Mandarin Tone Classification in Spoken Speech with EEG Signals

Mingtao Li
University Key Laboratory of Advanced Wireless Communications of Guangdong Province
Department of Electrical and Electronic Engineering
Southern University of Science and Technology
Shenzhen, China
limt@mail.sustech.edu.cn

Sio Hang Pun
State Key Laboratory of Analog and Mixed-Signal VLSI
University of Macau
Macau, China
lodgepun@um.edu.mo

Fei Chen
University Key Laboratory of Advanced Wireless Communications of Guangdong Province
Department of Electrical and Electronic Engineering
Southern University of Science and Technology
Shenzhen, China
fchen@sustech.edu.cn

***Abstract*—As a naturalistic form of communication, direct-speech brain-computer interfaces (DS-BCIs) give users the possibility of 'reading the mind'. The understanding of brain processing in the spoken speech is the bridge to the ideal DS-BCI, and lexical tones as an important element in tone languages like Mandarin are desirable to be well explored. This work studied the classification of four Mandarin tones in spoken speech by using electroencephalogram (EEG). Specially, a speech pronunciation experiment was performed to include imagined, intended, and spoken states. The multiple combinations of vowels, consonants, and tones constituted monosyllables as the stimuli. Common spatial pattern (CSP) and Riemannian manifold were used as feature extraction methods, and linear discriminant analysis as classifier. Result showed that the four-class classification accuracy of the Riemannian manifold-based method across all participants was 42.9%, which was 12.3% higher than that of the CSP-based method. This work suggested that the spoken Mandarin tones were decodable with corresponding EEG signals.**

Keywords—Mandarin tones, electroencephalogram (EEG), spoken speech, common spatial pattern (CSP), Riemannian manifold

I. Introduction

Brain-computer interface (BCI) systems in recent studies often used indirect approaches (e.g., motor imagery [1], P300 and steady-state visual evoked potentials [2], etc.) to control external devices (e.g., keyboard [3], screen cursors [4], wheelchairs [5], etc.). Although the results of these methods were effective, the relatively low information transmission rate and additional user training were clear limitations. Meanwhile, the direct interaction in BCI systems attracted the attention of many studies in these years [e.g., 6-8].

Speech is considered as one of the most intuitive communication forms in daily life. Decoding the speech pronunciation process is helpful to the design of the direct-speech BCI (DS-BCI). Three types of speech used in the experiments have been categorized [9], namely as imagined, intended, and overt. The imagined speech is to imagine the pronunciation of words without any outputs. The intended speech results in the corresponding movement of articulators, but does not produce any audible output either. The overt speech is the natural speech production. The DS-BCI based on imagined speech is the ideal system. However, there are challenges for speech types without sound [10]. First of all, it is difficult to distinguish the speech parts from the non-speech parts in brain signals. Then, the temporal information of speech is absent. In addition, the lack of auditory feedback also makes studies difficult. Therefore, most relevant studies focused on overt speech.

The brain signals of phoneme pronunciation "BA/WA" or "RA/LA" had been successfully classified by Blakely et al. [11]. By using local field potential (LFP), Kellis et al. classified a small set of words at well above chance levels[12]. Speaking five Japanese monophthongal vowels was the task of Ibayashi et al.'s research [13]. Single-unit activity (SUA), LFP, and electrocorticography (ECoG) were recorded simultaneously, and their accuracies of vowel classification in each recoding technique were 37.7%, 40.7%, and 41.0%, respectively. Ramsey et al. [14] classified four spoken phonemes. The classification accuracy reached 71.9%, based on 30 repeated times for each task. These works concentrated on the stimuli like single phoneme, vowel, and word. Mugler et al. focused on the spoken phoneme classification within different words [15]. The study completed by Pei et al. decoded vowels and consonants separately from spoken monosyllables with consonant-vowel (CV) pairs [16]. The above two studies were more complex and challenging. So far, plenty of spoken speech classification studies had been performed with phonemes, vowels, consonants, etc., but not lexical tones.

