Decoding of Chinese phoneme clusters using ECoG

Chen Song, Rui Xu, Bo Hong*, *IEEE Member*

Abstract— **A finite set of phonetic units is used in human speech, but how our brain recognizes these units from speech streams is still largely unknown. The revealing of this neural mechanism may lead to the development of new types of speech brain computer interfaces (BCI) and computer speech recognition systems. In this study, we used electrocorticography (ECoG) signal from human cortex to decode phonetic units during the perception of continuous speech. By exploring the wavelet time-frequency features, we identified ECoG electrodes that have selective response to specific Chinese phonemes. Gamma and high-gamma power of these electrodes were further combined to separate sets of phonemes into clusters. The clustered organization largely coincided with phonological categories defined by the place of articulation and manner of articulation. These findings were incorporated into a decoding framework of Chinese phonemes clusters. Using support vector machine (SVM) classifier, we achieved consistent accuracies higher than chance level across five patients discriminating specific phonetic clusters, which suggests a promising direction of implementing a speech BCI.**

I. Introduction

Human speech consists of sequence of building blocks including phonemes, syllables, words or phrases [1]. But how our brain recognize these units from continuous speech is still largely unknown. In speech BCI studies, decoding methods at the level of phonemes, syllables or directly based on spectrogram of acoustic signals have been studied[2-4], but these systems usually explore the neural code in speech production or use the discretely presented speech units. Here we use segmented continuous speech material to study whether it is feasible to implement a speech BCI to decode Chinese (Mandarin) phonemes from neural activities over speech perception cortex.

A typical Chinese syllable consists of three parts of phonemes, the initial, final and tone. The initial is usually a consonant, and final is usually combination of vowels. The final can be further segmented into three parts, medial, kernel and coda[5], containing usually one phoneme in each part. For convenience, we call the final in one Chinese syllable a "Chinese phoneme", indicating its integrity in both production and perception. Besides, we denote the medial, kernel and coda as "head", "body" and "tail" in a typical Chinese phoneme.

*This work was supported by National Program on Key Basic Research Projects of China (2011CB933204) and Ministry of Science and Technology Grant 2012BAI16B03.

B. Hong is with the Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China(e-mail: hongbo@tsinghua.edu.cn).

C. Song and R. Xu are with the Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China

In phonology, the distinctive feature system has clustered these Chinese phonemes into categories, represented mainly by their composing phoneme's place of articulation and manner of articulation[6]. Recently, it has been suggested that during perception of continuous speech, our brain also processes the phonemes by population encoding of these distinctive features [7, 8]. Meanwhile, these studies show the feasibility of decoding phoneme clusters directly from the ECoG signals in human auditory cortex. However, due to the different syllable structure in Chinese and English, it has been proposed that the Chinese syllable as a whole, instead of its composing phonemes, is the fundamental unit[9, 10]. In terms of neural encoding, whether Chinese phonemes are organized into clusters as being defined by the distinctive features of vowels and consonants is still an unsolved question.

In this study, we investigate whether the Chinese phoneme clusters can be classified from ECoG signals and how they are organized in the space of neural representation. We hypothesize that there exist brain areas responding to specific Chinese phoneme clusters, and these clusters may resemble the phonological categories of phonemes. ECoG data were obtained from 5 epilepsy patients with subdural electrodes. During the experiment, the patient attentively listened to continuous speech materials. Selective responses to phonemes in the speech stream mainly appear in the posterior temporal and inferior parietal cortex. Using wavelet features and SVM classifiers, we have tested the accuracies of classifying Chinese phoneme clusters into different categories defined by place of articulation and manner of articulation. All accuracies are significantly higher than chance level, suggesting the feasibility of building a speech brain-computer interface (BCI) system based on Chinese phoneme clusters.

II. Methods

A. Paradigm and Procedure

9 narrative stories (2~3min each) read by three native Mandarin speakers (one female) were presented to the patients. Patients were asked to attentively listen to the stories and answer several story content related questions afterwards. The experiment program was implemented in Matlab (the Mathworks, USA) using Psychophysics Toolbox 3.0 extensions[11].

B. Patient and ECoG Recordings

The ECoG data were collected from 5 patients (4 male, 1 female, from 12~30 years old) who suffered from intractable epilepsy and underwent temporary placement of ECoG electrode arrays to localize seizure foci prior to surgical resection. Prior to the implantation of electrodes, the patient gave written informed consent for his involvement in research. The experiments were carried out during stable interictal

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

periods. No seizure had been observed 2 hour before or after the tests in any of the patients. For each patient, around 64 surface electrodes were implanted. The detailed demographic and clinical information, including the placement of electrodes is shown in Table I. The study was approved by the Research Ethics Committee of Tsinghua University and the affiliated Yuanquan Hospital.

