A Study of Deep Learning Based Classification of Mandarin Vowels Using Spoken Speech EEG Signals

Wenyuan Cui*, Xinyu Wang*, Mingtao Li*†, Sio Hang Pun† and Fei Chen*
* Southern University of Science and Technology, Shenzhen, China
E-mail: 11912304@mail.sustech.edu.cn, 12232141@mail.sustech.edu.cn, limt@mail.sustech.edu.cn,
fchen@sustech.edu.cn Tel/Fax: +86-755-88018554
† University of Macau, Macau, China
E-mail: YB97473@umac.mo, lodgepun@um.edu.mo Tel/Fax: +853-882244

2023 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC) | 979-8-3503-1672-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSPCC59353.2023.10400255

Abstract—Brain computer interface (BCI) based on imagined speech provides an alternative access for human-computer interaction and can help patients who are unable to speak to communicate with the outside world. However, the complexity and variability of speech-related electroencephalographic (EEG) signal make decoding speech information from EEG signals a challenging task. This study aimed to explore feature extraction methods and classifiers that were suitable for speech-related EEG classification. Six feature extraction techniques and three classifiers were used to classify spoken speech EEG signals of four Mandarin vowels. The classification results showed that the accuracy reached 68.7% when using channel cross-covariance (CCV) with Riemannian manifolds as input features and deep convolutional neural network (Deep ConvNet) as the classifier, which was better than classification performance of other methods and previous studies. This study provides insights into the effectiveness of various feature extraction and classification methods in EEG-based speech classification. It suggests the potential of CCV combined with Riemannian manifold features and a Deep ConvNet based classifier for speech imagery studies.

Keywords—speech imagery, electroencephalogram (EEG), brain computer interfaces (BCI)

I. INTRODUCTION

The direct-speech brain computer interface (BCI) is an advanced system that uses neural signals to facilitate direct communication between human brain and external devices [1]. This emerging technology is particularly suitable for individuals who suffer from language impairments, such as locked-in syndrome [2] and aphasia [3]. Through BCI systems, individuals who are limited in their ability to communicate can effectively express their thoughts, intention, and desires without the need for conventional verbal communication. Electroencephalography (EEG), as a non-invasive neuroimaging technique with very high temporal resolution and easy handling, is an important tool in neuroscience research.

The current studies cover a variety of speech-related EEG signals, mainly including spoken speech, intended speech and imagined speech [e.g., 4]. The sound generated via human articulation can be perceived by the speaker himself and nearby listeners, which is overt speech, i.e., spoken speech. Covert speech includes intended speech and imagined speech. Intended speech refers to attempt to speak verbally without producing sounds. Imagined speech involves internal vocalizations, but does not produce audible sounds or corresponding movements. Previous studies showed that imagined speech and spoken speech had similarities in spatial and temporal features [e.g., 5], which implied that spoken speech research could promote the development of imagined speech research.

Speech-related EEG signals are highly complex, reflecting patterns of neural activity in the brain, which makes direct analysis of these signals difficult and less effective. Through feature extraction, complex neural signals can be transformed into meaningful information that can be used for analysis and interpretation. There have been a number of studies using feature extraction in EEG classification tasks [e.g., 6-8]. The study by Lu et al. was based on discrete wavelet transform (DWT) analysis of EEG signals and approximate entropy for epilepsy detection [6]. Zhang et al. used the common spatial patterns (CSP) to extract features of EEG signals for a motor imagery classification task [7]. Dae-Hyeok et al. improved the decoding accuracy of imagined speech by 7.42% over imagined speech EEG by fusing spoken speech EEG features and imagined speech EEG, which suggested the positive and effective role of spoken speech EEG features in enhancing the decoding accuracy of imagined speech [8].

Most of the captured EEG features do not have large data volume and complexity. Therefore, machine learning methods are often applied in the classification stage. In the study by Le et al., support vector machine (SVM) was applied to the imagined word classification task with covariance-based connectivity features, achieving a maximum accuracy of 76.4% [9]. The study by Moctezuma et al. conducted comparisons between four different machine learning algorithms, and the results showed that SVM had the best accuracy [10].

