# EEG-based Classification of Imaginary Mandarin Tones

Xinyu Zhang, Hua Li, and Fei Chen *Senior Member, IEEE*

*Abstract*— **Speech imagery based brain-computer interface (BCI) has the potential to assist patients with communication disorders to recover their speech communication abilities. Mandarin is a tonal language, and its tones play an important role in language perception and semantic understanding. This work studied the electroencephalogram (EEG) based classification of Mandarin tones based on speech imagery, and also compared the classification performance of speech imagery based BCIs at two test conditions with visual-only and combined audio-visual stimuli, respectively. Participants imagined 4 Mandarin tones at each condition. Common spatial patterns were applied to extract feature vectors, and support vector machine was used to classify different Mandarin tones from EEG data. Experimental results showed that the tonal articulation imagination task achieved a higher classification accuracy at the combined audio-visual condition (i.e., 80.1%) than at the visual-only condition (i.e., 67.7%). The results in this work supported that Mandarin tone information could be decoded from EEG data recorded in a speech imagery task, particularly under the combined audio-visual condition.**

## I. Introduction

Brain-computer interface (BCI) is one of the research hotspots in the field of biomedical engineering in recent years [1-2]. Among existing non-invasive BCI systems, electroencephalogram (EEG) is much cheaper than others and has been widely studied. Due to different brain activities, commonly-used EEG signals include event-related (de)synchronization (ERD/ERS) [3], mental task [4], steady-state evoked potentials (SSEPs) [5], P300 evoked potentials [6], etc. ERD/ERS of sensory-motor rhythms has been evoked in overt motor execution and motor imagery. Although BCI based on motor imagery has good classification accuracy, its largest classification categories are limited (e.g., 4 in [7]). The SSEPs and P300 based BCIs rely on external devices to provide stimulus input and require users to pay attention for long periods of time, making them difficult in practical use for some users, especially patients.

Compared with BCIs using the above-mentioned brain signals, the BCI based on speech imagery is a simple method that requires no special training and could fulfill multi-classification tasks [8]. Many speech imagery based BCI studies have been carried out. Leuthardt controlled BCI through electrocorticography (ECoG) speech network [9], and DaSalla et al. realized the classification by imagining the pronunciation and corresponding mouth shapes of vowels /a/ and /u/ [10]. Matsumoto et al. studied speech imagery to classify two of 5 Japanese vowels (i.e., /a/, /i/, /u/, /e/, and /o/), and achieved good classification accuracy (i.e., 77%) [8]. Yang et al. realized the dichotomy of different vowels and consonants in Mandarin by imagining their pronunciations, and their results supported the feasibility of speech imagery based on EEG.

Feature extraction and feature classification are important components of the BCI technology based on speech imagery. Since the critical information is contained in different frequency bands of EEG signals, EEG signals can be separated into different specific frequency bands for analysis [11]. Among the currently studied feature extraction and feature classification algorithms, common spatial patterns (CSPs) are the feature extraction method with the best performance and the most extensive applications [e.g., 8, 10, 12]. CSP converts the recorded EEG data to a new space, maximizing the variance of the data between classes and minimizing the variance within classes. Support vector machine (SVM) was widely used to classify the EEG data of speech imagery, and better classification results were obtained in early studies [e.g., 8, 10, 13]. Meanwhile, neural network is a common classification method for multi-classification problems, but it has not been widely applied to EEG signals based on speech imagery until recently. Saha et al. introduced deep learning models to categorize vowels and words based on speech imagery and achieved good classification results (i.e., 73.5%) [14].

Compared with English, Mandarin has fewer single finals and simple syllable structure. Generally, Mandarin syllables are composed of initial, final and tone. In term of Mandarin tones, tone information attached to finals has the function of distinguishing sound meaning. Therefore, the recognition and reconstruction of Mandarin tones are important to improve the intelligibility of synthesized Mandarin speech [15-16]. In Mandarin Chinese, there are 4 lexical tones which are the flat tone, the rising tone, the falling-rising tone, and the falling tone, respectively (or called tone-1, tone-2, tone-3, and tone-4) [15-17]. Errors in the pronunciation of these lexical tones can lead to semantic and syntactic errors. For example, four tones of Mandarin syllable /ba/ can be read as /bā/, /bá/, /bǎ/, /bà/ and respectively mean ‘eight’, ‘pull’, ‘target’, ‘father’. Thus, studying Mandarin tone classification based on speech imagery may provide an important cue for the future development of Mandarin speech imagery based BCI system.

