Deciphering phonemes from syllables in BOLD signals in human superior temporal gyrus

<table>
<thead>
<tr>
<th>Journal:</th>
<th>European Journal of Neuroscience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>EJN-2015-11-23242(R).R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Research Report</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Zhang, Qingtian; Tsinghua University, Department of Computer Science and Technology  
Hu, Xiaolin; Tsinghua University, Department of Computer Science and Technology  
Luo, Huan; Peking University, Department of Psychology  
Li, Jianmin; Tsinghua University, Department of Computer Science and Technology  
Zhang, Xiaolu; Tsinghua University, Department of Computer Science and Technology  
Zhang, Bo; Tsinghua University, Department of Computer Science and Technology |
| Key Words: | speech, auditory processing, fMRI, pattern classification |
Title: Deciphering phonemes from syllables in BOLD signals in human superior temporal gyrus

Running title: Phoneme encoding in human STG

Key words: speech perception, auditory processing, fMRI, multivariate analysis, pattern classification

Author affiliation:

Qingtian Zhang, 1 Xiaolin Hu, 1,2 Huan Luo, 3,4 Jianmin Li, 1 Xiaolu Zhang, 1 Bo Zhang 1,2
1Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
2Center for Brain-Inspired Computing Research(CBICR), Tsinghua University, Beijing 100084, China
3Department of Psychology, Peking University, Beijing 100871, China
4IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

Corresponding author:

Xiaolin Hu
Room 4-504, FIT Building, Tsinghua University, Beijing 100084, China
Tel: +86-10-62795869
Email: xlhu@tsinghua.edu.cn
Website: www.xlhu.cn

Number of pages: 25

Number of figures: 7

Number of tables: 0

Number of equations: 1

Number of words in Abstract: 195 (maximum 250)

Number of words in Introduction: 500 (maximum 500)

Number of words in the whole manuscript: 7130 (excluding figures)

The authors declare no competing interests.
Deciphering phonemes from syllables in BOLD signals in human superior temporal gyrus

Qingtian Zhang,1 Xiaolin Hu,1,2 Huan Luo,3,4 Jianmin Li,1 Xiaolu Zhang,1 Bo Zhang1,2
1 Tsinghua National Laboratory for Information Science and Technology (TNList), Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
2 Center for Brain-Inspired Computing Research (CBICR), Tsinghua University, Beijing 100084, China
3 Department of Psychology, Peking University, Beijing 100871, China
4 IDG/McGovern Institute for Brain Research, Peking University, Beijing 100871, China

Abstract

Linguistic units such as phonemes and syllables are important for speech perception. How the brain encodes these units is not well understood. Many neuroimaging studies have found distinct representations of consonant-vowel (CV) syllables that shared one phoneme and differed in the other phoneme (e.g., /ba/ and /da/), but it is unclear whether this discrimination ability is due to the neural coding of phonemes or syllables. We combined functional magnetic resonance imaging (fMRI) with multivariate pattern analysis to explore this question. Subjects listened to nine mandarin syllables in a consonant-vowel (CV) form. We successfully decoded phonemes from the syllables based on the blood oxygenation level dependent (BOLD) signals in the superior temporal gyrus (STG). Specifically, a classifier trained on the cortical patterns elicited by a set of syllables, which contained two phonemes, could distinguish the cortical patterns elicited by other syllables that contained the two phonemes. The results indicated that phonemes have unique representations in the STG. In addition, there was a categorical effect: the cortical patterns of consonants were similar, and so were the cortical patterns of vowels. Further analysis showed that phonemes exhibited stronger encoding specificity in the mid-STG than in the anterior STG.
Introduction

Speech perception is one of the most important skills for human beings. We can recognize the speech rapidly and accurately despite dramatic variations of speaker identity, intensity, occlusion and noise. This ability is supported by our speech perception system which can extract relevant features from the speech (McClelland, 1986; Belin et al., 2004; Formisano et al., 2008; Noris and McQueen, 2008). It is thought that the auditory pathway processes acoustic features, phonemes, syllables and some more complex patterns in a hierarchical fashion, but whether each level of these linguistic units is encoded and how they are organized in the brain remains unclear (Hickok and Poeppel, 2007).