Recently, neural networks have become increasingly popular in EEG studies. Deep learning methods enable algorithms to automatically extract useful features for classification, minimizing information loss during feature extraction [11]. Moreover, they can be directly applied to large datasets, efficiently batching and parallelizing the data, and speeding up the training process of models. Tamm et al. adopted a low-complexity convolutional neural network (CNN) and achieved 32.75% accuracy in the task of vowel classification [12]. Many well-known EEG classification architectures are also based on CNNs, such as ATCNet [13], EEGNet [14], Deep ConvNet and Shallow ConvNet [15]. Chengaiyan et al. used recurrent neural network (RNN) and deep belief network (DBN) for vowel classification, with accuracy rates of 72% and 80%, respectively, showing that

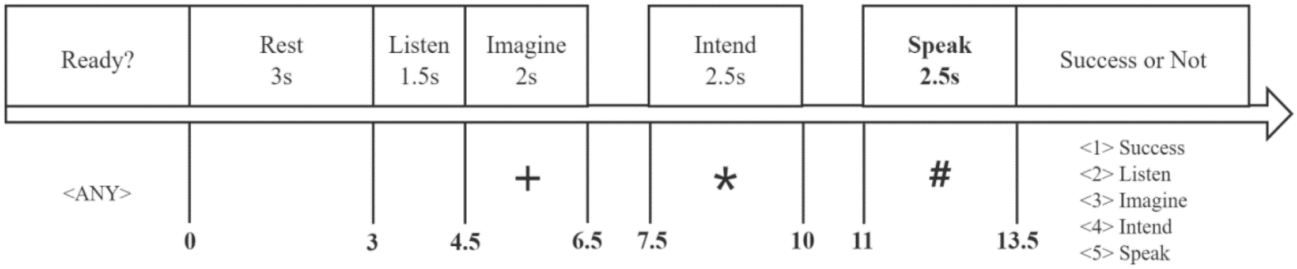

Fig 1. The experimental paradigm [21].

DBN was superior to RNN in speech EEG recognition [16]. The research by Saha et al. combined two typical networks, i.e., fused CNN and long-term short-term network, and achieved a significant improvement of 22.5% compared with previous speech-related EEG binary classification methods [17]. Although deep learning methods require large amounts of data and computing resources, they often outperform traditional machine learning methods due to their ability to automatically extract meaningful information from datasets. In contrast, manual feature extraction used in traditional machine learning methods may be limited by the loss of important information [18].

There have been several studies on speech-related EEG decoding. The work of Young-Eun et al. utilized Transformer to classify 13 classes of overt speech and obtained an accuracy of 49.5% [19]. The study of Rojas et al. classified two groups of grammatical classes of vocabulary using EEG signals and electromyography signals of overt speech, and the accuracy rates were 92.6±2.09% and 94.84±4.71% respectively [20]. Li et al. classified four classes of Mandarin vowels using a Riemannian manifold method and a linear classifier with an average accuracy of 63.1% [21].

Early studies did not explore the most suitable feature extraction methods and classification methods for speech-related EEG signal classification tasks. The purpose of this work is to explore the most suitable combination of feature extraction method and classifier for speech-related EEG signal classification tasks. This study classified four Mandarin vowels, i.e., /a/, /u/, /i/, and /ü/, on the collected spoken speech EEG dataset. In order to explore the best performing method, six commonly used features including time domain, frequency domain, spatial domain, and multi-channel features were compared. The extracted features were classified using three different classifiers, i.e., SVM, linear discriminant analysis (LDA), and Deep ConvNet. The results were validated using a 5-fold cross-validation.

II. Methods

A. Dataset

The dataset used in this experiment was from the Southern University of Science and Technology [21] and consisted of EEG data from 11 participants, 4 females and 7 males, with a mean age of (22.6 ± 3.2) years. One participant was excluded from the final analysis due to a data saving error. The stimuli used were 70 monosyllabic Mandarin words, including 54 words combining four vowels (/a/, /u/, /i/, and /ü/) and five consonants (/b_/, /f_/, /j_/, /l_/, and /m_/), as shown in Table I. The stimuli also included 16 words consisting of four individual vowels and four Mandarin tones. The dataset consisted of seven states, as shown in Fig. 1. This study used data from the speak state. Participants were asked to speak the phonemes during a 2.5-second speak state after listening each stimulus.

Scalp EEG signals were recorded using a 64-channel electrode cap (Neuroscan Inc.) with a sampling rate of 500 Hz. Electrodes were placed according to the extended 10-20 system, with the ground electrode on the forehead, the reference electrode on the tip of the nose, and additional reference electrodes on the bilateral mastoid. Two electrodes above and below the left eye measured the electrooculography. More details of the dataset can be found in [21].