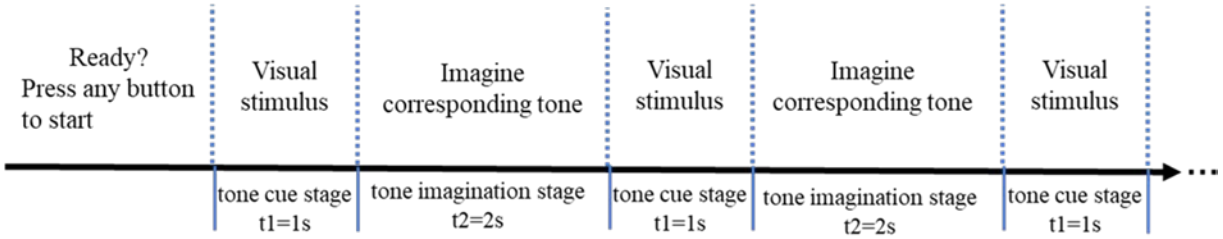
The aim of this work was to verify the feasibility of the Mandarin tone classification task based on speech imagery. The Mandarin tones were classified by SVM with Gaussian kernel. In addition, since most of the current speech imagery experiments used visual stimuli, this work further assessed the influence of different types of stimuli on EEG-based classification of imaginary Mandarin tone. The classification accuracies of visual-only stimuli and combined audio-visual stimuli were compared.

This work was supported by the National Natural Science Foundation of China (Grant No. 61971212). This work was done while Hua Li visited Dr. Fei Chen’s laboratory as a visiting student.

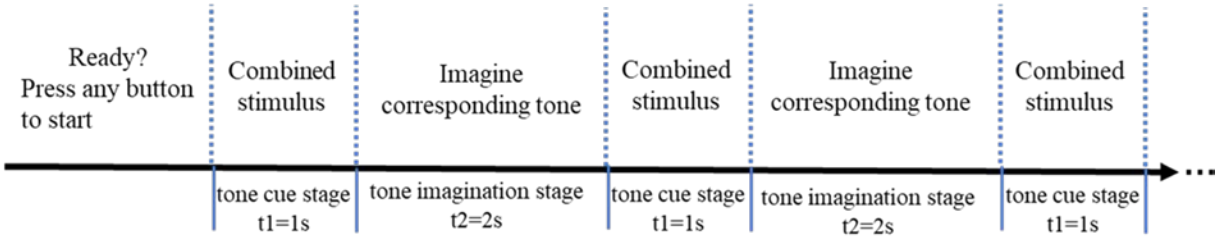
Xinyu Zhang and Hua Li are with Shenzhen University Health Science Center School of Biomedical Engineering, Shenzhen, China.

Fei Chen is with Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China (phone: +86-75588018554; fax: +86-75588018554; e-mail: fchen@sustech.edu.cn).


Figure 1. Visual stimuli used in this experiment, including Chinese syllable /ba/ in 4 lexical tones and the sign (i.e., '+') that prompts the subjects to imagine the tones.




(a) The experimental process of test condition 1.




(b) The experimental process of test condition 2

Figure 2. Experimental paradigms for (a) test condition 1 and (b) test condition 2.

## II. Methods

### A. Participants

Fourteen native Mandarin-Chinese speakers from 19 to 26 years of age (6 males and 8 females) participated in the experiment. All participants were students at Southern University of Science and Technology, and were in good general health condition with no report of neurological or psychological problems. All participants signed informed consent forms and were paid for their participations. The experimental procedure involving human subjects was reviewed and approved by the Research Ethics Committee of the Southern University of Science and Technology.

### B. Stimuli and experimental paradigm

Two types of test conditions were designed in this experiment, i.e., visual-only stimuli at test condition 1 and combined audio-visual stimuli at test condition 2. The visual stimuli at condition 1 were presented in the form of pictures, as shown in Fig. 1. All symbols/signs in pictures were presented at the same size to control the interference of image transformation. For the combined audio-visual stimuli at condition 2, the visual stimuli were the same as those used at condition 1, while the auditory stimuli were syllable /ba/ in four Mandarin tones pronounced by one adult female native Mandarin-Chinese speaker. The Mandarin syllables were recorded at a sampling rate of 16 kHz, and their durations were normalized to one second.