Several electrophysiological studies in patients (Chang et al., 2010; Pasley et al., 2012; Mesgarani et al., 2014) have identified representations of phonetic features in the brain. Complementing these results, many neuroimaging studies in normal humans have also revealed evidence of phoneme representations in the brain, which has primarily originated from two categories of studies. The first category of studies used single phonemes as stimuli and identified distinct cortical activity patterns associated with these stimuli (Obleser et al., 2006; Formisano et al., 2008; Bonte et al., 2014; Kim et al., 2014; Arsenault and Buchsbaum, 2015). One limitation of these studies is that only vowels or consonants were investigated. With single phonemes, it is impossible to identify how phoneme representations are integrated into syllable representations. The second category of studies used syllables as stimuli that comprised a few phonemes, and the most often used syllables were consonant-vowel (CV) syllables that shared one phoneme and differed in the other phoneme (e.g., /ba/, /da/ and /ta/). An fMRI-adaptation experiment suggested that /ga/ and /da/ had different representations in the temporal and parietal cortices (Joanisse et al., 2007). Multivariate pattern analyses have also revealed different cortical activities elicited by CV syllables (Raizada et al., 2010; Du et al., 2014). However, these studies cannot determine whether the discrimination ability is a result of the neural coding
of phonemes or syllables because a syllable-based representation can lead to the same results (Fig. 1A). A recent Electroencephalography (EEG) study (Wang et al., 2012) supports phoneme representation by showing that a classifier can be trained to distinguish two phonemes (e.g., /p/ versus /t/) with CV syllables that share different phonemes (e.g., /i/, /æ/, /u/ and /a/); however, since EEG has limited spatial resolution, it is difficult to identify the encoding locations in the cortex.

We used fMRI and multivariate analysis to investigate the representations of phonemes in human superior temporal gyrus (STG). The blood-oxygen-level-dependent (BOLD) responses elicited by Chinese mandarin consonant-vowel (CV) syllables (Fig. 1B) were successfully classified into different groups. By carefully manipulating the training set and test set we could judge whether the phoneme-based representation hypothesis or the syllable-based representation hypothesis is in accordance with the classification results (Fig. 1A). If the former holds, by characterizing the responses to the stimuli, we could also delineate the spatial patterns in the cortex that encode phonemes.

**Materials and Methods**

**Subjects**

Thirteen (median age = 25, all males) right-handed, healthy native Chinese adults providing written informed consent participated in the study. All participants had normal hearing and no neurological abnormalities. The study was approved by the ethics committee of the Institute of Biophysics at Chinese Academy of Sciences, Beijing, China, and conformed with the World Medical Association Declaration of Helsinki.

**Stimuli**

Nine CV syllables were recorded from a native Chinese mandarin speaker. Each syllable started with a
consonant /b/, /d/, or /m/ and ended with a vowel /a/, /i/, or /u/ (Fig. 1B). Each syllable corresponded to at least one Chinese word, and the first tone (yin ping) was pronounced. The design of stimuli was similar to that in (Formisano et al., 2008). In that study the two variables in the set of stimuli were vowel and speaker identity, but in this study the two variables were vowel and consonant. To introduce the acoustic variability, three different tokens for each syllable were included. The stimuli sampling rate was 44.1 kHz, saved with 16 bit resolution and 560 ms long. The stimuli were decorated with 100 ms linear onset and offset ramps to avoid a sharp cut-off and their average root mean square values were modulated to be equal (Formisano et al., 2008). The spectrograms of the nine syllable stimuli are shown in Fig. 1B. Prior to functional imaging, a behavioural experiment was conducted, and all subjects were able to associate each sound with the corresponding syllable easily and correctly. During the fMRI sessions, subjects were instructed to attentively listen to the sounds. After the fMRI sessions, subjects were debriefed, and all reported that they could distinctly hear and recognize all of the stimuli.

Data acquisition and preprocessing

Functional magnetic resonance images were collected on a SIEMENS Symphony 3 Tesla system at Beijing MRI Center for Brain Research. For each subject, anatomical images covering the whole brain were obtained using the T1-weighted sequence, and then two high-spatial resolution (1.5 mm × 1.5 mm × 2 mm) functional runs (360 volumes) were collected using a standard echo-planar imaging (EPI) sequence (repetition time [TR]=3.6 s; acquisition time [TA]=2.0 s; echo time [TE]=30 ms). The subjects lay supine in the MRI chamber and were instructed to remain the head as still as possible. The head was stabilised by padded restraints. Each brain volume consisted of 28 slices that covered the entire temporal lobe.

During the measurements, subjects listened to stimuli that were presented binaurally in the 1.6 s silent
gaps between two consecutive volume acquisitions (Hall et al., 1999; Joanisse et al., 2007; Formisano et al., 2008). There were two 520 ms silences between the EPI noise, the acoustic stimuli and the next EPI noise. Subjects wore headphones that attenuated the EPI noise to an acceptable level and allowed for stimulus presentation at a level perceived as higher than that of the preceding EPI noise. These measures made the acoustic stimuli and the scanner noise separated clearly in both spectral and temporal domain and ensured a clear perception of different syllables.

According to a slow event-related design, the average inter-trial-interval between two stimuli was 14.4 s (range 10.8~18 s, i.e., 3~5 TR). Each of the two functional runs included 10 trials per stimulus condition (90 trials per run), resulting in a total of 20 trials per stimulus condition. The sequence of stimuli was pseudo-randomised.

