Using phase to recognize English phonemes and their distinctive features in the brain

Rui Wang, Marcos Perreau-Guimaraes, Claudio Carvalhaes, and Patrick Suppes[1]

Center for the Study of Language and Information, Stanford University, Stanford, CA 94305

Contributed by Patrick Suppes, October 11, 2012 (sent for review June 4, 2012)

The neural mechanisms used by the human brain to identify phonemes remain unclear. We recorded the EEG signals evoked by repeated presentation of 12 American English phonemes. A support vector machine model correctly recognized a high percentage of the EEG brain wave recordings represented by their phases, which were expressed in discrete Fourier transform coefficients. We show that phases of the oscillations restricted to the frequency range of 2–9 Hz can be used to successfully recognize brain processing of these phonemes. The recognition rates can be further improved using the scalp tangential electric field and the surface Laplacian around the auditory cortical area, which were derived from the original potential signal. The best rate for the eight initial consonants was 66.7%. Moreover, we found a distinctive phase pattern in the brain for each of these consonants. We then used these phase patterns to recognize the consonants, with a correct rate of 48.7%. In addition, in the analysis of the confusion matrices, we found significant similarity–differences were invariant between brain and perceptual representations of phonemes. These latter results supported the importance of phonological distinctive features in the neural representation of phonemes.

phase synchronization | EEG classification

How the human brain processes phonemes has been a subject of interest for linguists and neuroscientists for a long time. From the beginning of the 20th century, behavioral experiments were carried out to explore the perceptual discrimination of phonemes under various conditions (1–5). Since 1978, Mismatch Negativity in EEG (6) has been used extensively to measure the neural discrimination of phonemes (7). These results suggest the existence of a language-specific phoneme representation (8). More recent findings (9–11) using magnetoencephalograph recordings support the brain's encoding of phonological features before the lexical information is retrieved. Invasive recordings of animal neural responses to human speech also exhibited temporal and spatial characteristics, reflecting the distinctive features of phonemes (12, 13). Related research has shown that animal recordings from the auditory cortex match animals' behavioral discrimination of phonemes (14) as well as the pattern of human psychological confusions (15). Functional MRI (fMRI) provides a noninvasive method to locate cortical activities of phoneme perception in healthy human brains (16). Success in recognizing fMRI evoked by isolated vowels has also been reported (17), but this work is difficult to extend because of the limited temporal resolution of fMRI.

In most EEG studies of phoneme perception, the interest has been in the temporal information rather than its spectral structure. However, it has been shown that EEG rhythms may be related to a specific cognitive state of the brain (18, 19). In the present study, we show that we can recognize, with some success, brain representations of specific phonemes using only the phase properties of brain waves in a narrow range of frequencies.

More specifically, in previous work (20) we investigated the EEG-recorded brain waves of language constituents using time-domain signals. Partial orders of similarity–difference were shown to be invariant between perceptual and brain representations of phonemes, words, and sentences. The present study extends this approach by using only the phases of the discrete Fourier transform (DFT) coefficients. Evidence from many studies suggests that phase modulation or phase synchronization can be the mechanism underlying the neural activities relative to the cognitive processes of language and memory retrieval (21, 22). Given what we know about the phase locking or synchronization of oscillators (23, 24), the synchronization of phases is conjectured to be the physical mechanism of retrieval. An unknown sound reaching the cortex is recognized as the specific spoken syllable /pi/ by the synchronization of its phase pattern with the stored phase pattern of /pi/ in verbal memory. Of course, our conjecture may be too simple, but the positive recognition results reported here, along with the earlier studies cited, suggest its plausibility, at least as a partial explanation. More general reviews of the brain literature on phase are in refs. 25–28.