C. Data Preprocessing and Feature Extraction

First, the continuous speech materials of narrative stories were segmented into syllables manually by three students with phonetics training. The onsets of syllables were marked (Fig.1), and extracted into trials of segments. Syllables that appear after a silent period lasting over 1s following the previous syllable were removed to avoid the effect of stimuli onset.

Then, the trials were decomposed into time-frequency atoms using Morlet wavelets by the Fieldtrip Toolbox[12]. The analyzed frequency band was chosen as 40~140Hz for the gamma and high-gamma broadband activities[13]. To amplify the high-frequency activities and overcome the 1/f effect in ECoG signal[14], we normalized each frequency bin for all trials, and then averaged the normalized wavelet coefficients over this frequency band to reconstruct frequency normalized signals.

The ECoG features for each trial were extracted from these frequency normalized signals by energy in time windows. Due to the ambiguous onset of each Chinese phonemes, 100ms window width and 50ms interval of centers, from onset of syllable to 400ms after the onset were chosen as features for selectivity index and classification.

TABLE I. PATIENT INFORMATION

Subject No.	Age, Sex	Hand	Grid
1	13,M	R	L temp,par,occ
2	22,M	R	L fr, temp
3	12,M	R	L temp,occ
4	30,F	R	L temp,par
5	22,M	R	R temp,occ

fr = frontal, par=parietal, temp=temporal, occ=occipital

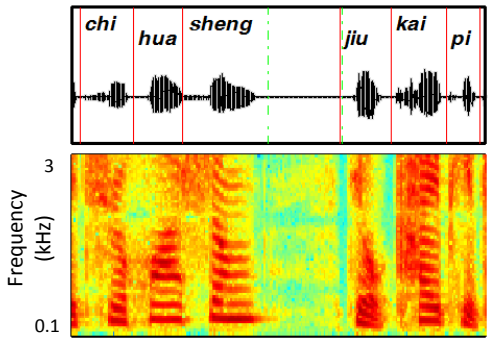

Fig.1 Segmentation of syllables in continuous speech

D. Feature Selection using Selectivity Index (SI)

The Selectivity Index was adopted to reflect the discriminability between different Chinese phonemes of a single feature. The number of pairs that passed the Wilcoxon rank-sum test was defined as selectivity index for each feature[7]. The features with SI higher than the manually selected threshold were collected for further phoneme classification. Currently, the SI threshold was selected to ensure that around 5~10 electrodes per subject had been selected and over 60% of these electrodes were located in the posterior temporal and inferior parietal areas through inspection of co-registrated cortex and electrodes using FreeSurfer[15], by merging the CT and MRI images[16]. To focus on neural activity over human speech perception areas, features in those electrodes outside posterior superior temporal and inferior parietal areas were excluded from the following analysis.

E. Selection of Chinese Phoneme Clusters for Classification

As mentioned before, a Chinese phoneme consists of three parts: 'head', 'body' and 'tail'. Categorization of a typical Chinese phoneme was performed according to the phonological distinctive features of phonemes in these parts.

For single electrode and multiple electrodes, the feature space of ECoG signals for each Chinese phoneme were also visualized using multidimensional scaling (MDS) method[17], which enabled us to find the clusters of Chinese phonemes with similar phonetic features, i.e. manner of articulation or place of articulation. Classification tasks were selected by inspection of MDS separable clusters.

F. SVM Classification

SVM is appropriate for classification problems with small sample size, thus suitable for our classification task. For each run of classification, the redundant samples of some categories were randomly chosen to be discarded for the balance of dataset. Also, original features were averaged for every 12 trials to reduce the variance due to the sparseness of neural response under continuous condition[18], generating approximately 100~200 trials for each category. Radial basis was used for SVM classification using LibSVM[19] and a simple grid search of best parameters for C and gamma was performed. Classifications had been performed 200 times on randomly balanced datasets to achieve stable accuracy, 90% for training and 10% for testing at each time after balancing. We then shuffled the labels of all trials and repeated the classification analysis as the permutation test of these accuracies.

III. RESULTS

The wavelet time-frequency responses exhibited activations and inhibitions of high frequency band between 50ms to 400ms post stimulus onset. Fig.2a and 2b shows two typical response patterns of single Chinese phonemes. For the gamma and high gamma broadband responses, frequency normalized signals shows smoother and often stronger

response than the band-pass filtered envelope signals, illustrated in Fig2c and 2d.

