論文 / 著書情報
Article / Book Information

Title	Multimodal recognition of speech and electrocorticogram
Author	Mitali, Shuji Komeiji, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, Toshihisa Tanaka
Journal/Book name	2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), , , pp. 546-550
Pub. date	2023, 11
DOI	https://doi.org/10.1109/APSIPAASC58517.2023.10317527
Copyright	(c)2023 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	This file is author (final) version.

Multimodal recognition of speech and electrocorticogram

Mitali Ahuja*, Shuji Komeiji†, Takumi Mitsuhashi‡, Yasushi Iimura‡,
Hiroharu Suzuki‡, Hidenori Sugano‡, Koichi Shinoda* and Toshihisa Tanaka†
* Department of Computer Science, Tokyo Institute of Technology
† Department of Electronic and Information Engineering, Tokyo University of Agriculture and Technology
‡ Department of Neurosurgery, Juntendo University School of Medicine

***Abstract*—Brain-Computer Interface (BCI) provides a novel way of communicating with computers but their poor performance in terms of speed and accuracy in comparison to the other modes of communication has been the biggest obstacle in their usage for practical applications. In this work, we aim to enhance the performance of BCI by utilizing speech data collected along with the electrocorticogram (ECoG) recordings when the person is speaking. While some BCI users may have difficulty in speaking at all, many of them can speak, even though their speech may be unclear. We propose that information from this speech data can help in improving accuracy and to employ such speech context to assist the decoding process, we apply a multimodal recognition method. We use speech data contaminated by noise in our evaluation to simulate the cases where the available speech quality is low. Our experiments using data from five subjects suffering from Epilepsy show that our method of using multimodal input has a significant improvement, an absolute reduction in phrase error rate by 1.1 points from recognition using speech alone and by 51.3 points from recognition using ECoG alone.**

I. Introduction

Brain-Computer Interfaces (BCIs) allow humans to communicate with computers directly through their brains by translating their brain activity into computer-readable and implementable formats. Their application areas are rehabilitation of disabled people along with non-medical fields such as Gaming, Internet of Things (IoT), and autonomous driving [1]–[3].

BCIs are not yet widely adopted in our society because of their sub-par performance due to challenges such as the trade-off between quantity and quality of brain data that can be recorded, complex and non-stationary brain signals, and difficult feature extraction process [1]. Past studies achieved acceptable results for several tasks such as controlling a cursor or prosthesis by translating the brain activity produced while imagining motor movements [4], [5]. However, such BCIs are still limited in terms of their applicability.

Lately, the focus has been on the decoding of speech from the brain as it further extends the scope of usability. Amongst different methods of brain data acquisition, Electrocorticography (ECoG), which measures the electrical activity of the brain from sensors placed directly on the cortical surface, has been frequently used due to its high temporal and spatial resolution which is required for fast-paced continuous speech recognition [6]. However, its invasive setup restricts its usage only under strict medical requirements resulting in the availability of limited amounts of data.

Most speech-based BCI studies perform spoken speech recognition as the first step towards implementing BCIs based on imagined speech since it is easy to monitor changes in brain activity when corresponding speech data is available. Ref. [7] was the first to decode continuously spoken speech from ECoG using the Gaussian mixture model.

Recently, advancements in the field of Automatic Speech Recognition (ASR) are being employed to BCIs even though the ECoG datasets are fairly smaller in size compared to frequently used speech datasets. For example, [8] used a Bi-directional Long Short Term Memory (BLSTM) based sequence-to-sequence encoder-decoder architecture to translate brain signals into sentences. Their approach was based on multi-task learning which required training the encoder with features obtained from the simultaneously produced speech and they achieved Word Error Rate (WER) as low as 3% for a subject. Ref. [9] extended this approach by switching the encoder architecture from BLSTM to Transformer as Transformers are better at learning long-range dependencies thereby achieving improved performance. Ref. [10] proposed a novel Brain2Char architecture that employed a Connectionist Temporal Classification (CTC) Decoder along with an external language model to avoid dealing with the alignment between speech and brain and they also achieved low WER of around 10% for a much larger vocabulary.

While these studies achieved good performance, their accuracy degrades with less training data or if the electrode arrays used to measure brain activity are less dense. To mitigate this problem, they assist their model with simultaneously recorded speech data by guiding their network with speech features so that it learns to make better predictions, but the improvement in performance is still not as significant. Another limitation is that none of them consider the scenario where clean spoken speech data is not available such as where the subject has difficulty in speaking.