Results

We focused on eight English consonants /p/, /t/, /b/, /g/, /f/, /s/, /v/, and /z/ and four English vowels /i/ (see), /æ/ (cat), /u/ (zoo), and /ɑ/ (father). The phonemes were selected to investigate three traditional phonological features of consonants [voicing, continuant (stop vs. fricative), and place of articulation] plus two features of vowels (height and backness). (More details about the distinctive features can be found in *SI Text*, section 1 and Tables S1 and S2.) We analyzed EEG-recorded brain data from two experiments. The first experiment focused on a set of 32 consonant–vowel (CV) syllables (eight consonants times four vowels). For each of the 32 syllables, we made 672 brain recordings from four participants (participant ids S1–S4) for a total of $672 \times 32 = 21{,}504$ trials. We placed the 128 EEG sensors on the scalp of a participant to record the brain data. (The channel locations are given in *SI Text*, section 2 and Fig. S1.) More details of the experiment setup are described in *Materials and Methods*. The second EEG experiment used the same experimental setup. In this case, we recorded brain wave responses to 1,792 isolated nonsyllabic stimuli of each of the four vowels from one participant (S4).

Results of Recognizing Initial Consonants Using Temporal and Spectral Representations of EEG. We first examined whether the spectral analysis can extract attributes of brain waves associated to phoneme perception. We tested the rates of recognizing the eight initial consonants when the signal of each channel was represented by one of three attribute vectors. The components of the first DFT representation vector were just the complex number representations of the sine wave decomposition of the original signal. The components of the second amplitude vector were the amplitudes of these sine waves. The third phase vector was the superposition of the sine waves with normalized amplitudes. Therefore, for this vector, only the frequencies and phases remained (details are in *Materials and Methods*). The attribute vectors of all of the monopolar channels were concatenated to form the long attribute vector of a trial. We applied principle component analysis (PCA) to reduce the vector

Author contributions: R.W., M.P.-G., and P.S. designed research; R.W. performed research; M.P.-G., C.C., and P.S. contributed new reagents/analytic tools; R.W. analyzed data; and R.W. and P.S. wrote the paper.

The authors declare no conflict of interest.

Data deposition: The dataset for the experiments reported in this manuscript consists of about 32 GB of EEG brain data. No Stanford University institutional database is available to us for a database of this size. In addition, there are strict rules of privacy at Stanford University and under current statutory law. EEG data are available upon request.

[1]To whom correspondence should be addressed. E-mail: psuppes@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1217500109/-/DCSupplemental.

from 7,688 to 200 components. (*SI Text*, section 3 and Fig. S2 describe the preliminary analysis.) Our recognition model used support–vector–machine (SVM) ensembles with bootstrap aggregating (or bagging). We used this scheme to recognize the averaged trials of initial consonants, vowels, and syllables.

For recognizing the initial consonants, we randomly drew 300 trials from 672 trials of each syllable and made 12 test samples from them. Each test sample was the average across 25 individual trials drawn without replacement. We combined all of the $12 \times 32 = 384$ averaged test samples to form the out-of-sample test set (OOS). The recognizer was trained using a training set (TR) made up of the remaining trials. The model using DFT representation vectors achieved a significant 38.0% recognition rate. It is similar to the rate of the model using the temporal signal. With a recognition rate of 11.2%, the amplitude-only model had a near-chance level rate. The phase-only model had a recognition rate of 36.5%, comparable with the rate of the DFT coefficient model.

By refining the recognition model, we improved the recognition rate. Representing the signal in the frequency domain makes possible the use of filtering techniques to remove the oscillation components that are unrelated to phoneme perception. We have shown in previous work that we can usually improve the recognition rate by finding an optimal frequency range over a grid of low- and high-frequency cutoffs (20). Using the phase-only model, we first looked for the approximate optimal frequency range for recognizing the eight initial consonants through a 10-fold cross-validation using only the trials from the training set. [Details are given in *SI Text*, section 4. The averages of the 10 cross-validation rates for each candidate frequency pair (L, H) are shown in Fig. S3.] The pair (2 Hz, 9 Hz) had the best average cross-validation rate of 44.1%. We retested the SVM with bagging recognizer using the optimal frequency band of 2–9 Hz. For the eight initial consonants, 193 of 384 test samples (50.3%) were recognized correctly ($P < 10^{-71}$). This result was significantly better than the 37.5% using the temporal signal.

The next improvement took advantage of the method in ref. 29 to linearly transform the original potential distribution into its surface Laplacian and the 2D scalp tangential electric field associated with it. The results of these derivations were combined into a single 3D wave, which we will refer to as the surface Laplacian and electric field (LEF) wave for simplicity. We then derived the phases in the optimal frequency range (2–9 Hz) as previously described. The resulting accuracy rate of recognizing the eight initial consonants was improved to 60.7% ($P < 10^{-100}$).

