Recognizing Tonal and Nontonal Mandarin Sentences for EEG-Based Brain–Computer Interface

Shiau-Ru Yang, Tzyy-Ping Jung, *Fellow, IEEE*, Chin-Teng Lin, *Fellow, IEEE*, Kuan-Chih Huang, Chun-Shu Wei, Herming Chiueh, *Member, IEEE*, Yue-Loong Hsin, Guan-Ting Liou, and Li-Chun Wang, *Fellow, IEEE*

***Abstract*—Most current research has focused on nontonal languages such as English. However, more than 60% of the world's population speaks tonal languages. Mandarin is the most spoken tonal languages in the world. Interestingly, the use of tone in tonal languages may represent different meanings of words and reflect feelings, which is very different from nontonal languages. The objective of this study is to determine whether a spoken Mandarin sentence with or without tone can be distinguished by analyzing electroencephalographic (EEG) signals. We first constructed a new Brain Research Center Speech (BRCSpeech) database to recognize Mandarin. The EEG data of 14 participants were recorded, while they articulated preselected sentences. To the best of our knowledge, this is the first study to apply the method of asymmetric feature extraction method for speech recognition using EEG signals. This study shows that the feature extraction method of rational asymmetry (RASM) can achieve the best accuracy in the classification of cross-subjects. In addition, our proposed binomial variable algorithm methodology can achieve 98.82% accuracy in cross-subject classification. Furthermore, we demonstrate that the use of eight channels [(F7, F8), (C5, C6), (P5, P6), and (O1, O2)] can achieve an accurate of 94.44%. This study explores the neurophysiological correlation of Mandarin pronunciation, which can help develop a tonal language synthesis system based on BCI in the future.**

***Index Terms*—Brain–computer interface (BCI), electroencephalography (EEG), feature extraction, lexical tone, machine learning, Mandarin, speech recognition.**

Manuscript received 1 December 2020; revised 5 April 2021 and 17 June 2021; accepted 29 November 2021. Date of publication 21 December 2021; date of current version 9 December 2022. This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 110-2221-E-A49-130-MY2, Grant MOST 109-2221-E-009-050-MY2, Grant MOST 110-2634-F-009-021-MY2, and Grant MOST 110-2221-E-A49-039-MY3; and in part by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. *(Corresponding authors: Li-Chun Wang; Tzyy-Ping Jung.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Board (IRB) of the National Chiao Tung University (NCTU) under Application No. NCTU-REC-108-127E.

Shiau-Ru Yang, Kuan-Chih Huang, Chun-Shu Wei, Herming Chiueh, Guan-Ting Liou, and Li-Chun Wang are with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu 30010, Taiwan (e-mail: lichun@nctu.edu.tw).

Tzyy-Ping Jung is with the Swartz Center for Computational Neuroscience, University of California at San Diego, San Diego, CA 92093 USA (e-mail: tpjung@ucsd.edu).

Chin-Teng Lin is with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW 2007, Australia.

Yue-Loong Hsin is with Neurology Department, Chung-Shan Medical University Hospital, Taichung 40201, Taiwan.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCDS.2021.3137251.

Digital Object Identifier 10.1109/TCDS.2021.3137251

I. Introduction

SPOKEN language is one of the most common forms of communication between people. However, for patients with locked-in syndrome (LIS), such as severe spastic quadriplegic cerebral palsy, stroke, and advanced amyotrophic lateral sclerosis, most of their voluntary muscles are paralyzed except for vertical eye movement communication or blink. Even if these patients are conscious, they cannot communicate through language. This may lead to undesirable long-term consequences, including reduced quality of life, reduced social interaction, and increased burden of caregivers [1]. Although patients with communication difficulties can benefit from long-term support and speech therapy, the long-term care needs of this population today. Despite motor abnormalities such as LIS or quadriplegia, their brains still function well.

With advances in sensor technologies, it is now possible to develop intelligent applications to manage, control, and automate our living environments without human intervention. The Internet of Things (IoT) is a good example of automation science and information technology. Many studies have combined the IoT and intelligent medical or rehabilitation systems with brain–computer interfaces (BCIs) [2]–[6]. In recent decades, BCI technologies have made many advances [7]–[14]. BCI is considered to be a new communication platform that utilizes the dynamics of the user's brain. Recognizing speech through neural signals has been an emerging research area in the past few years. Previous researches usually required the performance of hand motor imagery or some other conversation-irrelevant task [15], [16]. However, those BCI methods are all nonintuitive. Various methods of recording brain activity can be used as the basis for direct speech synthesis in brain–computer communication. Electroencephalography (EEG) is widely used in the field of BCI due to its high temporal resolution and low cost. Electrocorticography (ECoG) can provide more information about brain signals but requires invasive implantation of subdural electrodes [17].

Many speech perception studies focused on nontonal language (e.g., English or German). In fact, the neural evidence of lexical tone processing is scarce. More than 60% of the languages in the world are tone languages, and the words in them are distinguished by tonal features [18]. The "tones" of a word can represent different meanings [19], [20]. Mandarin is one of the most spoken languages in the world (about 1.1 billion people). Mandarin consists of many homophones, and

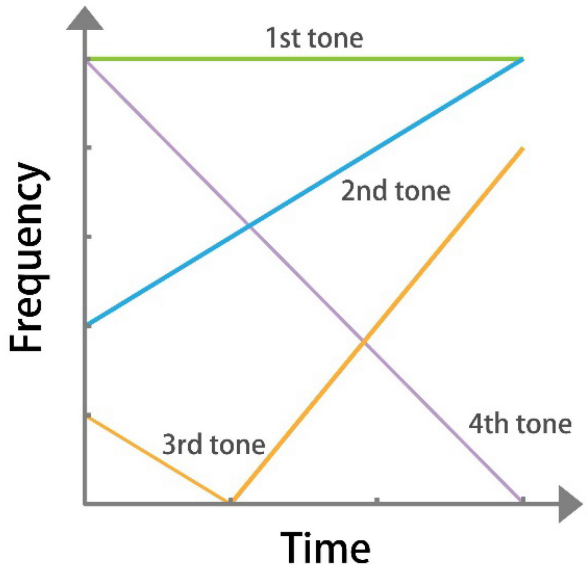

Fig. 1. Pitch frequency of the four tones in Mandarin.

it uses five tones, which is very different from nontonal language (see Fig. 1) [21]. In Mandarin, the meaning of words cannot be determined without tonal information. For example, the syllable /ma/ can be accented with four lexical tones (i.e., Tone 1—flat-level tone; Tone 2—mid-rising tone; Tone 3—mid-falling-rising tone; and Tone 4—high-falling tone) to represent four distinct meanings: mother "媽" hemp "麻" horse "馬" and curse "罵", respectively.

To the best of our knowledge, fewer BCI studies focused on BCIs that can translate brain signals into Mandarin than those BCIs for English [22]–[27]. The ultimate goal of this work is to develop a direct BCI for Mandarin. To achieve this goal we need to answer the following two questions first.

1) What is the difference in brain activity between speaking Mandarin with and without tones? (to find out the tone feature)?
2) What is the cognitive process in the brain when speaking a tonal language like Mandarin?

Therefore, this study proposes to investigate the functional difference in the brain while speaking in tonal and nontonal Mandarin. Native Mandarin speakers were asked to participate in two experiments: one is to speak normally, and the other is to speak flat-tone Mandarin. Then, we analyze the difference in EEG activities between the two. This experimental design is mainly to avoid observing the phenomenon caused by the second language [28]–[31]. Therefore, this study aims to use the same language to understand that the observed phenomena are caused by tones. One can use either a bottom-up or a top-down approach to decide neural signals during speech production. The bottom-up approach maps the basic language units [32], [33] (e.g., phonemes or syllables) onto articulation areas (e.g., motor cortex and premotor cortex). This study uses the top-down approach to decode Mandarin. We first map speech to sentence level and then corresponds to brain signals.

The purpose of this study is to investigate whether Mandarin spoken with and without tone can be distinguished based on the subject's EEG. The previous study [59] demonstrated that the left hemisphere is relevant for generating grammatical sentences and syntax rules, and the right hemisphere is key to participate in adding the emotional intonation to speech. However, the process of tone in the human brain is still not clear. Because of the functional hemispheric asymmetries, this study proposed an asymmetric feature extraction method to obtain the important features for tone. Also, this study proposed the binomial variable algorithm (BVA) to easily extract the significant features cross-subjects.

The remaining parts of this article are organized as follows. Section II introduces the related work. Section III discussed the experiment design method. Section IV presents our research method. Section V shows the evaluation results among the single and cross-subjects. Section VI discussed our numerical results. Section VII concludes this study.