The experimental paradigms for test conditions 1 and 2 are shown in Fig. 2. Each condition had four test blocks (the order of stimuli presentation was pseudo-random), and participants imagined 40 Mandarin tones in each block. Each speech imagery trial consisted of a tone cue stage and a tone imagination stage. When the participant was ready, s/he was prompted to press any key to start the experiment. On the tone cue stage, the visual-only stimuli (or combined audio-visual stimuli) appeared for 1 s. Then on the tone imagination stage, a prompt sign appeared to inform the participant to imagine the corresponding tone. The total duration of one test trial was around 8 mins. All stimuli were presented through a Sennheiser HD 25 earphone. During the whole experiment, participants sat in an acoustically and electrically shielded chamber. They were seated comfortably and instructed to pay attention to visual-only stimuli (or combined audio-visual stimuli).

### C. Procedure and EEG data recording

EEG data were obtained with a 64-channel electrode cap (Neuroscience Inc.). Through the extended international 10-20 system, the cap was placed at specific positions. The top of the nose served as a reference for all electrodes, and the ground electrode was attached to the forehead. The impedance between the reference electrode and any

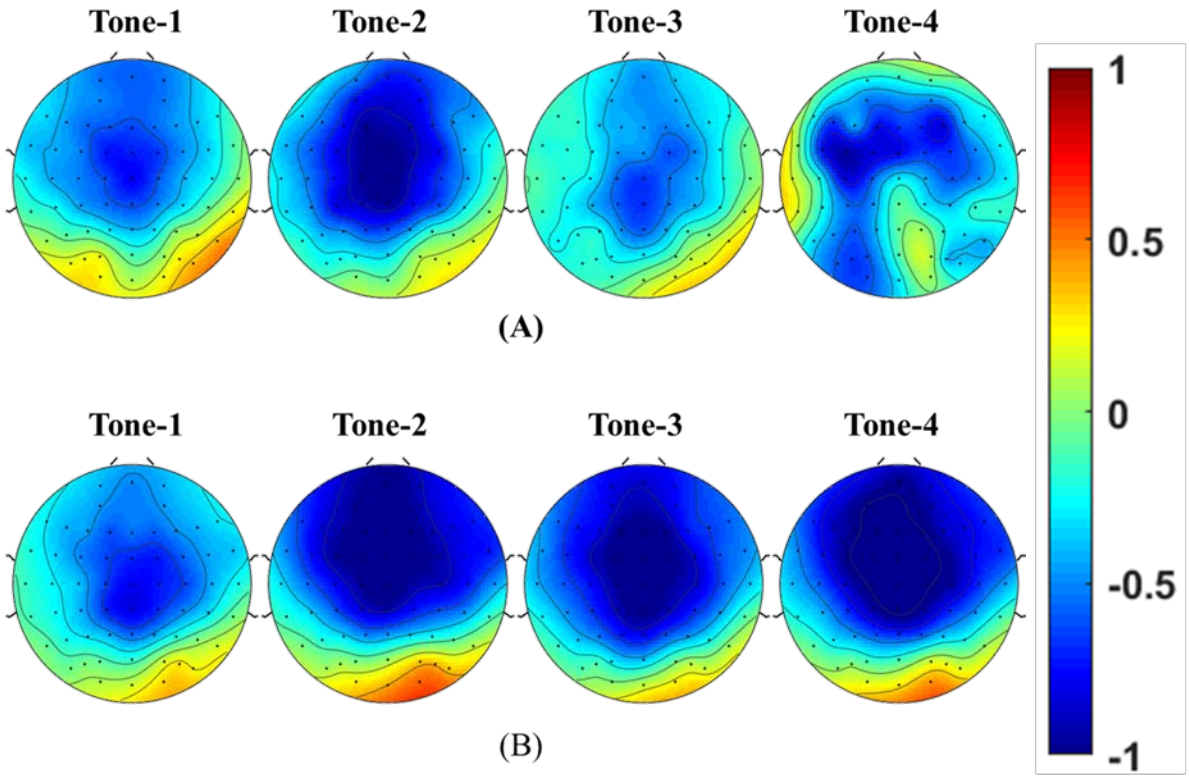



Figure 3. Average contour maps of four tones at 2 stimulus conditions. (A): Average contour maps under visual-only stimuli; (B): Average contour maps under combined audio-visual stimuli.

recording electrode was less than 5 kΩ. The sampling rate of EEG signal was 500 Hz. The participants were asked to minimize the body movements to avoid motion artifacts.