Taking inspiration from studies employing multimodal speech recognition [11], we propose to utilize the speech information by using it as input along with ECoG. Different from previous works where speech features were used as a target for intermediate layers, we combine information from

speech and ECoG features in a multimodal manner through a late fusion approach so that we can utilize the speech context efficiently. Additionally, to simulate the case when available speech quality is low, we propose to use noise-added speech in our experiments. Our aim with this approach is to show that if speech modality is available, irrespective of the quality of speech, it can be employed in BCIs to improve their performance.

II. Method

A. Dataset

Five subjects (js4, js5, js6, js7, and js8) undergoing brain monitoring through the intracranial placement of ECoG arrays for treatment of epilepsy at Juntendo University Hospital volunteered to be a part of this study. A written, informed consent approved by the Ethics Committee of Tokyo University of Agriculture and Technology, Tokyo Institute of Technology, and Juntendo University Hospital was obtained beforehand. The details about this dataset are summarised below but more information can be referred to from [9].

During the experiment, the subjects were asked to speak out loud the sentence being displayed on a screen in front of them while their corresponding speech and ECoG data were collected. All subjects were conscious during the experiments and had no trouble speaking.

The sentences used for the experiments consisted of 8 different Japanese language sentences. Each sentence is made up of three phrases. The first phrase is either 「わたしは」(I) or 「きみと」(with you), second is 「がっこうへ」(to school) or 「しょくばに」(to office) and the last one is 「いった」(went) or 「むかう」(go) .

Speaking one sentence constituted one trial, and 80 trials were recorded for each subject. The speech and ECoG data were digitized at a sampling frequency of 9,600 Hz. The number and placement of electrodes varied between subjects because it depended on their medical condition, so we adopted a subject-dependent approach by performing experiments for each subject separately.

B. Preprocessing

Simultaneously recorded speech and ECoG data for each subject were trimmed to a length equal to the maximum duration of the sentence spoken by that subject so that the model does not learn to predict the output based only on the length of the sentence.

To be inclusive of conditions where clean speech is not available or feasible, we added Signal-to-Noise Ratio (SNR) equal to -10 white noise to our raw speech data. Since it's extremely difficult to collect data from speech-disabled subjects with implanted ECoG arrays who also volunteer to participate in research, preparing such data by manually adding noise helped in simulating conditions close to the real-world scenario.

C. Network Architecture

Taking inspiration from recent works which showed end-to-end architectures can decode speech from ECoG efficiently [8]–[10], we decided on employing a similar model consisting of an encoder and a decoder with CTC and a Recurrent Neural Network (RNN) based language model for our task.

As shown in Fig. 1, our network consists of the following components:

1) **Encoder:** It contains stacked bi-directional LSTM layers which take input features $\mathbf{X}$ of dimension $[T, f]$ where T and f represent time frames and features respectively and outputs a hidden vector $\mathbf{h}$ of dimension $[T', H]$, where T' is subsampled time frames, and H is the number of projection units in the final layer of the encoder.
2) **Decoder:** Our decoder is composed of CTC block and an RNN-based language model.
   CTC: It consists of LSTM layers that take the hidden vector from the encoder as input and output probabilities over each token c_{t} in the dictionary for each time frame. The output probability for tokens in the dictionary is given as

$$p(c_t|X) = \mathrm{Softmax}(\mathrm{LinB}(\mathbf{h}_t)), \tag{1}$$

   where LinB and $\mathrm{Softmax}$ denote the linear layer and the softmax function, respectively.
   RNN language model: To assist CTC, we use an external RNN-based model that calculates the probability of a sequence of tokens as follows:

$$p(c_1, c_2, \ldots, c_n) = p(c_1)p(c_2|c_1)\ldots p(c_n|c_{1:n-1}) \tag{2}$$

During training, the encoder and CTC are trained to maximize the model's probability of predicting correct output by minimizing the corresponding CTC loss. RNN language model is trained separately using the text corpus.

During inference, the hidden vector from the trained encoder is fed to the decoder. For each time step t, the probabilities from the CTC block and RNN language model are fused, and the token with the maximum probability is selected. The obtained output for the current time step is then fed to the language model to predict the probabilities for the next token. Finally, when all time frames are exhausted, or <EOS> label is predicted, beam search decoding is performed to get the final predicted output.