II. Related Study

Recently, different neuroimaging modalities, such as functional magnetic resonance (fMRI), ECoG, EEG, etc., have been used to measure neural activities for decoding speech.

1) fMRI measures brain activities by detecting changes related to blood flow. This technique relies on the coupling of cerebral blood flow and neuron activation. Several studies have used fMRI to decode the spatial correction of speech [34]–[36]. However, the temporal resolution of fMRI (including that of the latest high-field fMRI) is limited to a few seconds, whereas the human speech articulation process takes less than a quarter of a second. However, human speech articulation involves the cooperation of different functional cortices, which originates from the mind and is manifested by the sensorimotor areas. The temporal correlation of different cortical locations and its relation to speech articulation cannot be decoded by fMRI with the low time resolution.
2) ECoG directly measures the electrical activities on the cortical surface. It has been used for accurate preoperative localization of epileptic seizures and provides high-density neural recordings. Recent studies have shown that ECoG can decode speech, including the ability to map speech evoked sensorimotor activations [37]; generate neural encoding mode of perceived phonemes [32], words [38], and sentences [22]; reconstruct acoustic properties of perceived [39]; generate natural-sounding synthetic speech from brain activity [24]; and immediately identify volunteers' spoken responses to a set of standard questions based solely on their brain activities [25]. Although recent studies have reported impressive progress in using neural signals for speech decoding, the complex dynamics, especially for Mandarin speakers, have yet to be fully elucidated.
3) EEG uses electrodes placed on the scalp to measure the electric potential of a large ensemble of simultaneously firing neurons. It is the most commonly used method for recording neural signals and has a huge advantage that it is noninvasive. EEG is widely used in BCI research because of its high temporal resolution and low cost [7]. It is easy to access, which helps the development of

a BCI-based system for language generation. Various studies have used EEG to convert vocal speech to imaginary speech of the English vowels [40], syllables [41], [42], and "yes" and "no" [43]. Some EEG-based BCIs have used deep-learning-based automatic recognition for English words [44], vowels [44], and vocabulary [45].

III. Experiment Design

A. Subjects

Fourteen healthy subjects aged 20–26 years (average age: 23.50 ± 1.99 years) were recruited to participate in this study. All subjects were native Mandarin speakers, right-handed and without neurological and mental illness, and no drug or alcohol abuse. The experiment was performed in accordance with the country's laws and approved by the Institutional Review Board (IRB) of the National Chiao Tung University (NCTU). Each participant provided written informed consent prior to participation. The participants were compensated approximately U.S. $25 after the experiments. The experimental protocol was approved by the IRB and assigned the number NCTU-REC-108-127E.

B. Experimental Paradigm

This study uses a "Focus Group Interview" [46] method to create a new Brain Research Center Speech (BRCSpeech) Database to analyze spoken Mandarin. Focus Group Interview is a method of collective discussion of specific research issues. During the interview, the interviewees are stimulated to construct ideas [46]. The study aimed to determine the differences in EEG activity when subjects spoke preselected sentences with and without tone in the BRCSpeech Database.

The BRCSpeech Database is a Mandarin-sentence database that included almost all Mandarin pronunciation characteristics. According to the research by Sagey [47], Ladefoged and Halle [48], and Longtin *et al.* [49], Mandarin involves six articulators: 1) labial; 2) coronal; 3) dorsal; 4) soft palate; 5) tongue-root; and 6) vocal cords. There are five tones in Mandarin, which are different from English: 1) Tone 1 (flat-level tone, with "–" symbol); 2) Tone 2 (mid-rising tone, with "/" symbol); 3) Tone 3 (mid-falling-rising tone, with "v" symbol); and 4) Tone 4 (high-falling tone, with "\" symbol). (Fig. 1) [21].

We adopt the contract method proposed by Duanmu's language experts [33]–[35]. The "Contract" refers to two words that sound different, that is, two words with different phonetic forms. We use two contracts. One is the tone contract. For example, tones 1 and 4 are the maximal pitch contract of tones. The other is the contract of articulators. For example, Labial and Dorsal are different articulators. In linguistics, these are the differences in speech. We assume that the differences will also be reflected in brain activities. The BRCSpeech Database we created also referenced the Texas Instruments/Massachusetts Institute of Technology (TIMIT)'s database [24], [50]. Because people's normal speakings mix both long and short sentences, the design of the 460 sentences in the TIMIT's database includes 3–12 words for each sentence. And our BRCSpeech Database is also composed of 3–12 words of each sentence.

Fig. 2. Procedure and experimental design. The sentences, each consisting of 3–12 words, were shown on a computer monitor. Each session of the experiment includes 191 sentences.

The BRCSpeech Database collected sentences composed of all Mandarin pronunciation, covering all combinations of Mandarin vowels and consonants. In this study, the BRCSpeech Database will be selected as the source of the sentences while speaking Mandarin in the tonal and nontonal experiments and the BRCSpeech Database contains combinations of various Mandarin characters' sounds, which can rich the data collection. In this study, the important features of tones are identified. In our future study, we will analyze the four types of Mandarin tones based on the results of this study.

In our EEG speech experiment, sentences were shown on the computer monitor. Each sentence is composed of 3–12 words, which were randomly selected from the BRCSpeech Database, and the duration of the sentence displayed on the monitor was adjusted according to the length of the sentence. The baseline between two consecutive trials is a white screen that is 2 s long. Each session of the experiment lasts 25 min (about 185–191 sentences), as shown in Fig. 2. In order to avoid the differences in degree of cognitive control with and without tonal information in Mandarin, the experimental design of this study is divided into two different sections. One section required the subject to speak normally (involving tone),

Fig. 3. Experimental system flowchart.

and the other required the subject to speak a flat tone like a robot without changing the tone (only Tone 1 was permitted). Each subject was asked to complete two different sections in random order.

Before recording data, the subjects would practice at least 16 trials in order to reduce the unexpected phenomenon caused by cognitive control with and without tone in Mandarin.

C. EEG Data Acquisition

This study used the SynAmps system (Australia Compumedics Ltd.) to record the EEG data, which has 64 unipolar sintered Ag/AgCl EEG electrodes placed on the scalp according to the international 10–20 system and referred to the linked mastoids (average of channel A1 and channel A2). Fig. 4 shows the layout of EEG electrodes on the cap. The impedance of all electrodes was kept below 5 kΩ. The EEG data were sampled at 1000 Hz with a 32-bit quantization. The spoken sentences were recorded with a microphone. (Sampling rate: 44.1 kHz/16 bit; Dimensions: 325-mm circumference).

D. Speech Phone Labeling

To avoid unwanted noise (other than the main physiological signals) [51] and simplify the experiment to obtain better results, we designed an experiment to adapt to the random speaking speed of the subjects. In addition, the subject did not need to perform other activities (i.e., press buttons) other than speaking [24]. Subjects were asked to speak the sentence shown on the monitor at their regular pace immediately after watching the display. This study used a high-quality microphone to record the speech and synchronized the audio recording with the subjects' EEG. Speech was synchronized by using Presentation (©2020 Neurobehavioral Systems, Inc.).

IV. Research Methods

In this section, we detail the processes applied to the EEG data speech recognition.

A. EEG Data Preprocessing

EEG signals were first filtered to 0.5–180 Hz, and then downsampled to 500 Hz for data compression. We used MATLAB R2019b (The Mathworks, Inc.), Python, and the open-source EEGLAB toolbox (http://sccn.ucsd.edu/eeglab) [52]. As shown in Fig. 3, by using the EEGLAB visualization tool, EEG signals containing electrode noise and a large number of muscle artifacts can be identified and simply removed to improve the signal-to-noise ratio.

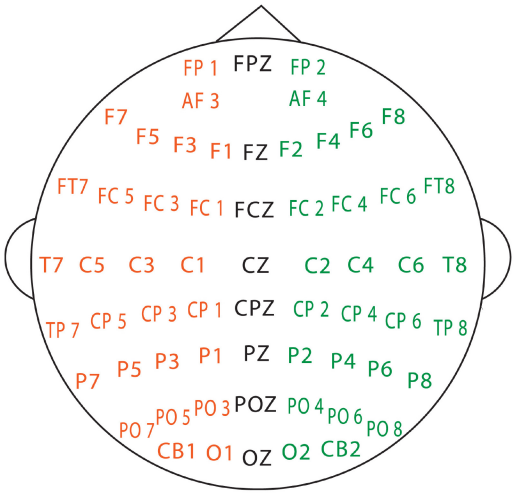

Fig. 4. Sixty-two channels layout of the EEG cap.