For tone languages like Mandarin, tones are very important in recognizing and understanding Chinese. They carry a large amount of intelligibility information [17]. Different tones give distinct meanings to individual syllables [18]. Each character in Chinese corresponds to one syllable. Each Mandarin syllable consists of the initial sound, final sound, and tone. Consonants are initial sounds and vowels are at the end. Mandarin has four tones, usually expressed as tone 1, tone 2, tone 3, and tone 4 [19-20]. The first tone is a high, even, and constant tone. The second tone is a rising tone that grows stronger. For the third tone, it is falling and fading firstly, and then rising and growing strong. The fourth tone is a quickly falling and fading tone. A lot of works [e.g., 18, 21] explored the neuropsychological mechanisms of lexical tones. Their results showed the different lateralization in tone processing of brain activities. Hence, it is potential to decode spoken tones in syllables using brain signals.

This work aimed to investigate whether the EEG signals of four spoken Mandarin tones were decodable. Since the

This work was supported by Shenzhen Sustainable Support Program for High-level University (2020), and High-level University Fund of Southern University of Science and Technology.

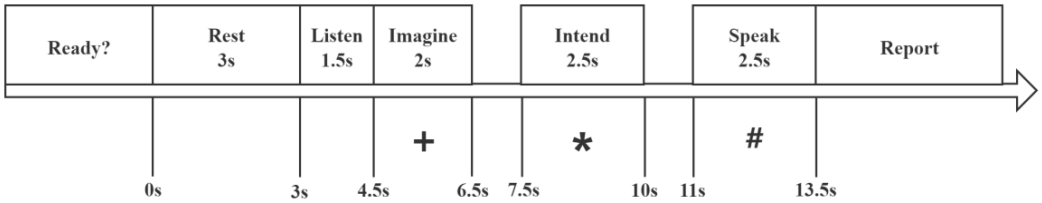

Fig. 1. The schematic of the experimental paradigm

overt speech has audible output and its exact speaking time is clear, the corresponding brain activity time can be constrained. This work recorded non-invasive EEG signals during the imagined, intended, and spoken states. The features of EEG data in the spoken state were extracted by the spatial-based common spatial pattern (CSP) method and the Riemannian manifold method, and classified by a linear discriminant classifier.

II. Methods

A. Subjects

Eleven (four females and seven males, aged from 20 to 30 years, mean age ± SD = 22.6 ± 3.2) participants were recruited from Southern University of Science and Technology. Ten participants were used in the analysis. All participants identified Mandarin as their first language and spoke Mandarin at a fluent level. All of them were in good health condition, and had no history of neurological or psychological disorders. The informed written consents were given before their participations. The experimental procedures were approved by the Institution's Ethical Review Board of Southern University of Science and Technology.

B. Stimuli and Experimental Paradigm

The stimuli consisted of seventy monosyllable words. Fifty-four of these seventy words were composed of one of four different vowels (i.e., /a/, /i/, /u/, and /ü/) with four tones and one of five consonants (i.e., /b_/, /f_/, /j_/, /l_/, /m_/). The other words were four different single vowels with four tones. These vowels are well separable in formant space and mouth shapes. The chosen consonants are the representation of plosives, nasals, fricatives, laterals, and affricates in Mandarin consonants. The number of stimuli in tone 1 is sixteen, and that in other tones is eighteen.

To avoid the confusion in pronunciation, the pseudowords whose pronunciations do not exist in Mandarin were rejected (i.e., /lü/, /mu/ in tone 1, and /bü/, /fi/, /ja/, /ju/, /mü/ in all four tones). The list of all monosyllabic words was shown in Table I. These vowels, consonants, and tones were integrated into a consonant-vowel-tone (CVT) structure. In this structure, each word was uniquely identified.

To better signal quality and reduce the fatigue of the subjects, the whole experiment was self-controlled by subject to determine the start of each trial. The schematic diagram of the experimental paradigm is shown in Fig. 1.

Each trial consisted of seven states:

- A prepared state, where the participant was instructed to prepare for this trial. The text 'Ready?' was shown on the screen. The participant pressed any button to start this trial when s/he was ready.
- A 3-second rest state. The participant needed to relax, look at the center of the black screen, restrain any movements, and clear any thoughts.
- A 1.5-second stimulus state. The auditory utterance was played about 700 ms from earphones at first. It was recorded by a female native Mandarin speaker. The remaining time was in silence.
- A 2-second imagined state, where a white cross symbol '+' appeared at the center of the screen. The participant imagined the pronunciation of the stimulus once without any articulator movements. After this state, a 1-second gap without symbol display was settled to show the completion of this state.
- A 2.5-second intended state. A star symbol '*' was shown in this state. The participant was instructed to silently utter the stimulus without producing any overt speech. The 1-second gap was also settled.
- A 2.5-second speaking state. A hash '#' symbol appeared at the center of the screen. The participant needed to speak out stimulus loudly and clearly. A microphone was used to record the spoken speech.
- A report state. The participants needed to report the completion of each state by using buttons.