EEG data were analyzed with EEGLAB 14.1.1. First, the data were re-referenced using the contralateral mastoid signals. Epochs containing artifacts exceeding ±75 μV were excluded from the averaging procedure. Next, the infomax independent component analysis was performed for artifact (e.g., eye blinks, horizontal eye movement, electrocardiographic activity) correction. After artifact removal, each epoch was selected between 100 ms pre-stimulus and 800 ms post-stimulus and corrected with the baseline of the pre-stimulus time window.

### *D. Feature extraction and tone classification*

The pre-processed EEG data were analyzed to select the optimal time range/frequency band to extract features. Firstly, fast Fourier transform was used to analyze the frequency domain characteristics of EEG data, and the two-dimensional event-related spectral perturbation (ERSP) was used to represent the average variation of EEG spectral energy relative to the baseline. A 256 ms sliding window was used for 200 times, with the output latency range of 100-750 ms and the frequency range of 1-30 Hz. Then, because tonal articulation imagination tasks were mainly involved Broca's area, Wernicke's area and sensorimotor cortex, the 6 electrodes involved in these areas (i.e., FC3, F5, CP3, P5, C3, and C4) were selected to calculate the average event-related potential waveform of all subjects, whose latency ranged from 200 to 500 ms. Due to the image transformation from the tone cue period to the tone imagination period, the EEG characteristics in the frequency range of 1-30 Hz and time range of 250-500 ms were finally extracted.

TABLE 1. Descriptive statistics of the classification accuracy of 4 Mandarin tones.

| Stimuli | Classification accuracy (%) (average ± standard deviant) |
|---|---|
| Visual-only | 67.7 ± 1.0 |
| Combined audio-visual | 80.1 ± 1.2 |

After extracting the characteristics of the optimal time range/frequency band, CSP was used to extract the EEG features, and the classification was carried out by using autoregressive SVM. CSP was widely used to improve the signal-to-noise ratio of EEG data. The CSP algorithm was originally designed for the two-classification BCI system, and was later extended to the multi-classification model and further improved to provide satisfactory experimental results [18-20]. SVM is a binary classifier; and for multi-classification tasks, SVM could also be implemented by designing multiple classifiers [21]. For the four-class classification task in this study, one classifier was trained for any pair of 2 classes chosen from the four classes, and the total number of two-class classifiers was six (i.e., tone-1 vs. tone-2, tone-1 vs. tone-3, tone-1 vs. tone-4, tone-2 vs. tone-3, tone-2 vs. tone-4, and tone-3 vs. tone-4). For EEG data to be classified, they needed to be predicted by all classifiers, and a

vote was used to determine the final class attribute of the Mandarin tone.

## III. Results

The average classification accuracies of 4 Mandarin tones are shown in Table 1, i.e., 67.7% with visual-only stimuli and 80.1% with combined audio-visual stimuli. There is a significant difference in classification accuracy between the two stimulus conditions ($F(1, 26) = 57.10$, $p<0.01$).

In order to explore the influence of stimulus conditions on phonologically imagined tone recognition, the average brain topographies of different tones are shown in Fig. 3. The time region when the first peak occurred within the time range of 250-500 ms was selected, and then the average EEG ERP of that region (i.e., 250-300 ms) was calculated. It can be seen from Fig. 3 that the sensorimotor cortex is significantly activated. In addition, the combined audio-visual stimuli activate a wider range of brain regions than the visual-only stimuli.

## IV. Discussion and Conclusion

In order to improve the classification accuracy of imaginary speech, it is necessary to consider the influence of different stimulating conditions on the experiment. Koheia et al. showed that the primary motor cortex was more excited during motor imagery in combined condition (visual and auditory) than in visual and auditory conditions alone [22]. Besle et al. studied the effect of visual stimuli on speech perception through electrophysiology, and the results showed that speech-related visual stimuli would have an impact on speech perception [23]. The results of high level of brain activation under audio-visual combined stimuli in Fig. 3 were consistent with early findings [e.g., 21-22]. Meanwhile, the higher classification accuracy of imaginary Mandarin tones under the combined audio-visual condition also indicated that the combined audio-visual stimuli were more favorable to the classification of imaginary Mandarin tones.

In conclusion, the present work examined the performance of EEG-based classification of imaginary Mandarin tones. Results showed that EEG data activated by the visual-only or combined audio-visual stimuli contained useful information for decoding Mandarin tones, which are important for Mandarin speech understanding and for the performance of speech imagery based BCIs. In addition, the combined audio-visual stimuli were more beneficial to the classification of imaginary Mandarin tones than the visual-only stimuli.

## V. References


[1] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, "Brain-computer interfaces for communication and control," Clinical Neurophysiology, vol. 113, no. 6, pp. 767-791, 2002.