Functional and anatomical images were analysed with SPM 8, which is available at www.fil.ion.ucl.ac.uk/spm/. Preprocessing consisted of realignment, co-registration, segmentation and normalisation. The realignment corrected the 3-dimensional motion. Then, the functional images were co-registered to the anatomical image. After segmenting the anatomical image into different brain functional areas, both the functional and anatomical data were normalised to the standard Talairach space (Talairach and Tournoux, 1988).

Univariate analysis

For each subject’s data, general linear models (GLM) were applied to identify activated voxels. Brain activation maps were generated from the GLM in which the BOLD signal was the dependent variable and the stimulus paradigm was the covariate. Then we formed the design matrix for the nine stimulus conditions using the hemodynamic response function (HRF)-adjusted time course predictors (Formisano et al., 2008).
The contrast analysis between different phoneme conditions did not show any significant effect (q=0.05, corrected for multiple comparisons with the false discovery rate (FDR) (Genovese et al., 2002) (Fig. 2).

Data preparation for multivariate pattern classification

Before the multivariate analysis was applied, the preprocessed brain images were divided into trials for every presentation. Same as in (Formisano et al., 2008), we defined the “responses” of voxels for each trial as the regressor coefficients in the GLM based on HRF. The response patterns were labelled by their corresponding syllables and classified into three consonant conditions (/b/, /d/ and /m/) and three vowel conditions (/a/, /i/ and /u/). Each condition contained the responses of three syllables which shared this phoneme, e.g., the /b/ condition contained the responses of /ba/, /bi/ and /bu/ and the /a/ condition contained the responses of /ba/, /da/ and /ma/.

Each volume of brain data included tens of thousands of voxels that had to be reduced for multivariate pattern classification. Two approaches were used in different analyses. The first approach was based on the permutation test (Nichols and Homes, 2002; Mourão-Miranda et al., 2005; Wang et al., 2007; Zhang et al., 2015), which is a general method for discovering relevant features for classification. This approach was used to obtain results in Figs. 4 and 5. For classifying any two phonemes, the labels of the samples in the training set were randomly permutated 1000 times, and accordingly 1000 standard (L2-norm) linear support vector machines (SVMs) (Burges, 1998) were trained with these labels (the control parameter $C$ was always set to 1). For each voxel, we estimated a Gaussian distribution to fit the linear weights in the 1000 classifiers (Kolmogorov-Smirnov test, p<0.001). Then we tested the null hypothesis that the weights were irrelevant to the true class label. If the weight learned with true labels was significantly different from the mean of the distribution of weights (one sample t-test), the null hypothesis was rejected and this voxel was treated as
“active”. Different p values will lead to different number of “active” voxels (smaller p, fewer voxels). By adjusting p (p<0.01) approximately 500 “active” voxels were selected each time for further analysis.

The second approach was based on the t-test for an irrelevant phoneme condition, which was only used to test a hypothesis that consonants (or vowels) have similar representations in the cortex (Fig. 6C,D). There were six phoneme pairs: three consonant pairs (/b/ versus /d/, /b/ versus /m/, /d/ versus /m/) and three vowel pairs (/a/ versus /i/, /a/ versus /u/, /i/ versus /u/). For each phoneme pair (e.g., /a/ versus /i/), there was always a phoneme (e.g., /u/) that did not appear in any of the six syllables used for classification. This phoneme condition consisted of three syllables (e.g., /bu/, /di/ and /mu/) and was used to select “active” voxels with t-test for the phoneme pair (e.g., /a/ versus /i/), each consisted of three syllables, say /ba/, /da/, /ma/ versus /bi/, /di/, /mi/). By adjusting the p value (p<0.05) approximately 500 “active” voxels were selected each time for further analysis.

To prevent overfitting, it is important to exclude the test samples in subsequent classification for selecting “active” voxels (De Martino et al., 2008). This was ensured by first splitting the samples into training set and test set, then performing voxel selection and classifier learning on the training set (Fig. 3).

Pattern classification

For each phoneme pair, after selecting “active” voxels, we performed pattern classification using a machine learning algorithm, the L1-norm SVM (Fung and Mangasarian, 2004). Given a set of training data \( \{x_i, y_i\}, i = 1, 2, \ldots, m \), where \( x_i \in \mathbb{R}^n \), \( y_i \in \{-1, 1\} \), the objective of the linear L1-norm SVM is to find a linear function \( f(x) = w^T x + b \) as the hyperplane that matches as more as possible training data, that is, \( \text{sign}(f(x_i)) = y_i \), and which could be satisfactorily generalised to unseen data. Training the classifier amounts to solving a constrained optimisation problem,
\[
\begin{align*}
\text{minimize} & \quad \|w\|_1 + C \sum_{i=1}^{m} \xi_i \\
\text{subject to} & \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, 2, \ldots, m
\end{align*}
\]

where \(C > 0\) is a constant, which was set to 0.1 in all experiments. This formulation differs from the standard (L2-norm) SVM in the first term in the objective function, which is the L1-norm of \(w\), while in standard SVM this term is \(\frac{1}{2}\|w\|_2^2\). We will discuss the advantage of the L1-norm term later.

