrode contacts positioned on the same electrode shafts. These locations are determined by co-registering pre-implantation magnetic resonance imaging (MRI) scans with post-implantation computed tomography (CT) scans.

rest. In addition, we closely monitored any potential risks during the trial and be ready to respond to emergencies at any time to maximize the safety and legal rights of the subjects.

1.2 Ethical Issues and Countermeasures

(1) Individuals participated in the study on a voluntary basis, and after ensuring that the subjects understand the relevant information, written informed consent were obtained from the subjects.

(2) All measures have been taken to protect the privacy of the subjects and keep personal information confidential.

(3) Each subject received sufficient information, including the purpose and methods of the study, any possible conflicts of interest, the researcher's organizational affiliation and potential risks, any discomfort that the study may cause, and any other information related to the study.

(4) Each subject was informed of his or her right to refuse to participate in the study and the right to withdraw consent to withdraw from the study at any time.

2 Dataset Structure

Our dataset collected 3200 samples from 3 volunteers, and then reserved 3069 samples, including 1493 samples from the first participant and 1576 samples from the second p articipant. Our data includes 27 folders. The outermost three folders are classified into BBS, HGA, and LFS to represent different frequency bands. The middle three folders are classified into Chinese Mandarin, English, and digits according to the type of words. It is essential to note that within each frequency band, we extracted samples from the initial pool of 3069, giving us a total of 9207 distinct samples across the full spectrum of frequency bands. This additional extraction process has allowed us to delve deeper into the data and create a comprehensive and detailed dataset.

As illustrated in Figure. 3, the innermost three folders are training set, validation set, and test set. In order to facilitate data users to view the basic information of each sample, we use a unified format to name the files of the training set, validation set, and test set, namely roundID_wordID_wordName, where round ID represents the round of experiments, word id represents the number of words read by the participant in this round of experiments, and word name represents the content of the words read by the participant. For ease of use, we provide the preprocessed sEEG signal and mel-spectrogram, both stored in npy format. It contains the following data:

(1) sEEG: a data matrix representing sEEG signals, ending with SEEG.npy, in the shape of T * F, where T represents the time dimension and F is the number of features. For HGA and LFS, the number of features is the same as the number of sEEG channels, and for BBS, the number of features is twice the number of channels. The number of valid channels for the first participant is 114, and the number of valid channels for the second participant is 158.

(2) Mel-Spectrogram: a data matrix representing the mel-spectroogram of audio signals, ending with MEL.npy, in the shape T*80, where T represents the time dimension and 80 represents the number of bin of the mel-spectrogram.

Additional dataset statistics are listed in Table 1. Note that the Total Number of Samples refers to the combined samples across all frequency bands (BBS, HGA, and LFS), while the Total Number of Words indicates the number of samples within any single frequency band.

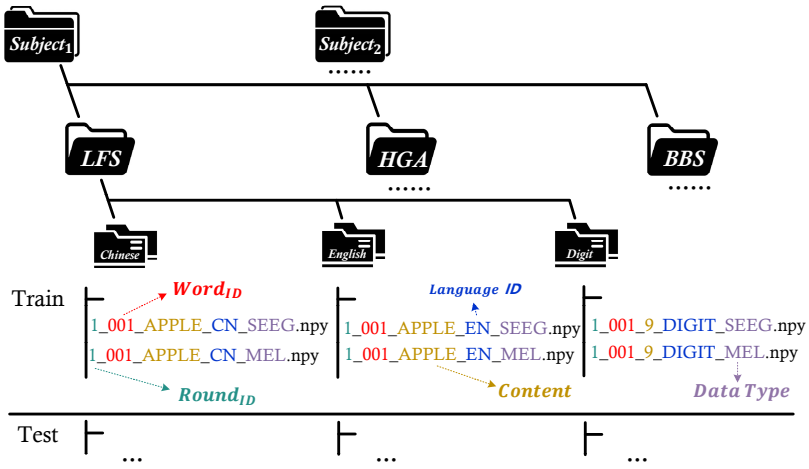

Figure 3: Dataset structure showing the organization of sEEG and audio data, in npy format.

Category	Data
Total Number of Participants	3
Gender Ratio	1:2
Total Number of Sample	9,207
Total Number of Words	3,069
Number of Language	2
Number of Word Types	3
Number of Categories	50

Table 1: CerebroVoice Dataset Card- This table enumerates dataset statistics, such as the total number of participants, gender ratio, total number of samples, total number of words, number of languages, word types, and categories. These factors collectively give an overview of the compiled dataset.