D. Decoding Approach

To test our proposed idea of using speech features as input along with ECoG will help in making the model learn better from speech modality, we performed speech recognition experiments in the following three steps. First, we used only noisy speech features as input and employed the architecture shown in Fig. 1 while discarding the ECoG branch to perform speech recognition. Next, we used only ECoG features as input while dropping the speech branch and performed the same experiment again. Finally, using noisy speech and ECoG features

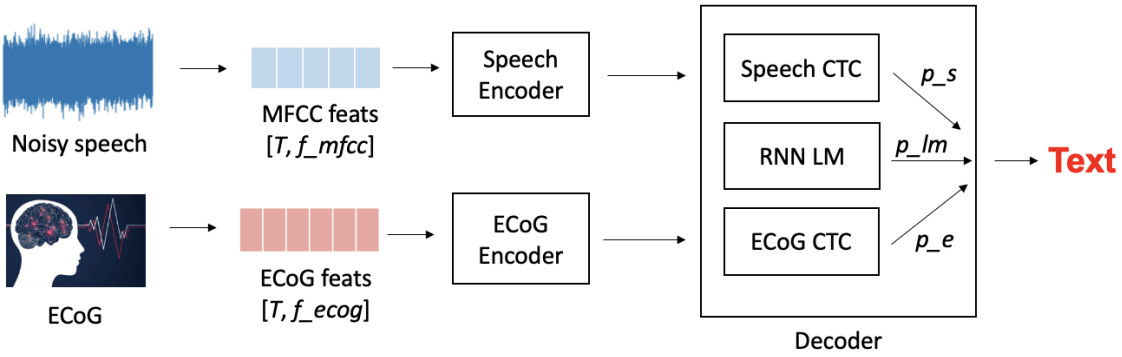

Fig. 1. Network architecture for multimodal recognition of speech and ECoG

together as input, we performed multimodal recognition and compared performance under these three scenarios.

For combining information from speech and ECoG, we used a late fusion method. We first retrieved the models trained on speech and ECoG modality individually. Then, in the decoder, we used their respective CTC scores and combined them with the language model score as follows:

$$\hat{c}_{\mathrm{t}} = \underset{c_{\mathrm{t}}}{\operatorname{argmax}} \{ w_{\mathrm{s}} \times \log p_{\mathrm{ctc}}(c_{\mathrm{t}}|X_{\mathrm{s}}) + w_{\mathrm{e}} \times \log p_{\mathrm{ctc}}(c_{\mathrm{t}}|X_{\mathrm{e}}) + w_{\mathrm{lm}} \times \log p_{\mathrm{lm}}(c_{\mathrm{t}}) \} \tag{3}$$

to get the most likely set of tokens for each frame. In (3), w_{s} and w_{e} are the weights assigned to the speech and ECoG CTC blocks, respectively, and w_{lm} is the weight for the language model. We believe that fusing information this way should prevent the model from becoming biased towards any particular modality while helping it in learning from each of these components effectively.

E. Feature Extraction from Speech

We extracted Mel Frequency Cepstral Coefficients (MFCCs) as acoustic features from speech data as they have been successfully used in many ASR as well as speech-based BCI-related studies [8]–[11]. For the calculation of MFCCs, a window of 20 msec with a frame period of 5 msec was used. We obtained MFCC features of dimension $[T, f_{\mathrm{mfcc}}]$ where T is the number of frames and f_{mfcc} is the dimension for features which was set to 13, similar to previous studies [8], [9]. The values for window length and frame period were set so that the output dimension of MFCC features across the time axis coincides with that of ECoG features.

F. Feature Extraction from ECoG

Our feature extraction process from ECoG has been inspired by past studies [8], [9] as they achieved state-of-the-art results in decoding speech from neural signals.

We began with removing channels that showed obvious signs of contamination by noise resulting in 54 channels for subjects js4, js5, js6 and js8 and 53 channels for subject js7. Next, a lowpass filter at 200 Hz followed by downsampling at 400 Hz was performed. A notch filter was applied at 50 Hz and 100 Hz to attenuate power line noise. To extract brain activity in the high gamma range (70–150 Hz), which is shown to be associated with speech [8], we used eight bandpass filters with centers between 70 and 150 Hz and averaged them across bands. ECoG data was further downsampled at 200 Hz and normalized using the z-score method to avoid bias due to the presence of any outliers.

The resultant ECoG features had dimension $[T, f_{\mathrm{ecog}}]$ where T is the number of time frames which is the same as that obtained for MFCC features and f_{ecog} is equal to the number of channels or electrodes used for that subject.

G. Implementation

We performed our experiments using ESPnet [12], an end-to-end speech processing toolkit. Our speech recognition model consisted of an encoder with four BLSTM layers; each layer had 320 units followed by a projection layer with 320 units. The subsampling factor for the encoder was set to 1/8 to reduce computational complexity. The CTC block consisted of one LSTM layer with 300 units. For the RNN language model, we trained a phrase-level model using one BLSTM layer with 1,000 units.