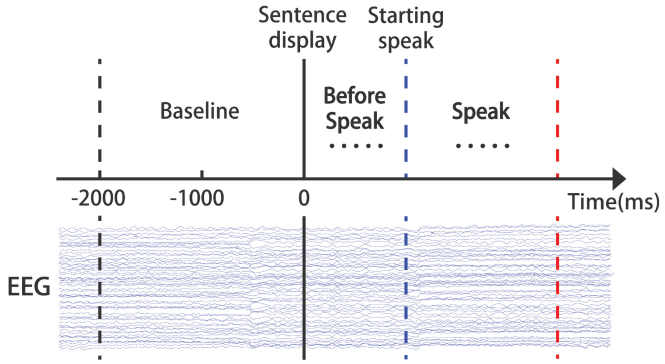

Fig. 5. Timing of a trial of the paradigm. EEG signals from baseline to end of speaking. This study analyzed the brain activities in three periods: 1) *BS*: 1 s before the sentence is displayed (the 2nd second of the baseline), 2) *Before Speak*: after the sentence is displayed and before the subject articulates the sentence, and 3) *Speak:* when the subject is articulating the sentence.

B. Feature Extraction

This study analyzed the brain activities in three periods: 1) *BS*: 1 s before the sentence is displayed (the 2nd second of the baseline); 2) *Before Speak*: after the sentence was displayed, before the subject articulates the sentence; and 3) *Speak:* when the subject is articulating the sentence (see Fig. 5). Many previous studies have shown that emotions cause differences in brain activities during rest [53], [54]. To exclude the influence of emotions, we removed the *BS* state data in the *Before Speak* and *Speak* states to ensure that the classification results were based on the tonal and nontonal language classification, while excluding the impact of the baseline emotion. In Fig. 5, the power in the range of 0.5–170 Hz of each time bin in *BS* was averaged at time bins as the average baseline power of 0.5–170 Hz. In addition, the average baseline power was subtracted from the power spectrum at each time bin of Before Speak and Speak states.

Zheng *et al.* [55] found that the following different features and electrode combinations are effective for EEG-based emotion recognition: 1) power spectral density (PSD); 2) differential entropy (DE); 3) differential asymmetry (DASM); and 4) rational asymmetry (RASM) features from EEG. As a

result, we used these features in this study. Further, this study also uses additional asymmetric (AASM) for feature extraction. The length of the window size used in this study was 0.5 s, which was based on the average time of each word (about 0.4–0.6 s in this study).

1) *Power Spectral Density (PSD):* The EEG signals of each trial for all 62 channels were first transformed into time–frequency domain to get EEG PSD using the short-time FFT. The six different frequency band power, delta–theta (0.5–7 Hz), alpha (8–12 Hz), beta (13–30 Hz), gamma (30–60 Hz), high gamma (60–170 Hz), and all bands (0.5–170 Hz), was selected and averaged as the feature of each time bins. This procedure resulted in 372 features (six frequency bands by 62 channels) for each trial (sentence).
2) *Differential Entropy (DE):* Shi *et al.* [56] found that EEG signals are subject to Gaussian distribution in a few subbands after band-pass filtering from 2 to 44 Hz. As such, the DE (denoted by $h(X)$) of the EEG signals (denoted by X) in the frequency band i can be derived by substituting the probability density function of a Gaussian random variable X into

$$h(X) = -\int f_X(x)\ \log(f_X(x))dx. \tag{1}$$

Then, we can obtain

$$h(X) = \frac{1}{2}\log\left(2\pi e\sigma^2\right) \tag{2}$$

where $f_X(x) = [1/(\sqrt{2\pi\sigma^2})]e^{-[(x-\mu)^2/(2\sigma^2)]}$ and σ^2 is the signal variance of X.
3) Some studies have shown that the lateralization between the left and right hemisphere is associated with emotions and language dominance [57]–[60]. This study investigates three asymmetric features: 1) AASM; 2) DASM; and 3) RASM.
   a) *Additional Asymmetry (AASM): AASM* $= DE(x_{\text{left}}) + DE(x_{\text{right}})$.
   b) *Differential Asymmetry (DASM): DASM* $= DE(x_{\text{left}}) - DE(x_{\text{right}})$.
   c) *Rational Asymmetry (RASM): RASM* $= DE((x_{\text{left}})/DE(x_{\text{right}})$.

   There are 27 pairs of asymmetric electrodes (x_{left}, x_{right}): (Fp1, Fp2), (AF3, AF4), (F1, F2), (F3, F4), (F5, F6), (F7, F8), (FC1, FC2), (FC3, FC4), (FC5, FC6), (FT7, FT8), (T7, T8), (C1, C2), (C3, C4), (C5, C6), (TP7, TP8), (CP1, CP2), (CP3, CP4), (CP5, CP6), (P1, P2), (P3, P4), (P5, P6), (P7, P8), (PO3, PO4), (PO5, PO6), (PO7, PO8), (CB1, CB2), and (O1, O2). Note that (x_{left}, x_{right}) denote the symmetric pair of electrodes [55], [61]–[63]. The dimensions of AASM, DASM, and RASM are 162 (6 frequency bands $*$ 27 pairs of asymmetric electrodes).

C. Dimensionality Reduction

The aim of this study is to implement a real-time BCI. Fewer features of real-time BCI correspond to more time-related calculation. This study used principal component analysis (PCA) to reduce the dimensionality. Also, we customized the BVA to extract the significant features cross-subjects, based on the binomial hypothesis test and multifactor-dimensionality reduction (MDR) [64]. There are two hyperparameters to be set in BVA as follows.

1) *O:* The proportion of the optimal feature.
2) *X:* The threshold of the number of interactions between individuals.

In the BVA method, two steps dimensionality reduction are as follows.

Step 1 [Selecting a Certain Amount (Hyperparameter O) of Features for Each Subject]: Taking the RASM feature extraction method, for example, there are 162 features, and each single subject was classified via a linear discriminant analysis (LDA) method based on the 162 features to obtain the corresponding classification accuracy. The hyperparameter O specifies the proportion of the optimal features to be selected out of the 162 features. That is, the values, $O = 10$, $O = 20$, and $O = 30$, are *the highest* 10% (16 features), 20% (32 features), and 30% (48 features) accuracy features, respectively. After setting the O value, the optimal features corresponding to the highest accuracy of each subject were determined.

Step 2 (Selecting the Important Features for the Cross-Subjects by Setting X Value):

f(s, n): The setting of optimal features for each subject.

Bf(n): To compute the important features for cross-subjects, where s is the subject number ranged from 1 to 14, and n is the feature number ranged from 1 to 162 (Note that there are 162 features by using the RASM method).

The value of function $f(s, n)$ equaled 1 ($f(s, n) = 1$) if $f(s, n)$ was selected as the optimal feature; otherwise, it equaled 0 ($f(s, n) = 0$), Then, $Bf(n)$ is defined as $Bf(n) = \Sigma_{i=1}^{s} f(i, n)$, if $Bf(n)$ was greater or equal to X, the feature n was regarded as the important feature for cross-subjects. For the pseudocode, see Algorithm 1.

A small value of O meant that the number of optimal features was small, that is, the feature achieving excellent classification performance of each subject was taken as the threshold; contrarily, a larger value of O indicated a larger range of thresholds. A larger value of X meant that a feature would be set as an important one if the feature was shared by multisubjects, which imposed relatively strict restrictions.

Therefore, both the values of O and X would influence the final number of selected important features. A larger number of important features were less favorable to the real-time BCI design, but were likely to enhance the classification effect, suggesting that the optimization of operating parameters was a crucial consideration.

We now describe how hyperparameters are selected in our methodology. Set the value of O (try $O = 5$, 10, 15, 20, and 25, respectively), and then set the value of X (14 subjects in total, and find X begins from 14, then 13, and substitute one by one successively). We will find the value of important features for each pair of hyperparameters O and X. Finally, we will choose the appropriate number of important features according to the results. In this study, it is expected that the final number of channels will be between 10 and 20, which will be favorable for future BCI development.

Algorithm 1 Binomial Variable Algorithm

Binomial Variable Algorithm (s, n, O, X, Accuracy, Feature)

s: the count of subjects
n: the count of features.
O: the proportion of the optimal feature.
X: the threshold of the number of interactions between individuals.
$Oset(n)$: using O to select features from each subject.
Accuracy (s, n): each single-subject was classified via an LDA method based on the features to obtain the corresponding classification accuracy for every feature.
Feature: all features.
$f(s, n)$: The setting of optimal features for each subject self.
$Bf(n)$: To compute the important features for cross-subjects.