3 Societal Impact

As we point out in Section 7 of the paper, we publish a sEEG-speech dataset that is specifically designed for the study of decoding speech from brain signals. The broad applicability of this dataset is crucial for explaining and predicting the neural mechanisms of human language. We not only confirm the quality and completeness of this dataset, but also verify the feasibility of sEEG-based brain-to-speech synthesis. This brain-to-speech synthesis technology provides new research paths at the intersection of neuroscience and artificial intelligence, especially in decoding spoken language, vocabulary categories, frequency bands, and the influence of decoding models.

Although our innovative research and the application of sEEG-speech datasets have demonstrated their obvious advantages, we need to point out some of the negative social impacts they may have. A major problem is that when not all EEG signals can be accurately decoded into understandable speech, this may limit the expression of the patient's true intentions to some extent. Medical staff often need to combine the patient's facial expressions and physiological reactions to more accurately understand their true intentions.

In addition, this technology may have an impact on patients' right to make their own decisions, as they may feel pressured to accept the technology, even though they may have their own concerns. Therefore, we are actively promoting the introduction of more relevant policies to respect and protect patients' right to choose whether to use this technology. We hope that such policies can help ensure the rights and interests of every individual, while providing an important reference for the use of similar technologies in the future.

4 Access to Dataset

The CerebroVoice dataset, which is available on Zenodo as a general-purpose open repository, is collected, updated, and maintained by team members from the Big Speech Data Laboratory of The xx. Users can fill out an application form via 、`https://forms.gle/xkKzYk5KZwZdaSLD9`, upon which the system will immediately and automatically provide a download link for the dataset. The code for dataset creation and experiments can be accessed at `https://github.com/Brain2Speech2/B2S2`.

5 Licence

We publish all data under CC-BY-4.0 licence. We include detailed instructions on how to obtain our data and provide preprocessing scripts in our GitHub repository. This dataset is intended for research purposes only and not for clinical usage.

6 Implementation Details

6.1 Experimental Parameter

In our experiments, to ensure uniformity and fairness across all experimental setups, we applied identical hyperparameter configurations for all comparison tests. Each model was trained over 300 epochs to guarantee convergence in every experiment. Specifically, we set the batch size to 16 and chose an initial learning rate of 0.0625. Utilizing the Adam optimizer with betas parameters of 0.9 and 0.98 allowed us to regulate the exponential moving average of both the gradient and its squared form, aiming to achieve a balance between training stability and speed. Additionally, we implemented a gradient clipping threshold of 1.0 to effectively mitigate the risk of gradient explosion. Additionally, we implemented a warm-up strategy to stabilize the training process.

6.2 Evaluation Metrics

PCC (Pearson Correlation Coefficient) is a statistical indicator used to measure the strength and direction of the linear relationship between two variables. PCC is the most commonly used metric in the field of sEEG-based speech decoding[1–4]. The value range of this indicator is between -1 and 1, where:

- If PCC is equal to 1, it means that the two variables are completely positively correlated, that is, when one variable increases, the other variable also increases, and the relationship between the two is linear.

- If PCC is equal to -1, it means that the two variables are completely negatively correlated, that is, when one variable increases, the other variable decreases, which is also a linear relationship.
- If PCC is equal to 0, it means that there is no linear relationship between the two variables.

7 Authorstatement

As the authors, we solemnly assure that we accept full responsibility for any possible infringements regarding the data compilation or related proceedings, and commit to promptly taking necessary steps - such as data removal - when dealing with such issues.

8 Information Sheet and Consent Form of Participants

In the following sections, we provide a detailed overview of the Consent Agreement and the Experiment Research Information Sheet. Each participant was required to thoroughly review the Experiment Research Information Sheet before consenting to participate. Upon agreeing to the terms outlined, participants signed the Consent Agreement prior to their involvement in the study.

9 The Comprehensive Performance Evaluation of VAD

Table 2: Comprehensive Performance Evaluation of VAD for Subject 1

sEEG feature	Models	Acc	MR	FAR	ER	Prec	Rec	F1	BA	AUROC
HGA	STANet	0.722	0.070	0.208	0.278	0.245	0.490	0.326	0.624	0.684
	EEGNet	0.728	0.060	0.212	0.272	0.269	0.566	0.365	0.660	0.722
	ECN	0.764	0.035	0.200	0.236	0.338	0.743	0.465	0.755	0.834
LFS	STANet	0.818	0.034	0.148	0.182	0.412	0.755	0.533	0.792	0.856
	EEGNet	0.813	0.033	0.154	0.187	0.405	0.764	0.530	0.792	0.852
	ECN	0.868	0.037	0.095	0.132	0.515	0.732	0.605	0.811	0.905
BBS	STANet	0.801	0.049	0.150	0.199	0.371	0.644	0.471	0.735	0.806
	EEGNet	0.813	0.028	0.159	0.187	0.409	0.797	0.540	0.807	0.867
	ECN	0.876	0.026	0.098	0.124	0.532	0.814	0.644	0.850	0.928