Similar to the procedure for optimizing the frequency range, we used the training set to select the optimal combination of channels. We estimated the rate of recognizing the eight consonants using the phase pattern of the LEF wave at the location of each electrode. [Fig. S4*A* shows the average rate of 10-fold cross-validation in a topographic map of the head at the corresponding locations. The topographic map of the cross-validation rate using the time-domain LEF waves band pass-filtered to 2–9Hz (shown in Fig. S4*B*) was very similar to the map using the phase representations.] The best cross-validation rate (53%) was obtained at the electrode locations close to the auditory cortical area. The rates at the frontal and posterior locations were very low. We ranked the locations by their average cross-validation rates and ran another cross-validation test using a combination of the N best locations. (The results are shown in Fig. S4*C*.) When N was greater than 5, the cross-validation rate did not change significantly with increasing N. Thus, we reconfigured the SVM with bagging recognizer using the phase patterns of the LEF waves at the five best electrode locations. This reconfiguration improved the recognition rate to 64.6% ($P < 10^{-100}$). Finally, by optimizing the temporal interval, we increased the rate to 66.7%. (More details are given in *SI Text*, section 6 and Fig. S5.)

For comparison, we also recognized the eight initial consonants using the time-domain LEF waves that were band pass-filtered to 2–9 Hz. We filtered the continuous time-domain brain waves of each session with a fourth-order Butterworth band-pass filter after down-sampling them to 62.5 Hz. Then, we converted the filtered potential signal to the LEF waves using the spatial linear transformation. Using the filtered time-domain LEF waves at the same five best locations, we correctly recognized 61.2% of the 384 test samples ($P < 10^{-100}$). All these results are summarized in Table 1.

Results of Recognizing the Vowels. Using the best configuration for recognizing consonants, we tested the model's performance on recognizing the four vowels under three different context conditions. For recognizing the vowels in isolation, we built 140 out-of-sample averaged test samples from the isolated vowel dataset and used all of the remaining trials to train the recognizer. It correctly recognized 126 of 140 test samples, with a rate of 90.0% ($P < 10^{-58}$). When recognizing the vowels using 384 test samples of the CV syllables, the accuracy rate was 45.8% ($P < 10^{-18}$). We also tested the vowel recognition accuracy when the consonant is known. In this case, we built eight vowel recognizers, each for an initial consonant, and we obtained the average recognition rate of 62.8% ($P < 10^{-54}$). The results suggest that the brain waves of initial consonants significantly affect the brain waves of vowels.

Brain Wave and Perceptual Similarity–Differences of Phonemes. To get a more quantitative analysis, we applied the methods in ref. 20 to analyze the confusion matrices of our present recognition model. Then, we compared brain wave and perceptual (2, 3) similarity trees of phonemes. (*SI Text*, section 7 discusses phoneme perception experiments. *SI Text*, section 8 and Table S3–S6 give confusion matrices.)

Fig. 1 *A* and *B* compares brain and perceptual similarity trees of vowels. Considering that, in the vowel perception experiments, the vowels were always presented with a context, we used the confusion matrix of recognizing the vowels in the context of CV syllables to derive the brain similarity tree. The perceptual confusion matrix was determined from the findings in ref. 3.

The perceptual similarity tree of vowels is approximately isomorphic to the brain similarity tree. In both cases, any vowel test sample is more similar to its own prototype than to any other vowel, which is consistent with high recognition rates. Also, the separation between open vowels /æ/ and /ɑ/ and closed vowels /i/ and /u/ is interesting. It suggests that, like with the perceptual finding, the brain representation of vowel height is more robust in the sense of less confusion than vowel backness. The results indicate that the brain waves reflect the low-frequency contrast around the frequency range of formant F1 (less than 1,000 Hz) better than the higher-frequency contrast around F2 (1,000–2,500 Hz). This finding is consistent with the fact that the human cochlea, where the sound pressure wave is converted to neural firings of the auditory nerve fibers, has higher resolution at low frequencies.