Inspired by (Formisano et al., 2008), two classification schemes were investigated, which differed in the selection of training samples and test samples. In the discrimination analysis, a six-fold cross-validation was performed for each phoneme pair for each subject. The 120 trials from two phoneme conditions were randomly divided into six equal parts. For each validation, one part of the trials was used for testing, and the other parts were used for training. The accuracy of the subject for this phoneme pair was defined as the average of the six test accuracies. In the generalisation analysis, a three-fold cross-validation was performed for each phoneme pair for each subject. The 120 trials from two phoneme conditions were divided into three equal parts, in which each part consisted of trials of two syllables that contained the two phonemes under study and shared another phoneme. For example, consider classifying /b/ and /d/, the three parts consisted of trials of \{/ba/, /da/\}, \{/bi/, /di/\} and \{/bu/, /du/\}, respectively. In this paradigm, the classifier was trained on four syllables (e.g., /ba/, /da/, /bi/ and /di/) but tested on the other two unseen syllables (/bu/ and /du/).

The accuracy of the subject for this phoneme pair was defined as the average of the three test accuracies.

The pipeline of the analyses was illustrated in Fig. 3. It is stressed that the initial voxel selection procedure was nested in the six-fold or three-fold cross-validation scheme which ensured that the test samples were only used in the test phase.

\textit{Final voxel selection}
After initial voxel selection, there were about 500 voxels left for each classification. But there were much redundancy among the voxels. It is desirable to identify those really useful voxels for each classification. For this purpose, some previous studies (Formisano et al., 2008; De Martino et al., 2008) recursively trained L2-norm SVMs and eliminated unimportant voxels, which is called the recursive feature elimination (RFE) algorithm. In this study, we employed a simpler method—train the L1-norm SVM only once to do classification and voxel selection at the same time. Note that the L1-norm SVM in equation (2) includes a L1-norm term of the weight vector \( w \). It encourages a sparse \( w \), i.e., most weights are zero. The features corresponding to nonzero components of \( w \) are regarded as discriminative for separating two classes. Therefore, by looking for the nonzero components of \( w \) after learning the classifier, we identified the useful voxels.

In the discrimination analysis, if a voxel was selected by the L1-norm SVM at least five times in the six-fold cross-validation, it was defined as an effective voxel. This is because if we consider the classification accuracy (about 60%, shown in Results) as the probability of a voxel selected by the classifier is effective, the probability of a voxel selected by the classifiers at least five times in the six-fold cross-validation is effective would be larger than 95% \( (1 - C^5_6 \times (1 - 0.6)^5 \times 0.6 - C^5_6 \times (1 - 0.6)^6 = 0.959) \). In the generalisation analysis, if a voxel was always selected by the L1-norm SVM in the three-fold cross-validation, it was defined as an effective voxel. The probability of such a voxel being effective is \( 1 - C^3_3 \times (1 - 0.6)^3 = 93.6\% \). The union of the effective voxels across phoneme pairs was defined as a subject’s individual effective map (IEM), and the union of the IEMs across subjects was defined as the overall effective map (OEM).

**Multivariate analysis based on Pearson correlation**

Motivated by (Du et al., 2014; Linke et al., 2011), for a given region of interest (ROI), we calculated a
phoneme-specificity index (PSI) as follows. First, data from two runs were randomly split into two partitions such that each partition included all six phoneme conditions and each condition included 30 trials. Second, within the ROI, a 6-by-6 voxel-wise correlation matrix was computed between the split halves using a Pearson correlation coefficient based on the averaged responses over the 30 trials. Third, PSI was calculated as the average of diagonal elements minus average of off-diagonal elements in the correlation matrix (Du et al., 2014). Clearly, PSI measures the difference in similarity between within-category and between-category correlations. This was then averaged over 10 random partitions of data. Clearly, higher PSI value indicates more specific encoding of phonemes for an ROI.

A spherical searchlight analysis (Kriegeskorte et al., 2006) was performed, but the analysis was restricted on a trajectory within STG such that the searchlight with a sphere of 10 mm radius could roughly cover the entire STG (Fig. 7A). At each location on the trajectory, the sphere was treated as an ROI while PSI was computed as described above.