Since the focus of our work was to test how speech and ECoG features impact the recognition performance, we kept the language model weight constant at 0.5 across all types of input while the remaining 0.5 came from the weights for speech and/or ECoG CTC blocks. For recognition using only noisy speech as input, $w_{\mathrm{s}} = 0.5$ and $w_{\mathrm{e}} = 0.0$, for ECoG only input, $w_{\mathrm{e}} = 0.5$, and $w_{\mathrm{s}} = 0.0$, and lastly, for the case of multimodal features, w_{lm} was kept fixed at 0.5, while we tried different combinations for w_{s} and w_{e}.

H. Evaluation

As discussed in Section II-A, we performed speaker-dependent experiments, which restricted us from pooling data across subjects resulting in around five minutes of data per subject. To mitigate the effects of data insufficiency, we performed a five-fold cross-validation where for each fold, we split our data (80 trials) into the train (56 trials), validation (eight trials) and test set (16 trials).

TABLE I
AVERAGE PER(%) FOR SPEECH RECOGNITION USING DIFFERENT INPUT FEATURES

Features	PER(%)
Speech	6.7
ECoG	56.9
Multimodal	**5.6**

We evaluated the performance of our model in terms of Phrase Error Rate (PER), which is given as

$$\text{PER} = \frac{S + I + D}{N}, \tag{4}$$

where S, I, and D are, respectively, the minimum number of substitutions, insertions, and deletions required to convert the predicted output into ground truth, and N is the total number of phrases in the actual output. Since this is an error rate, a lower PER is desirable.

III. RESULTS

The average PERs obtained across five subjects for speech recognition using different features as input (Speech, ECoG, and Multimodal) have been presented in Table I. Here, the results presented for multimodal features are when using the following combination of weights, $w_{\text{lm}} = 0.5$, $w_{\text{s}} = 0.3$ and $w_{\text{e}} = 0.2$.

As can be verified from Table I, recognizing speech using multimodal features has the lowest PER. Our late fusion method significantly reduced PER by 51.3 points from the case of using only ECoG features and by 1.1 points from the case of using only speech features. This proves that through proper utilization of information from speech, we can majorly enhance the accuracy of BCI systems. In particular, PER obtained using multimodal features was even lower than that obtained for using only speech features, which has been the goal of this research, i.e., to make the performance of speech recognition from ECoG comparable to that of ASR systems to make it possible to employ them for different applications.

We also tested our method for different combinations of Speech and ECoG CTC weights while keeping the language model weight fixed at 0.5, and the results have been presented in Fig. 2. When $w_{\text{e}} = 0.0$, w_{s} becomes 0.5, which correponds to the case of using only speech as input, whereas when $w_{\text{e}} = 0.5$, w_{s} becomes 0.0 corresponding to the case of using only ECoG features as input. For all subjects, multimodal features have the lowest PER and, therefore, better performance than using only speech or only ECoG features for different combinations of speech and ECoG CTC weights, thereby proving the effectiveness of our method.

IV. DISCUSSION

Our proposed method was able to achieve competing PERs for all subjects even with less training data and a lesser number of electrodes as compared to previous studies [8], [10]. Usually, it is difficult to directly compare these error rates because of the large gap in the experiment conditions,

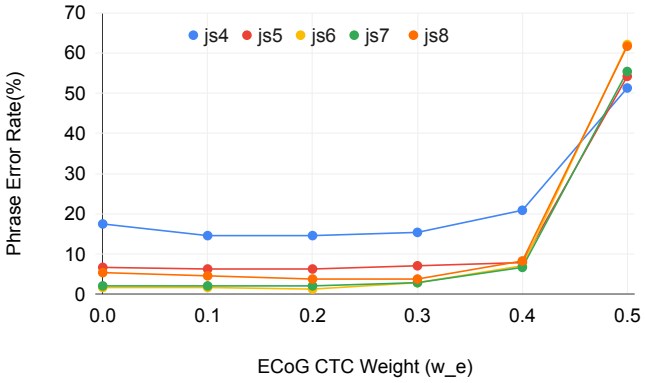

Fig. 2. PER(%) for different fusion weights for speech and ECoG

TABLE II
COMPARISON OF PER(%) FOR SPOKEN SPEECH RECOGNITION

Method	js4	js5	js6	js7	js8	Avg.
Komeiji [9]	18.7	35.8	33.1	30.9	41.9	32.1
Ours	**14.6**	**6.3**	**1.3**	**2.1**	**3.8**	**5.6**

such as vocabulary size, variability across subjects as well as the implementation architecture. However, just to evaluate the improvement in performance using our proposed approach, we compared our results with Komeiji et al. [9] since they have also used the same dataset and evaluation metric as ours. As presented in Table II, our method of using speech along with ECoG as input exceeds their performance where they used speech features to train the encoder by a large margin across all five subjects showing a promising application of our approach in future.