Note: Acc: Accuracy, MR: Miss Rate, FAR: False Alarm Rate, ER: Error Rate, Prec: Precision, Rec: Recall, F1: F1 Score, BA: Balanced Accuracy, AUROC: Area Under the Receiver Operating Characteristic Curve, ECN: EEGChannelNet

Table 3: Comprehensive Performance Evaluation of VAD for Subject 2

sEEG feature	Models	Acc	MR	FAR	ER	Prec	Rec	F1	BA	AUROC
HGA	STANet	0.576	0.073	0.351	0.424	0.239	0.604	0.343	0.587	0.622
	EEGNet	0.509	0.052	0.439	0.491	0.230	0.715	0.348	0.589	0.620
	ECN	0.546	0.045	0.409	0.454	0.252	0.752	0.377	0.626	0.675
LFS	STANet	0.584	0.044	0.371	0.416	0.272	0.757	0.400	0.651	0.699
	EEGNet	0.595	0.043	0.362	0.405	0.278	0.763	0.408	0.660	0.712
	ECN	0.618	0.038	0.344	0.382	0.296	0.790	0.430	0.684	0.752
BBS	STANet	0.629	0.060	0.311	0.371	0.284	0.673	0.399	0.646	0.695
	EEGNet	0.639	0.051	0.311	0.361	0.299	0.723	0.423	0.672	0.724
	ECN	0.666	0.031	0.303	0.334	0.334	0.831	0.476	0.730	0.803

Note: Acc: Accuracy, MR: Miss Rate, FAR: False Alarm Rate, ER: Error Rate, Prec: Precision, Rec: Recall, F1: F1 Score, BA: Balanced Accuracy, AUROC: Area Under the Receiver Operating Characteristic Curve, ECN: EEGChannelNet

Accuracy (Acc): The proportion of correctly identified instances (both true positives and true negatives) over the total number of instances. It provides an overall measure of the model's performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Miss Rate (MR): The proportion of actual positive instances (events where the subject is speaking) that are incorrectly identified as negative (missed). It is also known as the false negative rate.

$$\text{Miss Rate} = \frac{\text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

False Alarm Rate (FAR): The proportion of actual negative instances (events where the subject is not speaking) that are incorrectly identified as positive (false alarms). It is also known as the false positive rate.

$$\text{False Alarm Rate} = \frac{\text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Error Rate (ER): The proportion of all instances that are incorrectly classified. This includes both false positives and false negatives.

$$\text{Error Rate} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

Precision (Prec): The proportion of predicted positive instances that are correctly identified. It indicates the accuracy of the positive predictions.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

Recall (Rec): The proportion of actual positive instances that are correctly identified. It is also known as sensitivity or true positive rate.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

F1 Score (F1): The harmonic mean of precision and recall, providing a single measure that balances both concerns.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

Balanced Accuracy (BA): The average of the true positive rate and the true negative rate. It accounts for class imbalance by considering both recall of the positive and negative classes.

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2} \tag{8}$$

Area Under the Receiver Operating Characteristic Curve (AUROC): A measure of the model's ability to discriminate between positive and negative classes. It plots the true positive rate against the false positive rate at various threshold settings.

$$\text{AUROC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR}) \tag{9}$$

References

[1] 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.

[2] S. Duraivel, S. Rahimpour, C.-H. Chiang, M. Trumpis, C. Wang, K. Barth, S. C. Harward, S. P. Lad, A. H. Friedman, D. G. Southwell *et al.*, "High-resolution neural recordings improve the accuracy of speech decoding," *Nature communications*, vol. 14, no. 1, p. 6938, 2023.

[3] M. Angrick, M. C. Ottenhoff, L. Diener, D. Ivucic, G. Ivucic, S. Goulis, J. Saal, A. J. Colon, L. Wagner, D. J. Krusienski *et al.*, "Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity," *Communications biology*, vol. 4, no. 1, p. 1055, 2021.

[4] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker, "A neural speech decoding framework leveraging deep learning and speech synthesis," *Nature Machine Intelligence*, pp. 1–14, 2024.