The similarity trees of initial consonants are shown in Fig. 1 *C* and *D*. We make three observations:

i) Among the three distinctive features being examined, voicing is the most salient for both the brain and perceptual representations of consonants. The voiced and voiceless consonants merge only at the top level in the perception similarity tree. We observed a similar phenomenon in the brain, except for the sound /z/. The robustness of voicing for brain waves suggests that the temporal difference of the auditory input, such as the voice onset time, which is the primary acoustic cue for the voicing contrast (30), is well-preserved in the brain representation.

ii) For voiceless consonants, the feature continuant is more distinctive than place of articulation for both brain and perceptual representations. In fact, the place of articulation is the most confused feature for brain wave representations, because three of four pairs of the consonants that only differ on place of articulation (/p/ and /t/, b/ and /g/, and /f/ and /s/) are merged first.

iii) The major difference between the trees is in the grouping structure of the voiced consonants; /b/ is perceptually most confused with the voiced fricative /v/. This confusion persists in the brain wave results, except that it is not as strong as the confusion between /b/ and /g/. The contrast between /b/ and /g/ is mostly in the transition portion of the F2 formant of the

Table 1. Recognition rates for the eight initial consonants

Model	Rate (%)	*P* value
Temporal signal (TIME)	37.5	$<10^{-34}$
Spectral attributes		
DFT	38.0	$<10^{-35}$
Amplitude	11.2	0.801
Phase	36.5	$<10^{-31}$
Phase (2–9 Hz)	50.3	$<10^{-71}$
Phase (2–9 Hz) of the LEF wave	60.7	$<10^{-100}$
Phase (2–9 Hz) of the LEF waves at the five locations	64.6	$<10^{-100}$
Last phase model with optimal temporal interval	66.7	$<10^{-100}$
2- to 9-Hz filtered temporal LEF waves at the five locations	61.2	$<10^{-100}$

vowel that follows (2), whereas the primary perceptual cues to distinguish /b/ and /v/ are the abrupt onset of the stop sound of /b/ and the turbulent noises of the frictions of /v/ (31). Although the attraction between /b/ and /v/ is commonly seen in the perceptual consonant categorization data using masking noise (2, 4, 5), it is not clearly shown in the perceptual experiment of short-term memory (32) and the neural discriminations of animal responses to the human speech stimulation (15). Consequently, a possible explanation for this mismatch between the brain wave and perceptual confusions is that friction is more perceptually distorted by white noise than the formant transitions.

Recognizing the Phonological Distinctive Features. The similarity analysis of the phoneme recognition results shows that the distinctive features can also be seen in the brain waves of phonemes. For example, the brain waves of voiced consonants are not likely to be confused with the brain waves of voiceless consonants when they are represented as the phase pattern-based observation vector of our model. This finding naturally suggests that we can predict distinctive features using EEG-recorded brain waves.

To show this finding directly, we trained binary classifiers to predict the three distinctive features of initial consonants: voicing, continuant, and place of articulation (labial/nonlabial). The vowel height and vowel backness classifiers were trained and tested using eight sessions of the isolated vowel data. As in the previous analysis, the observation vector of each trial consisted of the 2- to 9-Hz component phases of surface Laplacian and scalp tangential electric field at the five best locations. The recognition results are summarized in Table 2.

We found that, for all five features, the recognition rates were highly significant. This finding suggests a possible brain mechanism for the recognition of phonemes: the phonemes are identified by using the phonological distinctive features. The scheme is computationally efficient, because only a relatively small set of distinctive features is required to distinguish all of the phonemes within one language. Additionally, the computation for each distinctive feature can be done independently, allowing efficient parallel processing.

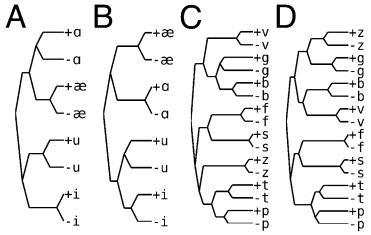

Fig. 1. The similarities of brain and perceptual representations of phonemes. (*A*) The similarity tree of brain wave representation of vowels. The test samples and the prototypes of phonemes are marked using − and +, respectively. (*B*) The perceptual similarity tree of vowels derived from results of the psychological experiments in ref. 3. (*C*) The similarity tree of the brain wave representation of the eight consonants. (*D*) The perception similarity tree of the eight consonants derived from results of the psychological experiments in ref. 2. The detailed construction and interpretation of the trees is given in *Methods and Materials*.