V. CONCLUSION

In this paper, we proposed multimodal recognition of ECoG for improved BCI communication. To our knowledge, we are the first to employ speech features as input and use noise-added speech data in our experiments. We achieved an average PER of 5.6% across subjects which outperforms the previous state-of-the-art results.

Currently, our method requires spoken speech data from subjects to perform multimodal recognition. However, we should also cater to scenarios where it is not available such as when the subject has completely lost their ability to speak or for decoding imaginary speech from ECoG. We would like to extend our method so that it does not rely on the necessity of having spoken speech data from the same subject during inference. We would also like to evaluate our method with non-invasive brain data acquisition techniques such as EEG or fMRI where we can explore a bigger dataset with many subjects and a larger vocabulary size in the future.

ACKNOWLEDGMENT

This work was supported in part by JSPS KAKENHI 20H00235.

REFERENCES

[1] Abdulkader S N, Atia A, Mostafa MS M, "Brain computer interfacing: Applications and challenges", in *Egyptian Informatics Journal*, Volume 16, Issue 2, 2015, Pages 213-230, URL: http://www.sciencedirect.com/science/article/pii/S1110866515000237

[2] Ajmeria R et al., "A Critical Survey of EEG-based BCI Systems for Applications in Industrial Internet of Things," in *IEEE Communications Surveys Tutorials*, URL: https://ieeexplore.ieee.org/abstract/document/10000406

[3] Blankertz B, Tangermann M, Vidaurre C et al., "The Berlin Brain–Computer Interface: Non-Medical Uses of BCI Technology", in *Frontiers in Neuroscience*, Volume=4, 2010, URL: https://www.frontiersin.org/article/10.3389/fnins.2010.00198

[4] Pandarinath C et al., "High performance communication by people with paralysis using an intracortical brain–computer interface" in *eLife*, 6, e18554 (2017), URL: https://doi.org/10.7554/eLife.18554

[5] Attallah O, Abougharbia J, Tamazin M et al., "A BCI System Based on Motor Imagery for Assisting People with Motor Deficiencies in the Limbs" in *Brain Sciences*, 2020, 10, 864, URL: https://doi.org/10.3390/brainsci10110864

[6] Schalk G, Leuthardt E C, "Brain-Computer Interfaces Using Electrocorticographic Signals," in *IEEE Reviews in Biomedical Engineering*, vol. 4, pp. 140-154, 2011, URL: https://ieeexplore.ieee.org/abstract/document/6047564

[7] Herff C, Heger D, De Pesters A, "Brain-to-text: Decoding spoken phrases from phone representations in the brain" in *Frontiers in Neuroscience*, Volume=9, Year=2015, Pages=217, URL: https://doi.org/10.3389/fnins.2015.00217

[8] Makin J G, Moses D A, Chang E F, "Machine translation of cortical activity to text with an encoder-decoder framework", *Nature Neuroscience*, 23, 575–582 (2020), URL: https://doi.org/10.1038/s41593-020-0608-8

[9] Komeiji S et al., "Transformer-Based Estimation of Spoken Sentences Using Electrocorticography", in *International Conference on Acoustics, Speech and Signal Processing, (ICASSP) 2022*, pp. 1311-1315, URL: https://doi.org/10.1109/ICASSP43922.2022.9747443

[10] Sun P, Anumanchipalli G K and Chang E F, "Brain2Char: A Deep Architecture for Decoding Text from Brain Recordings", in *Journal of Neural Engineering*, 2019, volume=17, URL: https://doi.org/10.48550/arXiv.1909.01401

[11] Shruti Palaskar, Ramon Sanabria and Florian Metze, "End-to-end multimodal speech recognition", in *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 5774-5778, 2018, URL: https://doi.org/10.48550/arXiv.1804.09713

[12] Watanabe S, Takaaki H, Shigeki K, Tomoki H et al., "Espnet: End-to-end speech processing toolkit" in *arXiv e-prints*, 2018, URL: http://arxiv.org/abs/1804.00015

[13] Martin S, Brunner P, Iturrate I et al., "Word pair classification during imagined speech using direct brain recordings" in *Sci Rep* 6, 25803 (2016), URL: https://doi.org/10.1038/srep25803