In Fig3a, electrodes from subject 1 with SI above threshold are shown. All circled electrodes have an SI above 170, i.e. over 85 pairs of Chinese phonemes are separable on any of these electrodes. Among these electrodes, those marked in red circles were utilized in classification because their anatomical location over speech areas, such as the posterior temporal and inferior parietal cortex. The detailed patterns of response intensities are illustrated in Fig3b and 3c, which indicates a tendency of phoneme clustering. The response intensities stand for the mean values of energy in the 100~200ms time window. Responses of electrode 1 to /en/,/in/,/an/ and /van/ are similar, indicating a selectivity to nasals (/n/) in the tail. Besides, similarity of responses to /ua/,/van/,/ang/,/iang/ and /uang/ on electrode 2 indicates a selectivity to the vowel /a/ in the body of a Chinese phoneme.

Using MDS on the features selected by SI on each subject, we have revealed the representation of multiple Chinese phonemes in the ECoG feature space. Results from subject 2 are shown in Fig.4. Colors are assigned to different phonological clusters of Chinese phonemes for inspection of separable Chinese phoneme clusters, and the Chinese phoneme /o/ was removed from MDS calculation due to extreme value in small sample size (6 trials). In Fig.4a, Chinese phonemes started with /a/,/u/,/o/,/e/ (low/back phonemes) are labeled in red, and those started with /i/,/v/ (high&front phonemes) are labeled in blue. It is the difference between place of articulation, mainly the high-front and low-back places of tongue, that separates these phoneme clusters. Meanwhile, the distinction between nasals (/n/, /ng/) and non-nasals is clear in Fig.4b. To sum up, 3 pre-defined types of classification tasks were selected by inspection into these results, containing low-back vs. high-front in the head, low-back vs. high-front in the body, and nasal vs. non-nasal in the tail.

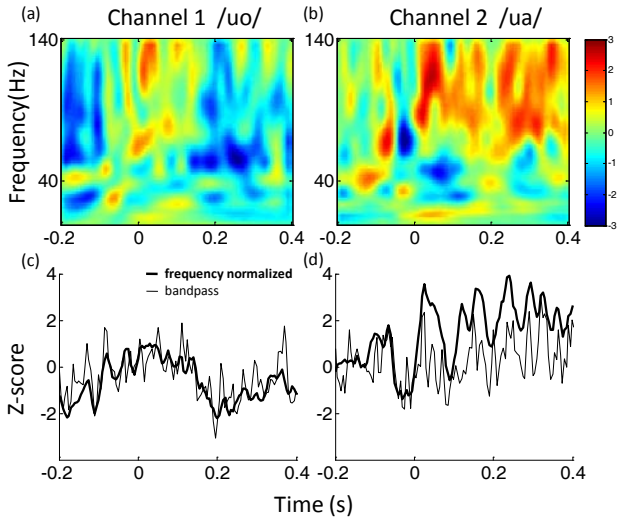

Fig.2 (a)Time-frequency responses to single Chinese phonemes (b) Difference between bandpass filtered envelope and reconstructed frequency normalized signal

SVM classifiers achieved above chance accuracy. Due to the sparseness of electrode coverage, we did not expect each type of Chinese phoneme clusters separable in each subject. Instead, the best classification result of the 3 types of classification is shown in Fig.5. All results shown were significant ($p<0.001$, permutation test). The mean accuracy reached 60.47% and maximum reached 65.19%. In retrospect, the locations of electrodes selected in the best subject are shown in Fig.5b.

IV. Discussion and Conclusion

The ECoG signals during speech perception contain rich spectral and temporal information. Especially, the gamma and high-gamma local field potential (LFP) (>40Hz) has been reported to correlate with spiking activities[20]. Therefore, the selective ECoG responses to specific phonemes we found might link to the single-unit recording in human auditory cortex[8].

Electrodes with high SIs, i.e. discriminative to Chinese phonemes, were mostly located around the posterior superior temporal and inferior parietal areas, consistent with previous results[21]. However, SI also picks some electrodes outside these areas. Some of them may also contain information for classification, which are under further investigation.