Results

Phonemes can be decoded from syllables in BOLD signals

We investigated whether the same phoneme in different syllables had a unique representation in the elicited BOLD signals. Consistent with previous studies (Fiez et al., 1995; Binder et al., 2000; Dehaene-Lambertz et al., 2006; Obleser et al., 2006; Formisano et al., 2008; Okada et al., 2010, Bonte et al., 2014; Arsenault and Buchsbaum, 2015), all syllables elicited significant responses in the superior temporal cortex, including regions in the Heschl’s gyrus, STG and superior temporal sulcus (STS) (Fig. 2). Because each phoneme condition (e.g., /b/) consisted of 60 trials for three syllables (e.g., /ba/, /bi/ and /bu/), we were able to
perform a contrast analysis between the three pairs of consonants (/b/ versus /d/, /b/ versus /m/, and /d/ versus /m/) and three pairs of vowels (/a/ versus /i/, /a/ versus /u/ and /i/ versus /u/), based on the elicited BOLD signals, with 60 samples for each phoneme for each pair. The univariate statistical analysis, however, did not show any significant difference when contrasting the phonemes (q=0.05, corrected for multiple comparison with FDR (Genovese et al., 2002); Fig. 2). We therefore adopted a multivariate pattern analysis method. Two complementary analyses (Formisano et al., 2008), called “discrimination analysis” and “generalisation analysis”, were performed, differing only in the selection of training samples. In each analysis, an L1-norm SVM (Fung and Mangasarian, 2004) was trained for each phoneme pair on the initially selected voxels using a permutation test.

In the discrimination analysis, a six-fold cross-validation was performed for each phoneme pair and each subject. All subjects’ accuracies for any pair of phonemes were higher than chance, 50% (p<10^{-4}) (Fig. 4A). The highest and lowest mean accuracies across all subjects were achieved for /a/ versus /u/ (61.35%) and /d/ versus /m/ (59.10%), respectively. The mean accuracy for vowel pairs (60.98%) was slightly higher than that for consonant pairs (59.24%) (p<0.01).

In the generalisation analysis, a three-fold cross-validation was performed for each phoneme pair and each subject. All subjects’ accuracies for any pair of phonemes were higher than chance (p<10^{-4}) (Fig. 4B). The highest and lowest mean accuracies across all subjects were achieved for /a/ versus /i/ (61.41%) and /b/ versus /d/ (58.08%), respectively. The mean accuracy for vowel pairs (61.22%) was also higher than that for consonant pairs (59.10%) (p<0.01).

These results show that syllables sharing a phoneme, either a consonant or a vowel, also share representations in the brain. This extends previous findings (Engineer et al., 2008; Mesgarani et al., 2008, 2014) demonstrating that phonemes sharing the same acoustic feature elicit similar cortical responses.
It should be mentioned that our average accuracy of the six phoneme pairs (about 60%) is lower than that of three vowel pairs in (Formisano et al., 2008) which is about 66%, but higher than that of three vowel pairs in (Bonte et al., 2014), which is about 55%. In fact, in the latter work, the /a/-/i/ pair could not be reliably classified (accuracy about 52%, p>0.05). Many factors may contribute to this discrepancy, e.g., different stimuli and different analysis methods.

**Spatial patterns on the temporal lobe**

Similar to (Formisano et al., 2008), we investigated the layout and consistency of the spatial patterns in the temporal lobe across subjects, which allows this decoding. The critical task is to select effective voxels for classification. To this end, our investigation was based on the feature selection ability of L1-norm SVM (Fung and Mangasarian, 2004). The effective voxels of one subject in the cortex are presented in Fig. 5A,B. We then defined the set of effective voxels for each subject as the individual effective map (IEM) and the set of effective voxels for all subjects as the overall effective map (OEM). The OEM resulting from either discrimination analysis or generalisation analysis was widely distributed bilaterally in STG/STS (Fig. 5C,D). Among the OEM voxels, a large portion (approximately 1/3–1/2), mainly in the STG, was selected in at least five subjects.

We were interested in the differences between the spatial patterns for consonant and vowel representations. Two complementary approaches were used. First, we quantified the similarity of spatial patterns in the cortex between the six phoneme pairs (Fig. 6A,B). The phoneme pairs were split into two categories, consonant pairs and vowel pairs. Higher similarities were observed within each category (top-left and bottom-right regions in the similarity matrices) than between categories (top-right and bottom-left regions in the similarity matrices), indicating a clustering effect of consonant and vowel representations.
Second, instead of using a permutation test to initially select voxels, we employed a univariate t-test to select the most responsive voxels elicited by one consonant (or vowel) condition and performed discrimination analysis and generalisation analysis to distinguish the other two consonant (or vowel) conditions. The hypothesis was that if consonants (or vowels) have similar representations in the STG, the voxels activated by one consonant (or vowel) would encode much of the information about the other consonants (or vowels), and therefore the classifier trained on these voxels would correctly distinguish the other two consonants (or vowels). This hypothesis was validated in our analysis (Fig. 6C,D), although the average classification accuracies were not as high as those using the permutation test as the initial voxel selection method (Fig. 4). Taken together, different consonants and different vowels exhibited similar representations in the STG, but consonants and vowels exhibited different representations. This is generally consistent with a magnetoencephalographic experiment (Mesgarani et al., 2014) in which electrodes responded differentially to consonants and vowels.