This distinctive feature-based mechanism can also be used in recognizing the brain waves of phonemes. In this model, we used an ensemble of binary SVM classifiers based on the phonological distinctive features of stimuli to implement the multiclass SVM in the SVM with bagging recognizer. The distinctive feature classifiers can be organized in parallel or hierarchically. When the parallel mode was used, we estimated the parameters of each binary classifier and made the distinctive feature prediction independently. In the hierarchically organized recognition model, the distinctive features were predicted in sequence (*SI Text*, section 9 and Fig. S6). For instance, for recognizing the four vowels, we first predicted the vowel height feature (open or closed). Then, we used different binary vowel backness classifiers for the open and closed vowels. We denote this prediction order as vowel-height→vowel-backness.

We recognized 140 out-of-sample test samples of the four vowels presented in isolation using the parallel model and the hierarchical model with two feature orders. The results are shown in Table 3. Using the parallel model, 80.0% ($P < 10^{-42}$) of the test samples were recognized correctly. The hierarchical model with the prediction order vowel-backness→vowel-height had a similar rate of 82.9% ($P < 10^{-46}$). Both were lower than the rates of the hierarchical model with the order vowel-height→vowel-backness (90.0%; $P < 10^{-58}$). The results suggest a dependency between vowel-height and vowel-backness. More specifically, the distinctions on vowel-backness are different for open vowels and closed vowels.

A similar approach can be used to determine the dependency of other distinctive features. For example, Table 3 shows the recognition results of classifying the 384 test samples of the eight initial consonants into four categories [voiceless stops (/p/,/t/), voiced stops (/b/, /g/), voiceless fricatives (/f/, /s/), and voiced fricatives (/v,/, /z/)] using the distinctive feature voicing and continuant.

The recognition rate of using continuant→voicing hierarchy (76.3%) was higher than using parallel model (71.4%) and voicing→continuant hierarchy (72.1%). This finding supports the view that the decision on voicing depends on the continuant category. When there are more than two features to be predicted, how to determine the optimal hierarchy is a very complicated problem and worth additional investigation (33). Intuitively, the distinctive feature that achieves the highest classification rate in the binary classification experiment should be predicted at first to provide the best foundation for additional prediction. This result is also consistent with our results of the two-feature recognition tasks discussed above. To further test this idea, in this study, we recognized the eight initial consonants using the order continuant→voicing→place and recognized the 32 CV syllables using the order continuant→voicing→place→vowel-height→vowel-backness. Both rates, 63.3% and 37.5%, were significantly higher than the corresponding results using parallel models (Table 3).

Phase Pattern of Initial Consonants. In our study, the best recognition model of brain waves of individual phonemes used the 2- to 9-Hz phase patterns of the LEF waves at the five best electrode locations. Using the brain wave recordings from one participant (S4), we estimated the mean phases of the surface Laplacian of the scalp potential, measured at the previously defined best location (electrode E41 in Fig. S1), at ~4, 5, 6, and 7 Hz, and we show them as unit length vectors in polar coordinates in Fig. 2. For each frequency, we estimated the mean phase of a phoneme as the average phase of the corresponding sine wave components of all of the individual trials associated with this phoneme. The average phase of k phase angles $\phi_1, \ldots, \phi_k$, denoted as $\bar{\phi}$, was calculated as the angle of the complex number $\sum_k \cos\phi_k + i\sum_k \sin\phi_k$. The similarity between the phoneme pairs that differed only on place of

NEUROSCIENCE

Table 2. Recognizing the five distinctive features

Distinctive feature	Grouping		No. of test samples	Rate (%)	*P* value
Voicing	Voiceless	/p, t, f, s/	384	83.5	$<10^{-42}$
	Voiced	/b, g, v, z/			
Continuant	Stop	/p, t, b, g/	384	85.4	$<10^{-47}$
	Fricative	/f, s, v, z/			
Place (labial)	Labial	/p, b, f, v/	384	74.4	$<10^{-20}$
	Nonlabial	/t, g, s, z/			
Height	Open	/æ, ɑ/	140	93.6	$<10^{-28}$
	Closed	/i, u/			
Backness	Front	/æ, i/	140	85.0	$<10^{-17}$
	Back	/ɑ, u/			

articulation (for example, /p/, /t/ and /b/, /g/) can also be seen in Fig. 2. (Some histograms of averaged phases are shown in *SI Text*, section 10 and Fig. S7.)