%The setting of optimal features for each subject self.

for $i = 1$ to s **do**
  for $j = 1$ to n **do**
    if $Accuracy(i, j) \in Oset(i)$ **then**
      $f(i, j) = 1$
    else
      $f(i, j) = 0$
    end if
  end for
end for
$Bf \leftarrow 0$
for $j = 1$ to n **do**
  for $i = 1$ to s **do**
    $Bf(j) = Bf(j) + f(i, j)$
  end for
end for

%According to the X value to compute the important features for cross-subjects.

for $j = 1$ to n **do**
  if $Bf(j) \geq X$ **then**
    Important features $\leftarrow$ Feature(j)
  end if
end for
return Important features

Let us consider RASM as an example.

When $O = 5$: $X = 14$, 13, 12, 11, 10, 9, 8, 7, and 6, there is no important feature; $X = 5$, there are four important features; and $X = 4$, there are four important features.

When $O = 10$: $X = 14$, 13, 12, 11, 10, 9, there is one important feature; $X = 8$, there are six important features; $X = 7$, there are nine important features; $X = 6$, there are 14 important features; $X = 5$, there are 19 important features; and $X = 4$, there are 38 important features.

When $O = 20$: $X = 14$, 13, 12, 11, 10, 9, there are five important features; $X = 8$, there are seven important features; $X = 7$, there are 12 important features; $X = 6$, there are 23 important features; $X = 5$, there are 31 important features; and $X = 4$, there are 42 important features.

Therefore, if O is set to five, it will be too harsh, and if O is set to 20, it will be too loose. In the study, we set $O = 10$. After setting O, we set X again. Because more important features represent that an additional number of channels are required, it will be difficult to implement BCI. However, less important features may cause poor classification effect. Therefore, we selected the situation that is more likely to realize BCI for analysis, in which $X = 7$ has nine important features, and $X = 6$, there are 14 important features, therefore this setting has a good chance of achieving BCI, hence this study selected $X = 7$ and $X = 6$ for further analysis.

TABLE I
MEAN ACCURACIES (%) AND STANDARD DEVIATIONS OF LDA-BASED CLASSIFICATION RESULTS FOR PSD, DE, AASM, DASM, AND RASM IN THE *BS*, *Before Speak*, AND *Speak*

Feature		*BS*		*Before Speak*		*Speak*	
		Mean	Std.	Mean	Std.	Mean	Std.
PSD	***372***	50.17	2.66	50.03	0.49	50.09	0.34
DE	***372***	85.75	11.70	56.70	5.59	59.46	4.71
AASM	***162***	94.69	5.98	62.72	6.63	69.86	7.76
DASM	***162***	97.12	3.19	67.32	9.18	69.63	7.69
RASM	***162***	94.79	4.55	**99.06**	1.22	**99.40**	1.05

D. Classification

This study applied LDA, K-nearest neighbor (KNN) algorithms and fivefold cross-validation to the EEG features to classify the spoken Mandarin with versus without tones. K-Fold evaluation is a popular and easy to understand technique. It ensures that every observation in the original data set has a chance to appear in the training and test sets. This study evaluated the associations between EEG power in different frequency bands at different channels and Mandarin speech tones.

V. EVALUATION

In this section, the classifier is combined with different feature extraction methods to classify tonal Mandarin versus nontonal Mandarin. First, we classify a single subject and compare the impacts of the feature extraction methods and classification on different frequency bands.

Tables I and II exhibit the results of single subjects, and Tables III and IV show the results of the classification of 14 subjects by the leave-one-out cross-validation (LOOCV) method (13 were trained, and another one was tested, with each subject being a test subject once). Table III shows the classification results of EEG activities measured in the *baseline (BS), Before Speak*, and *Speak,* and Table IV compares the classification results of the PCA dimensionality reduction method and BVA dimensionality reduction method.

A. Evaluating Single-Subject Performance

For each EEG feature (data point) of each subject's trial, fivefold cross-validation was used to estimate the accuracy of

TABLE II
MEAN ACCURACIES (%) OF LDA-BASED CLASSIFICATION RESULTS FOR PSD, DE, AASM, DASM, AND RASM FROM DIFFERENT BAND IN THE *BS*, *Before Speak*, AND *Speak*

Feature	Period	Delta Theta	Alpha	Beta	Gamma	High Gamma	All band
PSD	*BS*	49.38	50.58	48.90	50.42	50.17	50.12
	Before Speak	48.01	50.11	49.17	50.16	50.01	50.31
	Speak	48.69	50.19	49.79	50.42	49.83	50.23
DE	*BS*	93.26	90.10	91.80	91.47	91.24	92.48
	Before Speak	62.37	57.99	58.98	61.47	62.72	64.50
	Speak	69.11	64.88	64.13	67.51	66.93	69.83
AASM	*BS*	88.01	88.15	87.34	88.36	86.78	90.19
	Before Speak	59.90	61.87	60.48	60.04	59.19	63.22
	Speak	68.13	68.63	65.62	65.00	64.83	66.85
DASM	*BS*	85.27	86.08	85.06	88.94	87.95	89.03
	Before Speak	59.13	60.98	57.22	60.06	61.74	64.52
	Speak	62.49	62.78	60.06	61.51	62.80	65.10
RASM	*BS*	97.43	96.68	96.49	98.18	97.97	98.80
	Before Speak	96.51	96.20	95.33	97.68	98.84	98.03
	Speak	98.09	97.93	96.51	98.40	**99.00**	**98.88**

TABLE III
USING ALL FEATURES TO RUN LOOCV TO TRAIN THE LDA-BASED CROSS-SUBJECT MODEL; THIS TABLE SHOWS THE CROSS-SUBJECT MEAN ACCURACY

Periods		PSD	DE	AASM	DASM	RASM
Numbers of Feature		***372***	***372***	***162***	***162***	***162***
BS	Mean	55.67	86.91	83.01	84.96	**97.93**
	Std.	2.66	1.39	1.75	1.81	0.55
Before Speak	Mean	50.01	61.51	59.58	58.69	97.72
	Std.	2.33	2.46	2.46	2.92	0.75
Speak	Mean	50.49	64.31	62.20	59.56	**98.82**
	Std.	2.57	2.44	2.66	2.23	0.66

TABLE IV
COMPARISON OF THE CLASSIFICATION RESULTS USING THE RASM FEATURE EXTRACTION METHOD IN CONJUNCTION WITH THE DIMENSIONALITY REDUCTION ALGORITHM (PCA\BVA)

Periods	Classifier	RASM	PCA	BVA $X \geq 7$	BVA $X \geq 6$
Numbers of Feature		***162***	***16***	***9***	***14***
BS	LDA	97.93	84.51	77.71	85.09
	KNN	96.20	96.41	99.00	99.22
Before Speak	LDA	97.72	73.91	71.22	77.52
	KNN	76.80	81.82	91.41	94.30
Speak	LDA	98.82	72.90	72.62	76.90
	KNN	76.89	94.22	93.00	**96.70**

LDA classification. We randomly used 370 trials for classification from the total trials. Therefore, there are 370 trials in each subject (185 "nontonal" and 185 "tonal" sentences). We split the data into five subsets, and each subset has 74 trials. We trained the model with 296 trials, and tested it on the remaining 74 trials. We repeated this procedure five times and averaged the accuracy obtained by each subject.

Table I shows the classification results of five different feature extraction methods using all the features (PSD, DE, AASM, DASM, and RASM), where 14 subjects were evaluated based on the LDA model through fivefold cross verification and each subject had 370 trials in the *BS, Before Speak*, and *Speak* states. By averaging the classification results of the 14 subjects based on the LDA model and calculating the standard deviation, it can be found that the RASM method performed best among the results based on a single subject. Some important observations can be summarized as follows.

1) The accuracy by using RASM for the *BS, Before Speak*, and *Speak* is 94.79%, 99.06%, and 99.45%, respectively.
2) The PSD feature extraction method has the worst performance and its accuracy is 50.17%, 50.03%, and 50.09% in the *BS, Before Speak*, and *Speak*, respectively.
3) In the *BS* state, the DASM led to the best performed and the PSD, DE, AASM, DASM, and RASM methods can achieve an accuracy of 50.17%, 85.75%, 94.69%, 97.12%, and 94.79%, respectively.
4) In the *Before Speak* state, the RASM method performed the best with the accuracy of 99.06%, while the accuracy of the PSD, DE, AASM, and DASM methods is 50.03%, 56.70%, 62.72%, and 67.32%, respectively.
5) In the *Speak* state, the RASM method performed the best with an accuracy of 99.40%, comparing to the accuracy of the PSD, DE, AASM, and DASM methods at 50.09%, 59.46%, 69.86%, 69.63%, and 99.40%, respectively.
6) The best accuracy of the LDA classifier is 99.40% using the RASM in the *Speak* state.