**Phoneme encoding at different locations**

Finally, we investigated phoneme encoding at different locations in STG using a multivariate analysis based on voxel-wise correlations (Linke et al., 2011; Du et al., 2014). A spherical searchlight analysis (Kriegeskorte et al., 2006) with a sphere of 10 mm radius was performed (Fig. 7A). At each location, a phoneme-specificity index (PSI) was calculated. Higher PSI at a location indicates more specific encoding of phonemes at this location. There were two observations (Fig. 7B). First, the mean PSI were higher in the left hemisphere than in the right hemisphere when the values along the trajectory were combined together (p=0.0007). Second, from posterior to anterior STG in both hemispheres the mean PSI first increased then decreased.

To make detailed comparison between different regions in STG, four typical locations on both
hemispheres were selected from the searchlight trajectory at roughly equal distances (Fig. 7A) for further analysis. Locations 2 and 3 were at the mid-STG (lateral area of Heschl’s gyrus) (Dewitt and Rauschecker, 2012), and locations 1 and 4 were posterior and anterior to the mid-STG, respectively. By varying the ROI radius from 6 mm to 16 mm we analysed differences between locations 2 and 4 (mid- versus anterior), locations 3 and 1 (mid- versus posterior), and locations 1 and 4 (posterior versus anterior) in both hemispheres together. For all ROI sizes tested, the mean PSI was higher at location 2 than at location 4 (p<0.05) (Fig. 7C). But for most ROI sizes no significant difference in mean PSI was observed between location 1 and location 3 (Fig. 7D). These results indicate that the mid-STG encodes more phoneme information than the anterior STG.

Discussion

In this study, we used fMRI and machine learning techniques to investigate whether and how phonemes are encoded in the brain. Three major findings are presented. First, phonemes can be decoded from syllables in cortical activities in the human STG. Second, consonants and vowels exhibit distinct spatial patterns in STG. Third, phoneme processing is strongly associated with the mid-STG compared to the anterior STG.

Previous fMRI studies (Joanisse et al., 2007; Raizada et al., 2010; Du et al., 2014) using syllables as stimuli for investigating phoneme encoding in the brain assumed that different cortical activities elicited by two syllables that differ only in one phoneme would be due to different encodings of these phonemes. This assumption has excluded the possibility that these different cortical activities may be due to the encoding of syllables as a whole, considering that both hypotheses would be supported by the same results in those studies (Fig. 1A). However, direct evidence to support this assumption is scarce. This study, together with other studies (Wang et al., 2012), suggests that phonemes have unique representations in the brain by
showing that a robust separating plane can be established between two sets of syllables, while each set of syllables contains a phoneme under study (Fig. 4A). The more convincing evidence is that a classifier trained on certain CV syllables for discrimination of two phonemes was able to classify novel CV syllables containing those phonemes (Fig. 4B). This capability of the classifier must be due to the encoding of phonemes because the test syllables did not appear in the training set (Fig. 1A). These results confirmed the default assumption made in previous studies (Joanisse et al., 2007; Raizada et al., 2010; Du et al., 2014).

Our analysis of different phoneme pairs yielded significant differences. First, the average test accuracy over all vowel pairs was higher than that over all consonant pairs (Fig. 4 and Fig. 6C,D). This is likely due to the different durations of the vowels (which lasted at least 400 ms) and consonants (which lasted at most 60 ms) used in the experiment. Vowels elicited stronger BOLD signals than consonants, making the former easier to distinguish. Second, the spatial patterns of effective voxels for two consonant pairs were similar, as were the spatial patterns for two vowel pairs, but the spatial patterns for a consonant pair and a vowel pair were dissimilar on average (Fig. 6A,B). A different classification scheme confirmed this observation (Fig. 6C,D). The distinct spatial patterns between vowels and consonants may underlie the selective difficulties in producing vowels and consonants found in aphasics (Caramazza et al., 2000).

There exists electrophysiological evidence of phoneme representations in the brain (Engineer et al., 2008; Mesgarani et al., 2008; Chang et al., 2010; Pasley et al., 2012; Mesgarani et al., 2014), among which the results obtained using surface electrode arrays in human brains (Chang et al., 2010; Pasley et al., 2012; Mesgarani et al., 2014) are more convincing because speech is uniquely human. With high spatial resolution and temporal resolution, these techniques have identified representations of acoustic features, speech units even smaller than phonemes. But these invasive studies in patients undergoing neurosurgical procedures for epilepsy or brain tumor cannot be used to study normal individuals and that such patients may have other
compensatory mechanisms that make them unsuitable for characterizing the normal speech processing system. The present fMRI study confirms the existence of phoneme representation in normal humans.