The whole experiment was conducted in an acoustically and electrically shielded chamber. The stimuli and audios were recorded at a sampling rate of 16 kHz. All participants sat down in a comfortable chair and stared at the symbol on the computer screen. They performed once naturally in imagined, intended and spoken state, like what they heard in stimulus state.

Before the formal experiment began, the participant was trained to be familiar with the experimental procedure. For each participant, five blocks were expected to be completed. The break time between two blocks was about 15 minutes. In each block, all seventy trials were displayed in a pseudo-random order and not repeated. The orders were different among the five blocks.

C. Data Collection and Pre-processing

The scalp EEG signals were recorded from a 64-channel electrode cap (Neuroscan Inc.). The sampling rate was 500 Hz. The extended 10-20 system was used to place the scalp electrodes in the right place. The electrode on the forehead was regarded as the ground. The reference channel was an electrode attached to the nose tip. Two additional reference electrodes were placed at the bilateral mastoids. To measure the electrooculography (EOG) signals, two electrodes were

TABLE I. List of Monosyllabic Stimuli

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

TABLE II. CONFUSION MATRIX OF SPOKEN TONE CLASSIFICATION IN CSP METHOD

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	21.4±7.8	25.0±9.1	22.4±7.9	31.2±10.3
Tone 2	19.1±6.4	34.4±12.5	22.2±8.5	24.3±9.2
Tone 3	18.6±6.8	24.9±8.9	**32.8±11.6**	23.7±8.7
Tone 4	21.7±7.4	23.5±8.0	22.1±7.4	32.6±11.1

TABLE III. CONFUSION MATRIX OF SPOKEN TONE CLASSIFICATION IN RIEMANNIAN METHOD

Accuracy (%)	Tone 1	Tone 2	Tone 3	Tone 4
Tone 1	37.4±11.6	20.5±6.8	16.6±5.5	25.5±7.1
Tone 2	20.6±5.6	40.5±16.8	21.4±8.7	17.5±8.0
Tone 3	13.0±4.4	19.3±8.7	**48.3±10.8**	19.3±7.6
Tone 4	20.1±5.5	18.7±7.2	16.5±5.9	44.7±8.9

attached above and below the left eye. During the whole experiment, all electrode impedances were less than 5 kΩ.

EEGLAB 13.5.4b [22] was used to analyze the recorded EEG data. Firstly, the raw data were re-referenced using the channels of bilateral mastoids. Then, an FIR bandpass filter with cut off frequencies 0.5 and 70 Hz was applied. A notch filter with cut off frequencies 49 and 51 Hz was used to remove the power line noise. Independent component analysis (ICA) was implemented to remove other artifacts like electrocardiography (ECG), EOG, electromyography (EMG), etc. [23].

In this experiment, the 2500-ms EEG data of all channels in the speaking state was extracted and corrected with the baseline of 200-ms data before the speaking state start. The unusable data were deleted (e.g., the data with disordered marker, the uncompleted trials recorded in the report state, etc.). VOICEBOX toolbox in MATLAB was implemented to find the onset time. Based on the speaking time, the 2500-ms EEG data in each epoch were constrained to 800 ms, which started at the beginning of the overt speech. Most of the pronunciation time was covered among all subjects. The data of participant 6 was removed because of the massive bad blocks. The audio files in the first block of participant 11 were not saved correctly, and the relevant EEG data were also unavailable.

D. Classification

The two methods based on CSP and Riemannian manifold were used to classify the EEG signals of four spoken tones. The CSP-based method [24] is a widely used feature extraction method with many mature extensive applications. This method extracts EEG features by using spatial filters to maximize the variance of the data between classes and to minimize the variance within classes. The features with 16 dimensions are extracted.

The Riemannian manifold-based method [25-27] recently has been implemented on EEG classification tasks. In each epoch, the covariance matrices of EEG data are calculated and treated as sample points. Then they are projected to the Riemannian tangent space before fed into the classifier.

Linear discriminant analysis (LDA) is the classifier for the above two methods. For each participant, the LDA model was determined from 80% of the data as a training set and tested on the remaining test set (i.e., 20% of the data). This classification procedure was then repeated 20 times. In each time, different 20% of the data were used as the test set.