Brain activities in different frequency bands usually reflect their distinct cognitive activities [53], [54], [65]–[67]. Therefore, this study divided brain activities into six frequency bands and classified them with different feature extraction methods to see if there is a difference in classification across frequency bands. This study used the PSD, DE, AASM, DASM, and RASM methods to extract EEG features in the *BS, Before Speak*, and *Speak* states. Table II shows the average classification results obtained by the LDA model. In the case of using High Gamma (60–170 Hz), the RASM method achieved the highest accuracy (99.00%) in the *Speak* state. Even in the *Speak* state of using only a specific frequency band, the RASM method still outperformed the other feature-extraction methods. Using the RASM method, the best classification band in the *Speak* state is High Gamma (99.00%), followed by All band (98.88%), and Gamma (98.20%). For the RASM feature in the *BS* state, the all-band-based classification accuracy was 98.80%, followed by Gamma (98.38%) and High Gamma (97.97%). The best classification band in the *Before Speak* state using RASM features is High Gamma (98.84%), followed by All-band (98.03%) and Gamma (97.68%).

B. Evaluating Cross-Subjects Performance

The LOOCV method was used to perform an LDA-based cross-subject classification on the 14 subjects, of which data from 13 subjects were used for training and data from the one remaining subject were used for testing. Table III shows the results obtained from using different feature-extraction methods. The RASM feature-extraction method performed best in the cross-subjects classification result. It achieved a classification accuracy of 98.82% in the *Speak* state, with a standard deviation of 0.66, which is the minimum standard deviation among the five feature-extraction methods in the *Speak* state. The RASM feature-extraction method achieved the best accuracy of 97.93%, 97.72%, and 98.82% in the *BS, Before Speak*, and *Speak*, respectively. The accuracy (%) of the PSD, DE, AASM, DASM, and RASM-based classification in the three

states can be averaged to 52.06, 70.91, 68.26, 67.73, and 98.16, respectively. It suggests that RASM features achieved the highest classification accuracy (98.16%), followed by DE (70.91%).

Table III indicates that RASM achieved the highest classification accuracy under the LDA-based cross-subject classification. Table IV compares the classification results using the RASM method in conjunction with the dimensionality reduction algorithm (PCA versus BVA). 10% of the features (as RASM has 162 features; 10% $x162 = 16$) are taken in the PCA algorithm, that is, the first 16 principal components were used as the important features.

Under the LDA-based classifier, the PCA-based dimensionality-reduction method reached the accuracy of 84.51%, 73.91%, and 72.90% in the *BS*, *Before Speak*, and *Speak* state, respectively. With ($X \geq 7$), 9 important features were taken in the BVA-based dimensionality-reduction method, which achieved the accuracy of 77.71%, 71.22%, and 72.62% in the *BS, Before Speak*, and *Speak* state, respectively; with ($X \geq 6$) 14 important features were taken in the BVA-based method, which achieved the accuracy of 85.09 %, 77.52%, and 76.90% in the *BS, Before Speak*, and *Speak* state, respectively. For the LDA model-based classifier, we find that the numbers of features/components both PCA and BAV methods affect the performance significantly. For the PCA algorithm, when the dimension is reduced to 16, the accuracy (%) dropped from 97.93% to 84.51%, from 97.72% to 73.91%, and from 98.82% to 72.90% in the *BS, Before Speak*, and *Speak* phase, respectively. For the BVA ($X \geq 6$) algorithm, when the dimension is reduced to 14, the accuracy dropped from 97.93% to 85.09%, from 97.72% to 77.52%, and from 98.82% to 76.90% in the *BS, Before Speak*, and *Speak* state, respectively.

When we used BVA ($X \geq 6$), there are 14 important features. The results of the LDA-based classification with these 14 important features can be compared with the results of the PCA-based using 16 important features. The accuracy of using PCA during the *BS*, *Before*, and *Speak* states is 84.51%, 73.91%, and 72.90%, respectively. The accuracy of using BVA ($X \geq 6$) during *BS*, *Before Speak*, and *Speak* states is 85.09%, 77.52%, and 76.90%, respectively. We also performed *t*-test analysis to test the statistical significance between BVA and PCA methods. The BVA ($X \geq 6$) using 14 features outperformed the PCA using 16 components significantly ($\rho < 0.01$).

Using the PCA/BVA methods to reduce the dimensionality, the KNN classifier can obtain a higher accuracy because the important features are identified to improve the disadvantages of the traditional KNN algorithm. The traditional KNN classification has three limitations.

1) *High Calculation Complexity:* To find the k nearest neighboring samples by KNN, all the similarities between the training samples must be calculated. When there are few training samples, the calculation time is not overwhelming, but if the training set contains a large number of samples, the KNN classifier needs more time to calculate the similarity [68].

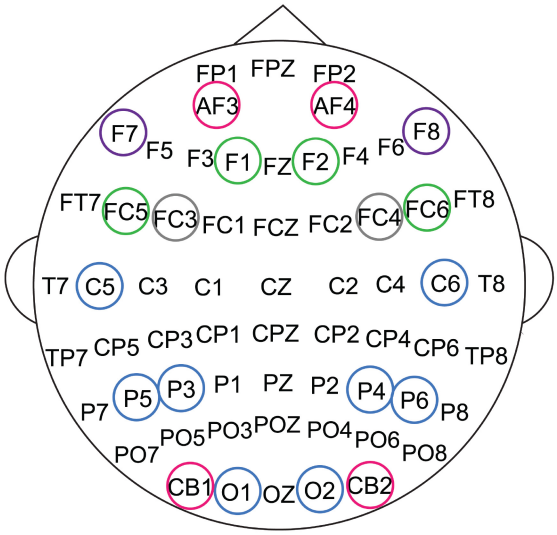

Fig. 6. Distribution of the 14 important features for the tone and nontonal sentence in the EEG classification using BVA ($x \geq 6$) were plotted in the cap. The blue circles represent All-band, the purple circles represent High Gamma, the gray circles represent Gamma band, the pink circles represent both Alpha and All-band, and green circles represent both Alpha and Gamma band.

2) *Dependence on the Training Set:* The classifier is only generated with the training samples and does not use any additional data. This makes the algorithm dependent on the training set excessively. It needs to be recalculated even if the training set has a small change.
3) *No Weight Difference Between Samples:* Because training samples are treated equally in the KNN, there is no difference between the samples with small and large amounts of data.

With the help of the BVA ($X \geq 6$) method using only 14 features, the KNN classifier can achieve an accuracy of 99.2%, 94.3%, and 96.7% in the *BS*, *Before Speak*, and *Speak* states, respectively. We also find that the BVA outperforms the PCA when $\rho < 0.01$ under a KNN model-based classifier. Table IV reveals that using all features, the LDA classifier could achieve higher accuracy. Presumably, the LDA can only learn simple linear boundaries among the data clusters. The high classification performance obtained by the LDA indicates that there were obvious differences in the data distributions of EEG signals under different conditions, which may reflect different cognitive activities of the brains.

As shown in Table IV, we speculate that the 14 features with BVA (interaction threshold is larger than 6, $x \geq 6$) are important for distinguishing tonal and nontonal sentences in the EEG classification. Fig. 6 displays the distribution of the representing channels of 14 feature pairs in the EEG cap. The blue circles mark *All-band* feature pairs, (C5, C6), (P3, P4), (P5, P6), and (O1, O2). The purple circles mark *High Gamma* band feature pair, (F7, F8). The gray circles mark *Gamma* band feature pair, (FC3, FC4). The pink circles mark both *Alpha and All-band* feature pairs, (AF3, AF4) and (CB1, CB2). The green circles mark both *Alpha and Gamma* feature pairs, (F1, F2) and (FC5, FC6)

In order to identify the critical features among the 14 important features obtained by the BVA ($X \geq 6$) method in

TABLE V
ANALYSIS OF 14 BVA-BASED IMPORTANT FEATURES; WITH EACH FEATURE REMOVED IN TURN; THE OTHER 13 FEATURES ARE USED EACH TIME. THE REMOVAL OF (C5, C6) CAUSES THE ACCURACY TO DROP SIGNIFICANTLY

Lack of Feature	Accuracy (%)
(F1,F2) Alpha	96.6
(AF3,AF4) Alpha	96.7
(FC5,FC6) Alpha	96.5
(CB1,CB2) Alpha	96.6
(F1,F2) Gamma	96.6
(FC3,FC4) Gamma	96.7
(FC5,FC6) Gamma	96.5
(F7,F8) HighGamma	96.3
(AF3,AF4) All-Band	96.4
(C5,C6) All-Band	**95.6**
(P3,P4) All-Band	96.6
(P5,P6) All-Band	96.4
(CB1,CB2) All-Band	96.6
(O1,O2) All-Band	96.4