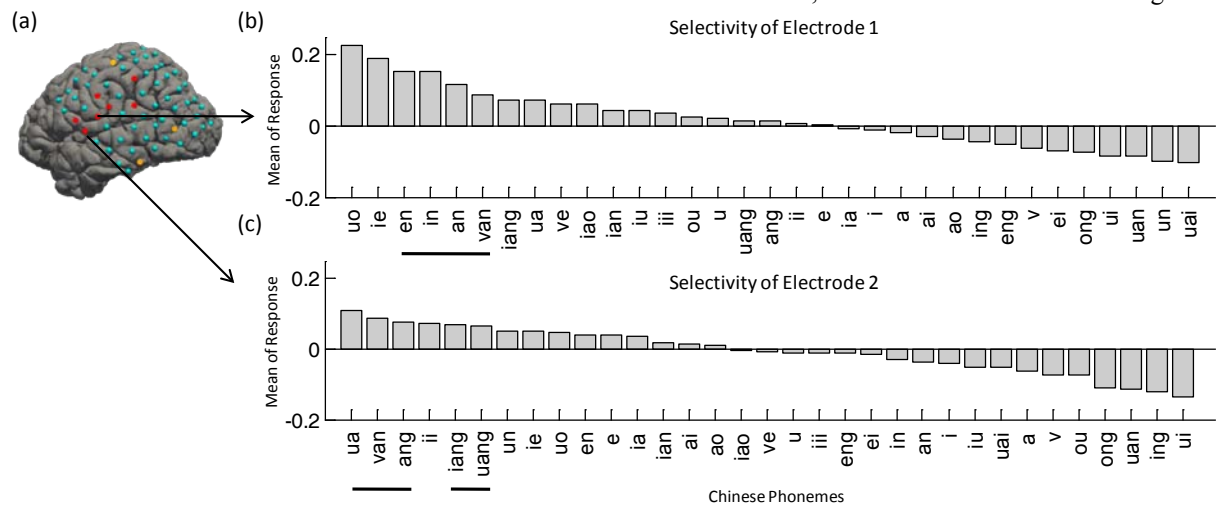

Fig.3 Feature selection using selectivity index. (a) Red and orange circles represent the electrodes selected by SI, subject 1. Red circles for electrodes posterior superior temporal and inferior parietal areas (b, c)Selectivity of electrodes to different Chinese phonemes

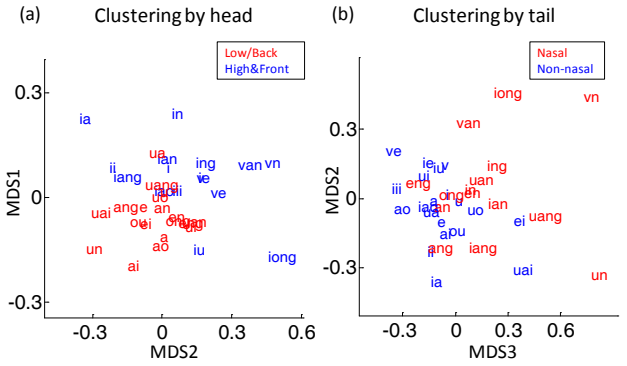

Fig. 4 MDS representation of Chinese phonemes in the ECoG feature space, subject 2

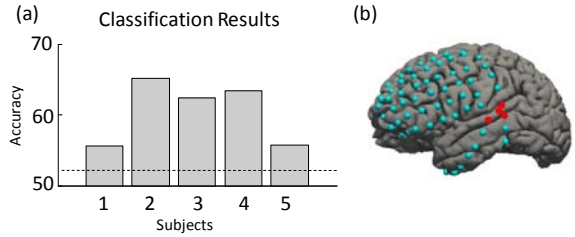

Fig. 5 Classification results of Chinese phoneme clusters. (a) classification accuracies. mean accuracy=60.47%, best accuracy=65.19%. (b)locations of electrodes for classification in the best subject

For the young patients, it is reported that 6 to 8-year-old Chinese first graders were able to complete phoneme-level tasks[22]. Their phoneme recognition system should be somehow developed at 12~13 years old. Besides, all of our subjects can normally communicate with experimenters.

The improvement of classification performance indicates that a clustered organization exists in the neural representation of Chinese phonemes revealed by ECoG activity. More importantly, the neural organization of Chinese phonemes is consistent with phonologically defined categories, which was also found in the neural encoding of English speech [7, 8].

In conclusion, spectrotemporal features of ECoG signal revealed a similar clustered organization of Chinese phonemes as compared with English, which can be used to improve classification accuracy of phonemes for speech BCI systems.

ACKNOWLEDGMENT

The authors wish to thank Dr. WenJing Zhou and his neurosurgery team at affiliated Yuquan Hospital, Tsinghua University and MRI imaging team of Tsinghua Biomedical Imaging Center.