We built a simple recognizer to test whether these mean phases can represent the brain waves of phonemes. In this analysis, each trial of the EEG-recorded brain waves was represented using the phase value of 2–9 Hz of the LEF waves at five locations. In total, we represent each trial with $15 \times 8 = 120$ angles. We divided all of the trials of the syllable experiment into training and out-of-sample test sets like in the previous analysis. We built a prototype for each initial consonant using the training set. The prototype of the j^{th} consonant denoted as (Eq. **1**)

$$\boldsymbol{\theta}^{(j)} = \left[\theta_1^{(j)}, \theta_2^{(j)}, \ldots, \theta_{120}^{(j)}\right]^T, \; -\pi < \theta_k^{(j)} < \pi \qquad [1]$$

consisted of the mean of each of the 120 angles estimated using all of the trials associated with that consonant in the training set. A test sample was calculated as the average of 25 out-of-sample trials. To recognize a test sample, we computed the squared angle differences between its phases and the corresponding estimated mean phases of each consonant. It was recognized as belonging to the consonant with the sum of squared angle differences that was minimal. Here, the angle difference $\text{diff}(\phi, \theta)$ was defined as the smaller angle between the unit length vectors representing ϕ and θ in polar coordinates. Using this phase pattern recognition model, 187 of 384 test samples (48.7%) were recognized correctly. This result supports use of phase patterns of brain waves to recognize phonemes. Although the model does not provide a physical mechanism, such as phase locking of oscillators, the recognition criterion of minimum mean phase differences is close to the useful criterion of quasisynchronization for oscillators (23).

Table 3. Phonemes recognition results using distinctive features

Model	Rate (%)	*P* value
Four vowels		
Parallel	80.0	$<10^{-42}$
Height→backness	90.0	$<10^{-58}$
Backness→height	82.9	$<10^{-46}$
Four consonant classes defined by voicing and continuant		
Parallel	71.4	$<10^{-79}$
Voicing→continuant	72.1	$<10^{-81}$
Continuant→voicing	76.3	$<10^{-97}$
Eight consonants		
Parallel	49.0	$<10^{-66}$
Continuant→voicing→place	63.3	$<10^{-100}$
32 syllables		
Parallel	19.0	$<10^{-33}$
Consonant→vowel features	37.5	$<10^{-100}$

Discussion

In this study, we used an SVM with bagging model to recognize EEG-recorded brain waves of auditory phoneme stimuli. Each brain wave was represented by *i*) its DFT coefficients; *ii*) only its amplitudes; or *iii*) only its phases. Results supported the view that eliminating the amplitude information of the DFT did not diminish the recognition rate of brain representations of different initial consonants. The phase pattern of sinusoidal components in the frequency range from 2 to 9 Hz seems critically important in recognizing the brain wave of a phoneme. This low-frequency range was consistent with other findings about the magnetoencephalograph phase patterns that discriminate speech in human auditory cortex (34). (*SI Text*, section 11 and Table S7 discuss our consonant recognition results using higher-frequency components.) Using the scalp tangential electric field and surface Laplacian vector close to the auditory cortical area can also improve the recognition rate.

This model works well in recognizing the initial consonants and vowels presented in isolation. It also recognized the vowels in CV syllables at a significant rate, although the rate was definitely lower than the rate for the isolated vowels. One possible reason is that, in a CV context, fluctuations evoked by the initial consonant brain waves impose extra noise on those brain waves of the vowels. Also, because the durations of initial consonants are different, the onsets of phases of the vowels' brain waves can be different. Not surprisingly, vowel recognition rates can be improved when the initial consonant is given.

In our analysis, the DFT-based phase analysis successfully extracted the attributes of brain signals that relate to phoneme perception. However, the spectral properties of brain waves are generally not time-invariant. The DFT method cannot characterize the changes of phase. Other time-frequency analysis methods, such as those methods of wavelets, may provide additional information.