TABLE VI
DISTRIBUTION OF THE IMPORTANT FEATURES FOR THE MANDARIN TONE IN THE EEG CLASSIFICATION USING BVA ($x \geq 6$) PROJECT FOR THE BRAIN AREA

Area	Count	A Channel Pair
Frontal	5	(F1,F2), (F7,F8), (AF3,AF4), (FC3,FC4), (FC5,FC6)
Tempotal	1	(C5,C6)
Parietal	2	(P3,P4), (P5,P6)
Occipital	2	(O1,O2), (CB1,CB2)

the KNN-based classification, we performed the leave-one-feature-out test, that is, using 13 of 14 features each time in the *Speak* state. The results are given in Table V. When the (C5, C6) All-band feature were removed, the accuracy dropped significantly; however, the removal of other features did not have a significant influence, indicating that (C5, C6) were the critical features in this study, which was consistent with prior literature results that C5 and C6 were most relevant to the brain areas in the *Speak* state [69]. Table V divided the 14 important features obtained through the BVA into the frontal lobe, temporal lobe, parietal lobe, and occipital lobe of a brain.

Aiming to reduce the dimensionality based on regions, we took a channel pair as feature values in each region, with a total of 20 combinations (Frontal: 5 x Temporal: 1 x Parietal: 2 x Occipital: $2 = 20$). The KNN method was used to classify the tones based on the brain dynamics in the *Speak* state. The results given in Table VII indicates that the combination of (F7, F8), (C5, C6), (P5, P6), and (O1, O2) achieved the highest accuracy (94.44%), and the frontal lobe-based data channel of the pair achieved the highest accuracy in conjunction with the combination of (C5, C6), (P5, P6), and (O1, O2) in each block in Table VII. Thus, it is implied that using eight channels can achieve an accuracy of 94.44%.

VI. DISCUSSION

Communications in the real life are achieved through sentences. The use of a sentence-level design to study speech can be applied to natural languages than word-level design. This study aims to understand the neural processing before and during speaking. We investigated the brain activities during the *BS* (baseline), *Before Speaking*, and *Speaking* a sentence in

TABLE VII
DATA CHANNEL PAIR CORRESPONDING TO EACH BRAIN AREA IN TABLE VI, WITH THE FEATURES EXTRACTED USING THE RASM METHOD. THE KNN CLASSIFIER IS USED TO CLASSIFY THE TONE AND NONTONAL SENTENCES ACCORDING TO THE BRAIN SIGNALS IN THE *Speak* STATE

Channels	Accuracy (%)	Std.
(F1,F2), (C5,C6), (P3,P4), (O1,O2)	92.60	1.94
(F1,F2), (C5,C6), (P3,P4), (CB1,CB2)	91.14	1.35
(F1,F2), (C5,C6), (P5,P6), (O1,O2)	**92.66**	**1.61**
(F1,F2), (C5,C6), (P5,P6), (CB1,CB2)	91.72	1.27
(F7,F8), (C5,C6), (P3,P4), (O1,O2)	93.96	1.25
(F7,F8), (C5,C6), (P3,P4), (CB1,CB2)	93.74	1.38
(F7,F8), (C5,C6), (P5,P6), (O1,O2)	**94.44**	**1.26**
(F7,F8), (C5,C6), (P5,P6), (CB1,CB2)	93.28	0.98
(AF3,AF4), (C5,C6), (P3,P4), (O1,O2)	91.10	1.73
(AF3,AF4), (C5,C6), (P3,P4), (CB1,CB2)	90.37	1.23
(AF3,AF4), (C5,C6), (P5,P6), (O1,O2)	**91.56**	**1.10**
(AF3,AF4), (C5,C6), (P5,P6), (CB1,CB2)	90.39	1.65
(FC3,FC4), (C5,C6), (P3,P4), (O1,O2)	92.51	1.01
(FC3,FC4), (C5,C6), (P3,P4), (CB1,CB2)	91.87	1.05
(FC3,FC4), (C5,C6), (P5,P6), (O1,O2)	**93.09**	**1.26**
(FC3,FC4), (C5,C6), (P5,P6), (CB1,CB2)	92.24	1.21
(FC5,FC6), (C5,C6), (P3,P4), (O1,O2)	93.03	1.33
(FC5,FC6), (C5,C6), (P3,P4), (CB1,CB2)	92.12	1.43
(FC5,FC6), (C5,C6), (P5,P6), (O1,O2)	**93.75**	**1.14**
(FC5,FC6), (C5,C6), (P5,P6), (CB1,CB2)	92.66	0.85

Mandarin. In Table VII, using KNN with BVA ($X \geq 6$) yielded accuracies that all exceeded 90% (highest is 94.44%) during the *Speak* state. Also, using the proposed real-time onset detection techniques for BCI, such as Matthews *et al.* [70], Chamanzar *et al.* [71], Chamanzar *et al.* [72], and with the aid of dimensionality reduction, our proposed BCI method can be used for accurate and real-time classification of the tonal versus nontonal language. This study will have the following potential impacts.

A. New Findings in Tone-Speaking Brain Dynamics

By analyzing the EEG recordings, the study demonstrates that it is possible to differentiate whether a native Mandarin speaker is using tone or not. Specifically, the cognitive processes of speaking tonal or nontonal Mandarin are different. Many Feature extraction methods for analyzing the EEG, have been repeated in the literature, such as the asymmetric feature extraction method [55], [61]–[63] and the single-channel-based method [40]–[45], [73]. Most of the previous EEG speech recognition research works were based on single-channel-based methods for feature extraction. To the best of our knowledge, this is the first study to apply the asymmetric feature extraction method for speech recognition through EEG signals. Additionally, this study finds that the RASM feature extraction method can achieve the best accuracy among these feature extraction methods (PSD, DE, AASM, DASM, and RASM) in the classification of cross-subjects. RASM is one of the asymmetric feature extraction methods and is just like a normalized process that changes the values of numeric columns in the data set to a common scale. It is obtained by dividing the left channel's value by the right channel's value. After RASM, the accomplishment of the best classification result by the classifier also proves the asymmetric cognitive process of the left and right hemispheres when the speech is delivered with tone or not [74] *(important result 1)*.

To exclude the influence of emotions, we removed the *BS* state data in the *Before Speak* and *Speak* state counterpart to ensure that the classification results were based on the tonal and nontonal language classification. According to previous studies, emotion might affect brain activities at the rest state [53], [54], [75]. From Table IV, we observe that there are great classification results between *Baseline* and *Before Speak* states. Therefore, the results support that speaking with or without tone is related to speaking motivation and articulation*(important result 2)*. Therefore, this study can help design the direct speech BCI and facilitate human–machine interaction (HMI). In the design of BCI or HMI, a prejudgment involving emotion before speaking is a very important design link for many patients with aphasia. It also truly implements the spirit of automation science and engineering [76].

When we communicate with others, no matter whether we speak a tonal or nontonal language, there may come with emotions when we speak. For example, when we are saying that I am very happy, we may have happy emotions, and the prosody of speech may also be changed. From the conclusions of *(important result 1)* and *(important result 2)*, we also speculate that the asymmetric feature extraction method may be not only helpful for the tonal language but also for the nontonal language when we are speaking in a natural situation. The hypothesis is worthwhile being verified in the future.

Previous research results indicated that the brain elicits high-Gamma (70–160 Hz) oscillations during linguistic phonetic processing [77], [78]. Although the cognitive process of speech in the brain is still unclear, we can speculate that high-Gamma will be an important feature for analyzing brain dynamics when speaking. Single-subject results in this study showed that when using the RASM method, the best classification band in the *Speak* state is High Gamma (99.00%) *(important result 3)*.

The results of this study have seen not only the language-related brain areas, such as parietal [79] and temporal [69] but also the frontal and occipital area, which may be triggered by the stimulus-driven executive control. We also found that channels (C5, C6) are the critical feature in this study, which is consistent with the prior research results mentioning that C5 and C6 are most relevant to the brain area when speaking (articulation) [69] *(important result 4)*.

From those results, this study obtained a satisfactory classification accuracy, indicating that different brain mechanisms may be used by the tonal and nontonal Mandarin in terms of cognitive behaviors. While the majority of previous studies have focused on the brain studies for the nontonal languages, this is the first study to analyze the presence or absence of tones in sentences based on the EEG signals for tonal languages. By using the machine learning classification approach, we confirmed brain activities (cognitive-behavioral differences) are different when people speak with or without tone.