III. RESULTS

The tone classification accuracies based on the CSP and Riemannian methods are shown in the confusion matrices in Table II and III. In Table II and III, the cell with the best tone classification accuracy is marked in bold. The column labels correspond to the correct tone, and the row labels correspond to the predicted tone. The values in each cell are classification accuracies in percent and the standard deviation across subjects. The average tone classification accuracies are 30.6% and 42.9% for the CSP method and the Riemannian manifold method, respectively. Under the CSP method, the best classification accuracy among four tones is achieved with tone 2, which is 34.4%. Tone 3 in the Riemannian manifold has the highest classification accuracy (i.e., 48.3%) compared with other tones. The accuracies of tone 1 in two methods are 21.4% and 37.4%, respectively, which are the minimal in both methods. The comparisons with Bonferroni correction were run between the accuracies of two feature extraction methods. The Bonferroni-corrected statistical significance level was set at $p<0.01$ ($\alpha = 0.05$). Analysis revealed that the average classification accuracy under the Riemannian manifold method was significantly higher ($p<0.01$) than that under the CSP condition.

IV. DISCUSSION AND CONCLUSION

This study carried out a Mandarin monosyllable pronunciation experiment combining imagined, intended, and spoken states. EEG signals of four spoken Mandarin tones were specially classified. The CSP and Riemannian manifold methods were used to extract features, and the classifier was based on LDA. To the best of our knowledge, no previous studies have demonstrated that four spoken Mandarin tones could be discriminated by using cortical EEG signals. It is difficult to directly compare our results with other multiclassification works as there were no common experiment materials and experimental paradigms among them. As shown in Table IV, single vowels [13] and single phonemes [14] were classified with 10 and about 30 repeat times of each trial, respectively. A large number of repeat times of the same trial increased the accuracy rate but reduced the complexity and challenge of the classification task. In [16],

TABLE IV. PERFORMANCE COMPARISON WITH OTHER STUDIES

Ref.	Recording Techniques	Number of Classes	Repeat Times	Accuracy
[13]	SUA, LFP, ECoG	5 vowels	10	37.7%, 40.7%, 41.0%
[14]	ECoG	4 phonemes	~30	71.9%
[16]	ECoG	4 vowels, consonants in CVs	~4	40.7%, 40.6%
This study	EEG	4 tones in CVT pairs	5	42.9%

4 vowels and 9 consonants embedded in spoken monosyllabic words were decoded from ECoG signals. The design of the present study was similar to [16]. The 4 classification accuracies in [16] for decoding vowels and consonants (i.e., 4 of 9 pairs were calculated separately) across all subjects were 40.7% and 40.6%, respectively. Compared with the results from [16], the results in this study (i.e., around 42.9% of 4-tone classification accuracy in the Riemannian manifold method) look promising. Further, the experiment materials used in [16] contributed one variable in their classification. When 4 vowels were the targets, the consonants were variable and vice versa. In this study, when 4 tones were the targets, the consonants and vowels were two variables, which were more complex and made the present task more difficult for classification. Besides, compared with ECoG signals, EEG signals have a much lower signal-to-noise ratio and fewer available frequency bands for signal analysis [10], which make the multi-classification task even harder. Moreover, considering the invasive property of ECoG signals, the EEG-based work is more suitable for future speech-related BCI studies.

Over the past few years, some phonetic representations in brain activities, like phoneme and vowel but not tone, have been decoded very well [11-16]. The attention paid to tone does not match its importance in tone language. This study might be a start. Relative time-accurate brain activities were located in the speaking state. Further studies could focus more on other states, like intended and imagined states.

In conclusion, this study demonstrated that the four Mandarin tones can be decoded in spoken consonant-vowel-tone pairs using EEG signals. The classification accuracies of the Riemannian manifold method were higher than that of the traditional CSP method.

REFERENCES

[1] P. Wierzgała, D. Zapała, G. M. Wojcik, and J. Masiak, "Most popular signal processing methods in motor-imagery BCI: a review and meta-analysis," *Frontiers in Neuroinformatics*, vol. 12, pp. 78, 2018.

[2] M. Xu, J. Han, Y. Wang, and D. Ming, "Optimizing visual comfort and classification accuracy for a hybrid P300-SSVEP brain-computer interface," in *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society on Neural Engineering*, pp. 363-366, 2017.