[2] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, "Brain-computer interface technology: A review of the first international meeting," IEEE Trans Rehabil Eng., vol. 8, no. 2, pp. 164-173,2000.

[3] G. Pfurtscheller, C. Brunner, A. Schlögl, F. H. Lopes da Silva, "Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks," NeuroImage, vol. 31, no. 1, pp. 153-159, 2006.

[4] F. Faradji, R. K. Ward, G. E. Birch, "Plausibility assessment of a 2-state self-paced mental task-based BCI using the no-control performance analysis," Journal of Neuroscience Methods, vol. 180, no. 2, pp. 330-339, 2009.

[5] P. Lee, C. Yeh, J. Y. Cheng, C. Yang, G. Lan, "An SSVEP-based BCI using high duty-cycle visual flicker," IEEE Trans Biomed Eng., vol. 58, no. 12, pp. 3350-3359, 2011.

[6] M. Salvaris, C. Cinel, L. Citi, R. Poli, "Novel protocols for P300-based brain–computer interfaces," IEEE Trans Neural Syst Rehabil Eng., vol. 20, no. 1, pp. 8-17, 2012.

[7] M. Naeem, C. Brunner, R. Leeb, B. Graimann, G. Pfurtscheller, "Seperability of four-class motor imagery data using independent components analysis," Journal of Neural Engineering, vol. 3, no. 3, pp. 208-216, 2006.

[8] E.C. Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, G. Schalk, "Using the electrocorticographic speech network to control a brain–computer interface in humans," Journal of Neural Engineering, vol. 8, no. 1, pp. 1-11, 2011.

[9] M. Matsumoto, J. Horib, "Classification of silent speech using support vector machine and relevance vector machine," Applied Soft Computing, vol. 20, no. 1, pp. 95-102, 2014.

[10] C. S. DaSalla, H. Kambara, M. Sato, Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," Neural Network, vol. 22, no. 9, pp. 1334-1339, 2009.

[11] B. Yang, H. Li, Q. Wang, Y. Zhang, "Subject-based feature extraction by using fisher WPD-CSP in brain computer interfaces," Computer Methods and Programs in Biomedicine, vol. 129, no. 1, pp. 21-28, 2016.

[12] K. P. Thomas, G. Cuntai, C. T. Lau, A. P. Vinod, K. A. Kai, "A new discriminative common spatial pattern method for motor imagery brain computer interfaces," IEEE Trans Biomed Eng., vol. 56, no. 11, pp. 2730-2733, 2009.

[13] L. Wang, X. Zhang, X. F. Zhong, Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," Biomedical Signal Processing and Control, vol. 8, no. 6, pp. 901-908, 2013.

[14] P. Saha, A-M. Muhammad, S. Fels, "Speak your mind! Towards imagined speech recognition with hierarchical deep learning," in Proc. of 20th Annual Conference of the International Speech Communication Association (InterSpeech), Graz, 2019, pp. 141-145.

[15] W. S. Y. Wang, "The Chinese language," Scientific American, vol. 228, no. 2, pp. 50-60, 1973.

[16] B. H. Repp and H-B. 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.

[17] 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.

[18] G. Dornhege, B. Blankertz, G. Curio, "Boosting bit rates in noninvasive EEG single-trial classifications by feature combination and multiclass paradigms," IEEE Trans Biomedical Eng., vol. 51, no. 6, pp. 993-1002, 2004.

[19] H. Ramoser, J. Muller-Gerking, G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," IEEE Trans Rehabil Eng., vol. 8, no. 4, pp. 441-446, 2000.

[20] M. Matsumoto, J. Hori, "Classification of silent speech using support vector machine and relevance vector machine," Applied Soft Computing, vol. 20, pp. 95-102, 2014.

[21] M. Grosse-Wentrup, M. Buss, "Multiclass common spatial patterns and information theoretic feature extraction," IEEE Trans Biomed Eng., vol. 55, no. 8, pp. 1991-2000, 2008.

[22] I. Koheia, H. Toshiob, S. Kenichic, T. Kounosukec, K. Hiroshid, K. Tatsuya, "The effect of visual and auditory enhancements on excitability of the primary motor cortex during motor imagery: a pilot study," International Journal of Rehabilitation Research, vol. 35, no. 1, pp. 82-84, 2012.

[23] J. Besle, O. Bertrand, M-H. Giard, "Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex," Hearing Research, vol. 258, no. 2, pp. 143-151, 2009.