B. Key Step to Direct-Speech BCI

This study found that the cognitive process of the brain while speaking with or without tone is different. Previous studies of speech synthesis have already indicated the articulation space of the brain when speaking English [24–27]. English is a nontonal language, according to this study, a nontonal language synthesis's model cannot be directly applied to a tonal language. However, over 60% of the world population use tonal languages [18], and Mandarin is one of the most widely spoken tonal languages. To ascertain the tonal feature is not only the key step to the direct-speech BCI of tonal languages but also the cross-language direct-speech BCI.

The largest difference between the tonal and the nontonal languages lies in their tones. If the tonal feature can be interpreted by physiological signal analysis, there is an opportunity to add tonal features based on the articulation space of English shown in past studies. Then, not only can the tonal languages be synthesized but the cross-language direct-speech BCI can also be achieved. BCIs can serve all ethnic groups and languages, which is the ultimate goal of Automation Science and Engineering [76]. We are looking forward to the invention of such a BCI.

VII. Conclusion

This study investigated the brain dynamics of human speech in tonal and nontonal Mandarin based on EEG recognition. In contrast to ECoG and fMRI, EEG signals have the advantages of low cost, mobility, fieldability, high-temporal resolution, and noninvasiveness. The brain activities corresponding to the tonal and nontonal Mandarin sentences exhibit different behaviors that can be distinguished by classifying EEG. To the best of our knowledge, this is the first study to apply the asymmetric feature extraction method for speech recognition through EEG signals. This study finds that the RASM feature extraction method can achieve the best accuracy in the classification of cross-subjects. Also, our proposed methodology, BVA, can achieve an accuracy of 98.82% in cross-subject classification. Furthermore, we show that using eight channels [(F7, F8), (C5, C6), (P5, P6), and (O1, O2)] can achieve an accuracy of 94.44%. The methods to discover different brain activities developed in this study will benefit and shed the light on the design of future BCI of speech synthesis for 60% of people in the world who use tonal languages.

Acknowledgment

The authors would like to thank Prof. Sin-Horng Chen, Prof. Chen-Yu Chiang, and Prof. Ching-Ching Lu for their insightful discussion on the design of linguistics experiment, and to Prof. Guan-Hua Huang, Dr. Jung-Tai Chin, Ta Yu Huang, and Shao-Ting Hsu for their help and encouragement during this research.

References

[1] K. Hilari, J. J. Needle, and K. L. Harrison, "What are the important factors in health-related quality of life for people with aphasia? A systematic review," *Archives Phys. Med. Rehabil.*, vol. 93, no. 1, pp. S86–S95, 2012.

[2] C. I. Penaloza, Y. Mae, F. F. Cuellar, M. Kojima, and T. Arai, "Brain machine interface system automation considering user preferences and error perception feedback," *IEEE Trans. Autom. Sci. Eng.*, vol. 11, no. 4, pp. 1275–1281, Oct. 2014.

[3] K.-H. Park, H.-E. Lee, Y. Kim, and Z. Z. Bien, "A steward robot for human-friendly human-machine interaction in a smart house environment," *IEEE Trans. Autom. Sci. Eng.*, vol. 5, no. 1, pp. 21–25, Jan. 2008.

[4] C. Kan, Y. Chen, F. Leonelli, and H. Yang, "Mobile sensing and network analytics for realizing smart automated systems towards health Internet of Things," in *Proc. IEEE Int. Conf. Autom. Sci. Eng. (CASE)*, 2015, pp. 1072–1077.

[5] S. N. Abdulkader, A. Atia, and M.-S. M. Mostafa, "Brain computer interfacing: Applications and challenges," *Egypt. Inform. J.*, vol. 16, no. 2, pp. 213–230, 2015.

[6] D. Mulfari, A. Celesti, M. Fazio, and M. Villari, "Human-computer interface based on IoT embedded systems for users with disabilities," in *Proc. Int. Internet Things Summit*, 2014, pp. 376–383.

[7] M. Spüler, "A high-speed brain–computer interface (BCI) using dry EEG electrodes," *PLoS One*, vol. 12, no. 2, 2017, Art. no. e0172400.

[8] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "An empirical mode decomposition based filtering method for classification of motor-imagery EEG signals for enhancing brain–computer interface," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, 2015, pp. 1–7.

[9] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "A multivariate empirical mode decomposition based filtering for subject independent BCI," in *Proc. Irish Signals Syst. Conf. (ISSC)*, 2016, p. 7.

[10] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "Enhanced motor imagery classification in EEG-BCI using multivariate EMD based filtering and CSP features," in *Proc. Int. Brain Comput. Interface (BCI) Meeting*, 2016, p. 98.

[11] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "A multi-class EEG-based BCI classification using multivariate empirical mode decomposition based filtering and Riemannian geometry," *Expert Syst. Appl.*, vol. 95, pp. 201–211, Apr. 2018.

[12] P. Gaur, K. McCreadie, R. B. Pachori, H. Wang, and G. Prasad, "Tangent space features-based transfer learning classification model for two-class motor imagery brain–computer interface," *Int. J. Neural Syst.*, vol. 29, no. 10, 2019, Art. no. 1950025.

[13] P. Gaur, R. B. Pachori, H. Wang, and G. Prasad, "An Automatic subject specific intrinsic mode function selection for enhancing two-class EEG-based motor imagery-brain computer interface," *IEEE Sensors J.*, vol. 19, no. 16, pp. 6938–6947, Aug. 2019.

[14] P. Gaur, G. Kaushik, R. B. Pachori, H. Wang, and G. Prasad, "Comparison analysis: Single and multichannel EMD-based filtering with application to BCI," in *Proc. Mach. Intell. Signal Anal.*, 2019, pp. 107–118.

[15] J. Kevric and A. Subasi, "Comparison of signal decomposition methods in classification of EEG signals for motor-imagery BCI system," *Biomed. Signal Process. Control*, vol. 31, pp. 398–406, Jan. 2017.

[16] J. Shin, K.-R. Müller, and H.-J. Hwang, "Near-infrared spectroscopy (NIRS)-based eyes-closed brain–computer interface (BCI) using prefrontal cortex activation due to mental arithmetic," *Sci. Rep.*, vol. 6, Nov. 2016, Art. no. 36203.

[17] K. Hilari, "The impact of stroke: Are people with aphasia different to those without?" *Disabil. Rehabil.*, vol. 33, no. 3, pp. 211–218, 2011.

[18] M. Yip, *Tone*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[19] J. M. Howie and J. M. Howie, *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge, U.K.: Cambridge Univ. Press, 1976.

[20] Z. Hua and B. Dodd, "The phonological acquisition of Putonghua (modern standard Chinese)," *J. Child Lang.*, vol. 27, no. 1, pp. 3–42, 2000.

[21] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1317–1320, Sep. 1990.

[22] C. Herff *et al.*, "Brain-to-text: Decoding spoken phrases from phone representations in the brain," *Front. Neurosci.*, vol. 9, p. 217, Jun. 2015.

[23] X. Pei, J. Hill, and G. Schalk, "Silent communication: Toward using brain signals," *IEEE Pulse*, vol. 3, no. 1, pp. 43–46, Jan. 2012.

[24] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[25] D. A. Moses, M. K. Leonard, J. G. Makin, and E. F. Chang, "Real-time decoding of question-and-answer speech dialogue using human cortical activity," *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[26] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG classification of covert speech for brain–computer interfacing," *Int. J. Neural Syst.*, vol. 27, no. 8, 2017, Art. no. 1750033.

[27] M. Angrick *et al.*, "Speech synthesis from ECoG using densely connected 3D convolutional neural networks," *J. Neural Eng.*, vol. 16, no. 3, 2019, Art. no. 036019.

[28] D. Perani *et al.*, "The bilingual brain. Proficiency and age of acquisition of the second language," *Brain J. Neurol.*, vol. 121, no. 10, pp. 1841–1852, 1998.

[29] A. Hahne, "What's different in second-language processing? Evidence from event-related brain potentials," *J. Psycholinguist. Res.*, vol. 30, no. 3, pp. 251–266, 2001.

[30] Z. Dörnyei, *The Psychology of Second Language Acquisition*. Oxford, U.K.: Univ. Press, 2009.

[31] S. Reiterer, E. Pereda, and J. Bhattacharya, "Measuring second language proficiency with EEG synchronization: How functional cortical networks and hemispheric involvement differ as a function of proficiency level in second language speakers," *Second Lang. Res.*, vol. 25, no. 1, pp. 77–106, 2009.