A hierarchical perspective on auditory processing (Rauschecker and Scott, 2009; Leaver and Rauschecker, 2010; Dewitt and Rauschecker, 2012) suggests that to translate acoustic sounds into speech, the brain must recognize acoustic features, elemental speech units (i.e., phonemes) and the concatenation of those units (i.e., syllables). According to this perspective, all of these units have representations somewhere in the brain and form a recognition pathway similar to that in the visual cortex (Riesenhuber and Poggio, 2002). This study together with other studies (Greenberg 1999, 2003, 2006; Luo and Peoppel, 2007, 2012; Engineer et al., 2008; Mesgarani et al., 2008, 2014) suggests that acoustic features, phonemes and syllables indeed have representations in the auditory cortex. How this pathway is organized in the auditory cortex remains unclear.

Our classification analysis demonstrated that the voxels effective for distinguishing phonemes were mainly located in the STG (Fig. 5). Further analysis based on Pearson correlations indicated that phonetic processing is stronger in the mid-STG than in the anterior STG (Fig. 7B,C), which is consistent with human electrocorticography studies (Chang et al., 2010, 2011) and imaging studies (Joanisse et al., 2007; Leaver and Rauschecker, 2010), which localise processing of phonemes to the mid-STG. Previous imaging studies have localised processing of word forms to the anterior STG/superior temporal sulcus (STS) (Binder et al., 2000; Marinkovic et al., 2003; Dehaene-Lambertz et al., 2006). Taken together, the results support an anterior-directed hierarchical model of speech perception (Binder et al., 2000; Dewitt and Rauschecker, 2012). Finally, our correlation analysis showed that the left STG was more involved in phoneme encoding than the right STG, which is generally consistent with many studies (Binder et al., 1997; Thierry et al., 2003; Joanisse et al., 2007; Leaver and Rauschecker, 2010; Dewitt and Rauschecker, 2012) indicating that the left hemisphere is biased for speech perception.
Acknowledgement

This work was supported in part by the National Basic Research Program (973 Program) of China (Grant Nos. 2012CB316301, 2013CB329403) and National Natural Science Foundation of China (Grant Nos. 61273023, 61332007, 91420201).

References


**Figure Legends**

**Figure 1.** Hypotheses and stimuli. A, Illustration of two hypotheses for syllable and phoneme encoding in human cortex. /AC/, /BC/, /AD/ and /BD/ denote four syllables consisting of phonemes /A/, /B/, /C/ and /D/. Without loss of generality, suppose that the syllables and phonemes are encoded by locations in the 2D plane and that the decision boundary between any two classes is a straight line. The phoneme-based representation hypothesis (top) implies that a boundary between phonemes /A/ and /B/ can be learned using samples of /AC/ and /AD/ versus samples of /BC/ and /BD/ (discrimination analysis). In particular, if the boundary is learned using samples of /AC/ versus samples of /BC/ then it can also separate samples of /AD/ versus samples of /BD/ (generalisation analysis). In contrast, the syllable-based representation hypothesis (bottom) implies that a correct decision boundary between samples of /AC/ and /AD/ versus samples of /BC/ and /BD/ may not exist. Consequently, the boundary between /AC/ and /BC/ may not separate /AD/ and /BD/. Note that in either hypothesis, a decision boundary can be learned to separate any two syllables. B, Spectrogram of the nine CV syllables used in experiments, formed by three consonants (/b/, /d/ and /m/) and three vowels (/a/, /i/ and /u/).

**Figure 2.** Results of statistical univariate analysis. Shown are activated voxels for all syllable sounds estimated with univariate general linear model analysis (t-test, single subject), visualised in coronal planes (upper) and inflated cortex
(bottom). However, only one voxel showed a statistically significant difference in the /d/ versus /m/ comparison, and no such voxels were found in other phoneme pair comparisons. L: Left hemisphere; R: right hemisphere.

**Figure 3.** Pipeline of multivariate analysis. In the discrimination analysis $n = 6$ and in the generalisation analysis $n=3$.

**Figure 4.** Results of phoneme decoding based on permutation test for selecting initial voxels. **A**, Test accuracies (median value and distribution) over all subjects in the discrimination analysis. For each phoneme pair, the mean accuracy was higher than chance (50%) (/b/ versus /d/=59.49%, $p=7 \times 10^{-4}$; /b/ versus /m/=59.17%, $p=2 \times 10^{-3}$; /d/ versus /m/=59.10%, $p=9 \times 10^{-5}$; /a/ versus /i/=60.38%, $p=1 \times 10^{-5}$; /a/ versus /u/=61.35%, $p=3 \times 10^{-6}$; /i/ versus /u/=61.22%, $p=1 \times 10^{-6}$). **B**, Test accuracies (median value and distribution) over all subjects in the generalisation analysis. For each phoneme pair, the mean accuracy was higher than chance (50%). (/b/ versus /d/=58.08%, $p=1 \times 10^{-5}$; /b/ versus /m/=58.91%, $p=2 \times 10^{-4}$; /d/ versus /m/=60.32%, $p=3 \times 10^{-5}$; /a/ versus /i/=61.41%, $p=2 \times 10^{-5}$; /a/ versus /u/=60.90%, $p=8 \times 10^{-6}$; /i/ versus /u/=61.35%, $p=4 \times 10^{-5}$). In both analyses, the mean accuracy for vowel pairs was higher than for consonant pairs ($p<0.01$).