[3] N. Birbaumer, A.R. Murguialday, and L. Cohen, "Brain–computer interface in paralysis," *Current Opinion in Neurology*, vol. 21, no. 6, pp. 634638, 2008.

[4] A. Kübler, B. Kotchoubey, T. Hinterberger, N. Ghanayim, J. Perelmouter, M. Schauer, C. Fritsch, E. Taub, and N. Birbaumer, "The thought translation device: a neurophysiological approach to communication in total motor paralysis," *Experimental Brain Research*, vol. 124, no. 2, pp. 223-232, 1999.

[5] B. Rebsamen, E. Burdet, C. Guan, H. Zhang, C. L. Teo, Q. Zeng, C. Laugier, and M.H. Ang, "Controlling a wheelchair indoors using thought," *IEEE Intelligent Systems*, vol. 22, no. 2, pp. 18-24, 2007.

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, and J. Brumberg, "Silent speech interfaces," *Speech Communication Journal*, vol. 52, no. 4, pp. 270-287, 2010.

[7] C. Herff, G. Johnson, L. Diener, J. Shih, D. Krusienski, and T. Schultz, "Towards direct speech synthesis from ECoG: a pilot study," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 1540-1543, 2016.

[8] D. Dash, P. Ferrari, and J. Wang, "Decoding imagined and spoken phrases from non-invasive neural (MEG) signals," *Frontiers in Neuroscience*, vol. 14, no. 290, 2020.

[9] 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.

[10] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Heff, and J. S. Brumberg, "Biosignal-based spoken communication: a survey," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2257-2271, 2017.

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

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

[13] K. Ibayashi, N. Kunii, T. Matsuo, Y. Ishishita, S. Shimada, K. Kawai, and N. Saito, "Decoding speech with integrated hybrid signals recorded from the human ventral motor cortex," *Frontiers in Neuroscience*, vol. 12, no. 221, 2018.

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

[15] E. M. Mugler, J. L. Patton, R. D. Flint, Z. A. Wright, S. U. Schuele, J. Rosenow, J. J. Shih, D. J. Krusienski, and M. W. Slutzky, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *Journal of Neural Engineering*, vol. 11, no. 3, pp. 035015, 2014.

[16] X. Pei, E. C. Leuthardt, C. M. Gaona, P. Brunner, J. R. Wolpaw, and G. Schalk, "Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition," *NeuroImage*, vol. 54, no. 4, pp. 2960-2972, 2011.

[17] A. D. Patel, Y. Xu, and B. Wang, "The role of F0 variation in the intelligibility of Mandarin sentences," in *Proceedings of Speech Prosody*, 2010.

[18] X. Si, W. Zhou, and B. Hong, "Cooperative cortical network for categorical processing of Chinese lexical tone," in *Proceedings of the National Academy of Sciences of the United States of America*, pp. 12303-12308, 2017.

[19] B. H. Repp, H. Lin, "Integration of segmental and tonal information in speech perception: a cross-linguistic study", *Journal of Phonetics*, vol. 18, no. 4, pp. 481-495, 1990.

[20] F. Chen, E. Y. W. Wong, "Mandarin tone identification with subsegmental cues in single vowels and isolated words", *Speech, Language and Hearing*, vol. 21, no. 3, pp. 183-189, 2017.

[21] X. Wang, M. Wang, and L. Chen, "Hemispheric lateralization for early auditory processing of lexical tones: Dependence on pitch level and pitch contour," *Neuropsychologia*, vol. 51, no. 11, pp. 2238-2244, 2013.

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

[23] G. Krishna, C. Tran, Y. Han, M. Carnahan, "Speech synthesis using EEG," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 1235-1238, 2020.

[24] M. Grosse-Wentrup and M. Buss, "Multiclass Common Spatial Patterns and Information Theoretic Feature Extraction," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 8, pp. 1991-2000, 2008.

[25] A. Barachant, S. Bonnet, M. Congedo and C. Jutten, "Multiclass Brain-Computer Interface Classification by Riemannian Geometry," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 4, pp. 920-928, 2012.

[26] A. Barachant, S. Bonnet, M. Congedo and C. Jutten, "Classification of covariance matrices using a Riemannian-based kernel for BCI applications", *NeuroComputing*, vol. 112, pp. 172-178, 2013.

[27] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," *Journal of Neural Engineering*, vol. 15, no. 1, pp. 016002, 2018.