TABLE I. The List of Monosyllabic Stimuli

CV	b_	f_	j_	l_	m_	_
/a/	ba	fa		la	ma	a
/u/	bu	fu		lu	mu	u
/i/	bi		ji	li	mi	i
/ü/			ju	lü		

B. Data Preprocessing

EEG data were preprocessed using EEGLAB 13.5.4b. The raw data were re-referenced using bilateral mastoid channels and then band-pass filtered between 0.5 and 70 Hz. Power noise was removed using a trap filter with a cutoff frequency of 49~51 Hz. The data are then subjected to independent component analysis (ICA) and artifacts are manually removed, including electrocardiography, electrooculography, electromyography, etc [22].

For analysis, 2500 milliseconds EEG data from all channels during the speak state were extracted and baseline corrections were made using 200 milliseconds data before the speak state. The VOICEBOX toolbox in MATLAB was used to determine the onset time of speaking. Based on the speaking time, the EEG data for each trial was limited to 800 milliseconds, covering most of the articulation time for all subjects. Data of the 6th subject was removed due to poor data quality. The first data block of the 11th subject was removed due to improper saving of the audio file. After preprocessing, the EEG data consisted of 3,297 trials, each containing 62 channels, with 1,250 sample points. This structure is represented as a three-dimensional array with dimensions of 3297 × 62 × 1250, indicating the number of trials, the number of channels, and the sample length of each trial.

C. Feature Extraction

EEG signals have a complex and abstract temporal structure. Valuable information needs to be extracted from these signals by feature extraction methods, which transform the raw neural signals into more comprehensible and manipulable feature representations. Six feature extraction methods were used and compared in this experiment, which are statistical features, short-time Fourier transform (STFT), Mel frequency cepstrum coefficients (MFCCs), DWT, CSP, and channel cross-covariance (CCV) using Riemannian manifolds.

In this experiment, a variety of time-domain statistical features of the EEG signal were extracted, including mean, median, standard deviation, variance, maximum, minimum, maximum±minimum, sum, spectral entropy, energy, skewness, kurtosis, and first- and second-order derivatives of these features [23].

STFT is a time-frequency analysis method that decomposes a signal into a series of spectral components within a short time window. MFCC is a feature extraction method for speech and audio signal processing. It divides frequencies into Mel scales by simulating the auditory properties of the human ear and calculates the energy or power of each Mel frequency channel. MFCC performs Mel filtering and logarithmic operations on top of STFT and then applies discrete cosine transform in order to extract the spectral coefficients. The DWT decomposes the signal into spectral components at different scales. By performing several iterations of low-pass filtering and high-pass filtering on the signal, DWT can obtain information about different frequency components.

CSP is a spatial domain feature extraction method. It obtains the spatial pattern of EEG activity in two different states by spatially filtering the EEG signals.

CCV is a multi-channel feature extraction method that extracts information exchanged between different brain regions [24], typically used for Riemannian manifolds. Riemannian manifolds are used as a per-channel feature extraction method, which is effective in capturing the spatial-temporal dynamics of EEG data [21]. The method first calculates for each trial the covariance matrix of the EEG signal, reflecting the correlation between the different electrodes. These matrices are normalized and mapped onto a Riemannian manifold. Local features are extracted by performing a convolution operation on the Riemannian manifold. Reduced dimensionality using principal component analysis method retains 90% of the original data information.

D. Classification

This experiment utilizes and compares some common machine learning techniques, including SVM, LDA and CNN-based neural network Deep ConvNet [15], aiming to explore the application of these techniques in understanding the speech EEG classification.

LDA is a widely used classification and dimensionality reduction method in speech EEG data for finding linear combinations of features that optimally separate multiple classes. It seeks the optimal projection of features that maximizes the separation between classes while minimizing the variance within classes. SVM is a well-established supervised machine learning technique known for its ability to work with high-dimensional data and provide robust classification even with limited training samples. SVM separates different classes by finding the optimal hyperplane by maximizing the margins between the classes in the feature space.

Deep ConvNet is a CNN-based neural network [15]. The term “deep” refers to the increased number of layers in the network, which allows it to capture complex patterns and hierarchical representations that may not be accessible to shallower CNNs. The network consists of four convolutional layers, each of which is followed by a max pooling layer to prevent the network from overfitting. The network concludes with a fully connected layer for the output of the classification task. In addition, the network differs from a typical CNN by the use of spatial filters, which helps to better process information from different channels in the data.

This experiment used five-fold cross-validation. Data from all subjects were divided into 5 folds, and in each fold the dataset was proportionally divided into three subsets, of which 80% was the training set, 10% was the test set, and 10% was the validation set.