[32] E. M. Mugler *et al.*, "Direct classification of all American English phonemes using signals from functional speech motor cortex," *J. Neural Eng.*, vol. 11, no. 3, 2014, Art. no. 035015.

[33] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[34] L. Naci, R. Cusack, V. Z. Jia, and A. M. Owen, "The brain's silent messenger: Using selective attention to decode human thought for brain-based communication," *J. Neurosci.*, vol. 33, no. 22, pp. 9385–9393, 2013.

[35] C. J. Price, "A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading," *Neuroimage*, vol. 62, no. 2, pp. 816–847, 2012.

[36] A. G. Huth, W. A. De Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant, "Natural speech reveals the semantic maps that tile human cerebral cortex," *Nature*, vol. 532, no. 7600, p. 453, 2016.

[37] C. Cheung and E. F. Chang, "Real-time, time–frequency mapping of event-related cortical activation," *J. Neural Eng.*, vol. 9, no. 4, 2012, Art. no. 046018.

[38] 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," *J. Neural Eng.*, vol. 7, no. 5, 2010, Art. no. 056007.

[39] B. N. Pasley *et al.*, "Reconstructing speech from human auditory cortex," *PLoS Biol.*, vol. 10, no. 1, 2012, Art. no. e1001251.

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

[41] K. Brigham and B. V. Kumar, "Imagined speech classification with EEG signals for silent communication: A preliminary investigation into synthetic telepathy," in *Proc. 4th IEEE Int. Conf. Bioinform. Biomed. Eng.*, 2010, pp. 1–4.

[42] K. Brigham and B. V. Kumar, "Subject identification from electroencephalogram (EEG) signals during imagined speech," in *Proc. 4th IEEE Int. Conf. Biometr. Theory Appl. Syst. (BTAS)*, 2010, pp. 1–8.

[43] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "EEG classification of covert speech using regularized neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2292–2300, Dec. 2017.

[44] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2019, pp. 1090–1094.

[45] G. Krishna, C. Tran, M. Carnahan, and A. H. Tewfik, "Advancing speech recognition with no speech or with noisy speech," 2019, *arXiv:1906.08871*.

[46] F. Rabiee, "Focus-group interview and data analysis," *Proc. Nutr. Soc.*, vol. 63, no. 4, pp. 655–660, 2004.

[47] E. C. Sagey, *The Representation of Features and Relations in Non-Linear Phonology*. Cambridge, MA, USA: Massachusetts Inst. Technol., 1986.

[48] P. Ladefoged and M. Halle, "Some major features of the international phonetic alphabet," *Language*, vol. 64, no. 3, pp. 577–582, 1988.

[49] C.-M. Longtin, J. Segui, and P. A. Hallé, "Morphological priming without morphological relationship," *Lang. Cogn. Processes*, vol. 18, no. 3, pp. 313–334, 2003.

[50] V. Zue, S. Seneff, and J. Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Commun.*, vol. 9, no. 4, pp. 351–356, 1990.

[51] C. Neuper, R. Scherer, M. Reiner, and G. Pfurtscheller, "Imagery of motor actions: Differential effects of kinesthetic and visual-motor mode of imagery in single-trial EEG," *Cogn. Brain Res.*, vol. 25, no. 3, pp. 668–677, 2005.

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

[53] P. Putman, "Resting state EEG delta-beta coherence in relation to anxiety, behavioral inhibition, and selective attentional processing of threatening stimuli," *Int. J. Psychophysiol.*, vol. 80, no. 1, pp. 63–68, 2011.

[54] J. Wu, R. Srinivasan, A. Kaur, and S. C. Cramer, "Resting-state cortical connectivity predicts motor skill acquisition," *Neuroimage*, vol. 91, pp. 84–90, May 2014.

[55] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Trans. Affective Comput.*, vol. 10, no. 3, pp. 417–429, Jul.–Sep. 2019.

[56] L.-C. Shi, Y.-Y. Jiao, and B.-L. Lu, "Differential entropy feature for EEG-based vigilance estimation," in *Proc. 35th IEEE Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2013, pp. 6627–6630.

[57] R. J. Davidson, "Anterior cerebral asymmetry and the nature of emotion," *Brain Cogn.*, vol. 20, no. 1, pp. 125–151, 1992.

[58] C. Helmstaedter, M. Kurthen, D. Linke, and C. Elger, "Patterns of language dominance in focal left and right hemisphere epilepsies: Relation to MRI findings, EEG, sex, and age at onset of epilepsy," *Brain Cogn.*, vol. 33, no. 2, pp. 135–150, 1997.

[59] D. L. Molfese, "Left and right hemisphere involvement in speech perception: Electrophysiological correlates," *Percept. Psychophys.*, vol. 23, no. 3, pp. 237–243, 1978.

[60] N. T. Alves, S. S. Fukusima, and J. A. Aznar-Casanova, "Models of brain asymmetry in emotional processing," *Psychol. Neurosci.*, vol. 1, no. 1, p. 63, 2008.

[61] D. O. Bos, "EEG-based emotion recognition," *Influence Visual Auditory Stimuli*, vol. 56, no. 3, pp. 1–17, 2006.

[62] R. Jenke, A. Peer, and M. Buss, "Feature extraction and selection for emotion recognition from EEG," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 327–339, Jul./Sep. 2014.

[63] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[64] M. D. Ritchie *et al.*, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Amer. J. Human Genet.*, vol. 69, no. 1, pp. 138–147, 2001.

[65] M. C. Bastiaansen, R. Oostenveld, O. Jensen, and P. Hagoort, "I see what you mean: Theta power increases are involved in the retrieval of lexical semantic information," *Brain Lang.*, vol. 106, no. 1, pp. 15–28, 2008.

[66] M. C. Bastiaansen, M. V. D. Linden, M. T. Keurs, T. Dijkstra, and P. Hagoort, "Theta responses are involved in lexical—Semantic retrieval during language processing," *J. Cogn. Neurosci.*, vol. 17, no. 3, pp. 530–541, 2005.

[67] R. Hannemann, J. Obleser, and C. Eulitz, "Top-down knowledge supports the retrieval of lexical information from degraded speech," *Brain Res.*, vol. 1153, pp. 134–143, Jun. 2007.

[68] Y. Wang and Z.-O. Wang, "A fast KNN algorithm for text categorization," in *Proc. IEEE Int. Conf. Mach. Learn. Cybern.*, vol. 6, pp. 3436–3441, 2007.

[69] B. J. Baars and N. M. Gage, *Cognition, Brain, and Consciousness: Introduction to Cognitive Neuroscience*. Amsterdam, The Netherlands: Academic, 2010.

[70] B. A. Matthews *et al.*, "Automatic detection of speech activity from neural signals in Broca's Area," in *Proc. Neurosci. Meeting Planner*, Washington, DC, USA, 2008.

[71] A. Chamanzar, M. Shabany, A. Malekmohammadi, and S. Mohammadinejad, "Efficient hardware implementation of real-time low-power movement intention detector system using FFT and adaptive wavelet transform," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 3, pp. 585–596, Jun. 2017.

[72] A. Chamanzar, A. Malekmohammadi, M. Bahrani, and M. Shabany, "Accurate single-trial detection of movement intention made possible using adaptive wavelet transform," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2015, pp. 1914–1917.

[73] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," *Biomed. Eng. Lett.*, vol. 10, pp. 217–226, Mar. 2020.

[74] T. A. Knaus, D. M. Corey, A. M. Bollich, L. C. Lemen, and A. L. Foundas, "Anatomical asymmetries of anterior perisylvian speech-language regions," *Cortex*, vol. 43, no. 4, pp. 499–510, 2007.

[75] J. A. Coan and J. J. Allen, "Frontal EEG asymmetry as a moderator and mediator of emotion," *Biol. Psychol.*, vol. 67, nos. 1–2, pp. 7–50, 2004.

[76] K. Goldberg, "What is automation?" *IEEE Trans. Autom. Sci. Eng.*, vol. 9, no. 1, pp. 1–2, Jan. 2012.

[77] M. Fukuda, R. Rothermel, C. Juhász, M. Nishida, S. Sood, and E. Asano, "Cortical gamma-oscillations modulated by listening and overt repetition of phonemes?" *Neuroimage*, vol. 49, no. 3, pp. 2735–2745, 2010.

[78] V. L. Towle *et al.*, "ECoG gamma activity during a language task: Differentiating expressive and receptive speech areas," *Brain*, vol. 131, no. 8, pp. 2013–2027, 2008.

[79] S. L. E. Brownsett and R. J. S. Wise, "The contribution of the parietal lobes to speaking and writing," *Cerebr. Cortex*, vol. 20, no. 3, pp. 517–523, 2010.