**Figure 5.** Spatial patterns of effective voxels for classification. **A** and **B**, Subject 1’s effective voxels (orange) selected by L1-norm SVM for six phoneme pairs in the discrimination analysis and generalisation analysis, respectively. These voxels are subsets of those initially selected by the permutation test (blue). The dashed boxes in the top panel indicate the locations of these results in the brain. **C** and **D**, OEM obtained in the discrimination analysis and generalisation analysis, respectively. Blue, red and yellow indicate selection on at least one subject, three subjects and five subjects, respectively.

**Figure 6.** Properties of phoneme encoding. **A** and **B**, Similarities between spatial patterns of effective voxels identified by the classifier in discrimination analysis and generalisation analysis, respectively. For each subject, the similarity between two phoneme pairs was defined as $|V_{PP1} \cap V_{PP2}|/|V_{PP1} \cup V_{PP2}|$, where $V_{PP1}$ and $V_{PP2}$ denote the sets of effective voxels identified by the L1-norm SVM for two phoneme pairs, and $|\cdot|$ denotes the cardinality. A similarity
matrix was calculated for each subject. Shown are the average results across all subjects. C, Test accuracies (median value and distribution) over all subjects in the discrimination analysis using the t-test for selecting initial voxels. For each phoneme pair, the mean accuracy was higher than chance (50%). /b/ versus /d/=57.12%, p=1×10^{-3}; /b/ versus /m/=57.31%, p=2×10^{-4}; /d/ versus /m/=56.92%, p=1×10^{-3}; /a/ versus /i/=57.31%, p=5×10^{-4}; /a/ versus /u/=58.53%, p=2×10^{-4}; /i/ versus /u/=58.91%, p=4×10^{-4}). D, Test accuracies (median value and distribution) over all subjects for generalisation analysis using the t-test for selecting initial voxels. For each phoneme pair, the mean accuracy was higher than chance (50%). /b/ versus /d/=56.60%, p=1×10^{-5}; /b/ versus /m/=56.28%, p=1×10^{-3}; /d/ versus /m/=57.50%, p=6×10^{-4}; /a/ versus /i/=58.08%, p=2×10^{-5}; /a/ versus /u/=58.78%, p=2×10^{-5}; /i/ versus /u/=58.53%, p=1×10^{-3}). In both analyses, the mean accuracy for vowel pairs was higher than that for consonant pairs (p<0.01).

Figure 7. PSI as a function of location in the STG. A, Location of the analyses. The continues curve denotes the centre of the searchlight trajectory in STG. ROIs were spheres centred at the locations on the curve. Four typical locations on the curve were highlighted in both hemispheres with symmetric coordinates [from locations 1 to 4 (x, y, z): (±60, -34, 10), (±60, -26, 8), (±58, -17, 5), (±56, -8, 1)]. B, PSI for the ROIs with radius 10 mm as a function of searchlight location in the left hemisphere and right hemisphere, respectively, averaged over all subjects. The four locations in (A) are indicated on the horizontal axes. C, Difference of mean PSI over both hemispheres between locations 2 and 4. For all sphere radii tested, the PSI difference was larger than zero. D, Difference of mean PSI over both hemispheres between locations 3 and 1. For most sphere radii the PSI difference was not significantly different from zero. Error bar: standard deviation; asterisk: p<0.05; double asterisks: p<0.005.
Figure 1
86x48mm (300 x 300 DPI)
Figure 2
113x75mm (300 x 300 DPI)
Figure 3
68x51mm (300 x 300 DPI)
Figure 4
85x33mm (300 x 300 DPI)
Figure 5
162x102mm (300 x 300 DPI)
Figure 7
111x62mm (300 x 300 DPI)
Graphical Abstract Text

Subjects listened to nine mandarin syllables in a consonant-vowel (CV) form starting with /b/, /d/ or /m/ and ending with /a/, /i/ or /u/ while their brains were scanned in the magnetic resonance imaging (MRI) machine. Using multivariate pattern analysis techniques, we successfully decoded the consonants and vowels based on the blood oxygenation level dependent (BOLD) signals in the superior temporal gyrus (STG). The results indicated that phonemes have unique representations in the STG.
Graphical Abstract Figure
113x34mm (300 x 300 DPI)