REFERENCES

[1] J. Obleser, A. M. Leaver, J. Vanmeter, and J. P. Rauschecker, "Segregation of vowels and consonants in human auditory cortex: evidence for distributed hierarchical organization," *Front Psychol,* vol. 1, pp. 232, 2010.

[2] T. Blakely, K. J. Miller, R. P. Rao, M. D. Holmes, and J. G. Ojemann, "Localization and classification of phonemes using high spatial resolution electrocorticography (ECoG) grids," *EMBS 2008.* pp. 4964-4967, 2008.

[3] X. Pei, D. L. Barbour, E. C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans," *Journal of neural engineering,* vol. 8, no. 4, pp. 046028, 2011.

[4] G. Toyoda, E. C. Brown, N. Matsuzaki, K. Kojima, M. Nishida, and E. Asano, "Electrocorticographic correlates of overt articulation of 44 English phonemes: Intracranial recording in children with focal epilepsy," *Clin Neurophysiol*, Nov 19, 2013.

[5] Z. Jialu, "The distinctive feature trees of Standard Chinese (Putonghua)[J]," *Acta Acustica,* vol. 3, pp. 001, 2006.

[6] H. M. Meng, and V. W. Zue, "Signal representation comparison for phonetic classification." pp. 285-288.

[7] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science,* vol. 343, no. 6174, pp. 1006-10, Feb 28, 2014.

[8] A. Tankus, I. Fried, and S. Shoham, "Structured neuronal encoding and decoding of human speech features," *Nature Communications,* vol. 3, pp. 1015, 2012.

[9] W. You, Q. Zhang, and R. G. Verdonschot, "Masked syllable priming effects in word and picture naming in Chinese," *PloS one,* vol. 7, no. 10, pp. e46595, 2012.

[10] Q. Qu, M. F. Damian, and N. Kazanina, "Sound-sized segments are significant for Mandarin speakers," *Proceedings of the National Academy of Sciences,* vol. 109, no. 35, pp. 14265-14270, 2012.

[11] D. H. Brainard, "The psychophysics toolbox," *Spatial vision,* vol. 10, no. 4, pp. 433-436, 1997.

[12] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational intelligence and neuroscience,* vol. 2011, pp. 1, 2011.

[13] N. E. Crone, A. Sinai, and A. Korzeniewska, "High-frequency gamma oscillations and human brain mapping with electrocorticography," *Prog Brain Res,* vol. 159, pp. 275-95, 2006.

[14] K. J. Miller, L. B. Sorensen, J. G. Ojemann, and M. den Nijs, "Power-law scaling in the brain surface electric potential," *PLoS Comput Biol,* vol. 5, no. 12, pp. e1000609, Dec, 2009.

[15] B. Fischl, "FreeSurfer," *Neuroimage,* vol. 62, no. 2, pp. 774-781, 2012.

[16] D. Zhang, H. Song, R. Xu, W. Zhou, Z. Ling, and B. Hong, "Toward a minimally invasive brain-computer interface using a single subdural channel: A visual speller study," *Neuroimage,* vol. 71C, pp. 30-41, Jan 10, 2013.

[17] E. F. Chang, J. W. Rieger, K. Johnson, M. S. Berger, N. M. Barbaro, and R. T. Knight, "Categorical speech representation in human superior temporal gyrus," *Nat Neurosci,* vol. 13, no. 11, pp. 1428-32, Nov, 2010.

[18] M. Dastjerdi, M. Ozker, B. L. Foster, V. Rangarajan, and J. Parvizi, "Numerical processing in the human parietal cortex during experimental and natural conditions," *Nat Commun,* vol. 4, pp. 2528, 2013.

[19] C.-C. Chang, and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, no. 3, pp. 27, 2011.

[20] A. Belitski, A. Gretton, C. Magri, Y. Murayama, M. A. Montemurro, N. K. Logothetis, and S. Panzeri, "Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information," *The Journal of Neuroscience,* vol. 28, no. 22, pp. 5696-5709, 2008.

[21] S. Molholm, M. R. Mercier, E. Liebenthal, T. H. Schwartz, W. Ritter, J. J. Foxe, and P. De Sanctis, "Mapping phonemic processing zones along human perisylvian cortex: an electro-corticographic investigation," *Brain Struct Funct*, May 26, 2013.

[22] E. H. Newman, T. Tardif, J. Huang, and H. Shu, "Phonemes matter: the role of phoneme-level awareness in emergent Chinese readers," *J Exp Child Psychol,* vol. 108, no. 2, pp. 242-59, Feb, 2011.