III. Results

Table II demonstrates the classification accuracy of the three classifiers for all six features. The best results are highlighted in bold. The highest overall accuracy of 68.7% was obtained using a Deep ConvNet and CCV combined with Riemannian manifold features. The confusion matrix for this method is shown in Table III. The classification accuracies for the four vowels /a/, /i/, /u/ and /ü/ are 80.6%, 71.5%, 71.7% and 52.1%, respectively. The accuracy of vowel /ü/ is lower than those of the other three vowels, but it is improved by 22.9 percentage compared to [21] (i.e., 29.2%). The vowel /ü/ is mainly misclassified as the vowels /i/ and /u/, with misclassification rates of 21.6% and 20.3%, respectively.

The results for all combinations exceeded the probability level (25%). Among them, the results for all three classifiers using the CCV feature are above 60%, higher than those using other features. It is noteworthy that the result using

TABLE II. Accuracy of The Three Classification Methods on Six Features

Feature	LDA	SVM	Deep ConvNet
Statistical Features	41.2	48.6	65.1
STFT	36.1	40.6	42.0
MFCCs	35.9	40.4	40.6
DWT	38.9	45.2	59.8
CSP	52.0	58.8	55.3
Riemannian Method	63.9	66.7	**68.7**

TABLE III. Confusion Matrix of Remannian Method Combined With Deep ConvNet

	/a/	/i/	/u/	/ü/
/a/	80.6±5.5	6.5±3.5	9.0±2.2	3.8±4.8
/i/	6.0±1.0	71.5±7.9	14.0±4.2	8.4±4.0
/u/	8.1±3.1	12.8±4.8	71.7±5.4	7.4±1.4
/ü/	6.0±4.0	20.3±2.7	21.5±5.1	**52.1±4.9**

statistical features and Deep ConvNet achieves an accuracy of 65.1%. The results of three classifiers for the CSP feature exceed 50%, which is higher than other features except CCV.

In addition, since the deep neural network can be directly applied to raw EEG data without feature extraction, we also used the raw EEG data as an input to compare the experimental results when using Deep ConvNet as a classifier. The accuracy was 60.1% when the input was the raw EEG data.

IV. Discussion and Conclusion

In this study, a vowel classification task was performed on the EEG dataset of spoken modality through a combination of different feature extraction methods and classifiers. Six feature extraction techniques were used in this experiment, including statistical features, STFT, MFCCs, DWT, CSP, and CCV with Riemannian manifold. Three common classifiers were used, including LDA, SVM and CNN-based Deep ConvNet. The results showed that among the six different feature extraction methods, CCV achieved the best accuracy with different classifiers. Among the three classifiers, Deep ConvNet outperformed the other two traditional machine learning algorithms on most features. When comparing the performance of SVM and LDA, SVM showed superior results. In terms of statistical features, the performance gap between the Deep ConvNet model and the SVM model was much larger, whereas the gap narrowed when using the CCV and Riemannian manifold methods.

Notably, Deep ConvNet achieved over 60% accuracy without feature extraction. This result was better than most of the results of feature extraction combined with two traditional machine learning classifiers. However, it was still lower than the accuracy of using the CCV based on Riemannian manifold, which suggested the advantage of this feature extraction method.

Compared to the results obtained by LDA applied to the same dataset [21], the method that obtained the best results in the present study was more stable in each class and achieved an improvement in the vowel /ü/. This suggested that the deep learning method was better at distinguishing vowels with a similar formant structure.

In conclusion, this study demonstrates the effects to classify vowels using spoken EEG when combining different feature extraction method and classifier. The results showed that the combination of CCV with Riemannian manifold features performed best among the 6 features, and the Deep ConvNet surpassed traditional machine learning methods. The accuracy of Deep ConvNet using CCV features reached 68.7%. This study suggested the effectiveness of Riemannian approach feature extraction and deep learning-based classification techniques in Mandarin vowel classification with spoken speech EEG signals.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61971212), the Basic Research Foundation of Shenzhen (Grant No. JCYJ20220818101217037), and Guangdong Basic and Applied Basic Research Foundation (Grant No. 2022B1515120056). Part of this study was the basis for the Bachelor's dissertation of the first author (W.Y.C.). W.Y.C. and X.Y.W. equally contributed to this paper.

References

[1] M. Rashid, N. Sulaiman, A. PP Abdul Majeed, *et al.*, "Current status, challenges, and possible solutions of EEG-based brain-computer interface: A comprehensive review," *Frontiers in Neurorobotics*, p. 25, 2020.

[2] E. M. Holz, J. Höhne, P. Staiger-Sälzer, M. Tangermann, and A. Kübler, "Brain–computer interface controlled gaming: Evaluation of usability by severely motor restricted end-users," *Artificial Intelligence in Medicine*, vol. 59, no. 2, pp. 111–120, 2013.

[3] S. Sarasso, S. Määttä, F. Ferrarelli, R. Poryazova, G. Tononi, and S. L. Small, "Plastic changes following imitation-based speech and language therapy for aphasia: A high-density sleep EEG study," *Neurorehabilitation and Neural Repair*, vol. 28, no. 2, pp. 129–138, 2014.

[4] C. Cooney, R. Folli, and D. Coyle, "Neurolinguistics research advancing development of a direct-speech brain-computer interface," *IScience*, vol. 8, pp. 103–125, 2018.

[5] S.-H. Lee, M. Lee, and S.-W. Lee, "EEG representations of spatial and temporal features in imagined speech and overt speech," *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26--29, 2019, Revised Selected Papers, Part II 5*, pp. 387–400, 2020.

[6] H. Lu, H.-L. Eng, C. Guan, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized common spatial pattern with aggregation for EEG classification in small-sample setting," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 12, pp. 2936–2946, 2010.

[7] R. Zhang, X. Xiao, Z. Liu, et al., "A new motor imagery EEG classification method FB-TRCSP+ RF based on CSP and random forest," *IEEE Access*, vol. 6, pp. 44944–44950, 2018.

[8] D.-H. Lee, S.-J. Kim, and S.-W. Lee, "DAL: Feature learning from overt speech to decode imagined speech-based EEG signals with convolutional autoencoder," *arXiv preprint arXiv:2107.07064*, 2021.

[9] P. Agarwal and S. Kumar, "Transforming imagined thoughts into speech using a covariance-based subset selection method," *Indian Journal of Pure & Applied Physics*, vol. 59, no. 3, pp. 180-183, 2021.

[10] L. A. Moctezuma and M. Molinas, "EEG-based subjects identification based on biometrics of imagined speech using EMD," *Brain Informatics: International Conference, BI 2018, Arlington, TX, USA, December 7--9, 2018, Proceedings 11*, pp. 458–467, 2018.

[11] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," *Journal of Neural Engineering*, vol. 16, no. 3, p. 031001, 2019.

[12] M.-O. Tamm, Y. Muhammad, and N. Muhammad, "Classification of vowels from imagined speech with convolutional neural networks," *Computers*, vol. 9, no. 2, p. 46, 2020.

[13] H. Altaheri, G. Muhammad, and M. Alsulaiman, "Physics-informed attention temporal convolutional network for EEG-based motor imagery classification," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 2, pp. 2249–2258, 2022.

[14] T. M. Ingolfsson, M. Hersche, X. Wang, N. Kobayashi, L. Cavigelli, and L. Benini, "EEG-TCNet: An accurate temporal convolutional network for embedded motor-imagery brain–machine interfaces," *2020 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2958–2965, 2020.

[15] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[16] S. Chengaiyan, A. S. Retnapandian, and K. Anandan, "Identification of vowels in consonant–vowel–consonant words from speech imagery based EEG signals," *Cognitive Neurodynamics*, vol. 14, no. 1, pp. 1–19, 2020.

[17] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the EEG manifold for phonological categorization from active thoughts," *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2762–2766, 2019.

[18] M. Kaya and H. S. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[19] Y.-E. Lee and S.-H. Lee, "EEG-transformer: Self-attention from transformer architecture for decoding EEG of imagined speech," *2022 10th International Winter Conference on Brain-Computer Interface*, pp. 1–4, 2022.

[20] S. J. B. Rojas, R. Ramırez-Valencia, D. Alonso-Vázquez, *et al.*, "Recognition of grammatical classes of overt speech using electrophysiological signals and machine learning," *2022 IEEE 4th International Conference on BioInspired Processing*, pp. 1–6, 2022.

[21] M. Li, S. H. Pun, and F. Chen, "A preliminary study of classifying spoken vowels with EEG signals," *International IEEE/EMBS Conference on Neural Engineering,* pp. 13–16, 2021.

[22] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech synthesis using EEG," *202 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1235–1238, 2020.

[23] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[24] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 992–996, 2015.

[25] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from EEG-a comprehensive review," *Frontiers in Neuroscience*, vol. 15, p. 392, 2021.