The results of Table 3 provide significant support for the importance of phonological distinctive features in the neural

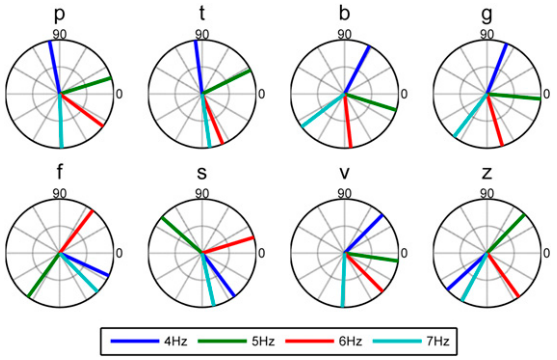

Fig. 2. Mean phases of the 4-, 5-, 6-, and 7-Hz components of the surface Laplacian of the scalp potential at the location of electrode E41 of the eight initial consonants for participant S4.

mechanism of phoneme perception. This finding, in turn, suggests a computationally efficient brain mechanism for recognizing phonemes and thus, words in verbal memory storage and retrieval. Providing such a mechanism seems essential to any physically concrete brain realization for the detailed but still abstract psychological models of human memory (35, 36). In particular, the success of recognizing the brain waves of phonemes using phases suggests that phase synchronization may be one of the physical mechanisms that the brain uses to retrieve words from verbal memory in speaking, listening, writing, or reading. Our empirical results provide useful experimental evidence to support what many researchers, some of whom we have cited, believe to be the case (i.e., phase synchronization in the brain is an important characteristic of many brain processes). We fully recognize that, at the present time, this view is controversial, and it will only be widely accepted on the positive outcome of future theoretical and experimental work in this area of brain research.

Materials and Methods

Subjects. Four right-handed adult subjects (three females and one male) participated in this experiment. All of subjects reported no history of hearing problems. Informed consent was obtained from all of the participants in accordance with the guidelines and approval of the Stanford University Institutional Review Board.

Stimuli. We recorded seven repetitions of each of the 32 consonant–vowel format syllables and four isolated vowels, which were read by a male American-English native speaker. The recordings were coded as 44.1-kHz mono-channel waveform audio files.

EEG Data Acquisition. The brain waves of 32 syllables were collected in our laboratory from 2008 to 2009. Using stereo speakers, we presented the $7 \times 32 = 224$ auditory stimuli to the four participants (S1–S4) in a random order. The participants, sitting in a dark booth, were instructed to listen to the sound attentively. No behavioral responses from participants were required. EEG signals were recorded using the 128-channel Geodesic Sensor Net (37) and EGI's Geodesic EEG system 300. To reduce the eye movement artifacts, we asked the participants to focus on a small cross at the center of the computer screen. The 128-channel brain data were recorded with a sampling rate of 1,000 Hz. There were 124 monopolar channels with a common reference Cz and 2 bipolar reference channels of eye movements. We had a total of 24 sessions from four participants. The numbers of trials recorded from these participants were 3,584 (S1), 4,480 (S2), 6,272 (S3), and 7,168 (S4). The major numerical difference per subject occurred, because we decided that, after running S2, it would be better to have more trials for testing. We combined the data from all of the 24 sessions to train and test the phoneme recognizer. The brain data of isolated vowels were recorded using the same experimental setup in 2010. In one recording session, the 28 stimuli were presented to the participant in random order 32 times. We recorded eight sessions from one participant (S4) for a total of $28 \times 32 \times 8 = 7{,}168$ trials.

Spectral Representations of EEG Signals. We used the N point DFT to derive the frequency-domain properties of the discrete time-domain real signal $x_n, n = 0, \ldots, N-1$ (Eq. **2**):

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk} \; k = 0, \ldots, N-1. \qquad [2]$$

Additionally, x_n can be reconstructed from X_k using the inverse transformation (Eq. **3**)

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{\frac{2\pi i}{N} nk} \; n = 0, \ldots, N-1. \qquad [3]$$

If we express X_k using its norm and phase ($X_k = A_k e^{i\phi_k}$), x_n can be written as the following equation when N is even (Eq. **4**):

$$x_n = \frac{1}{N} \left[A_0 + A_{\frac{N}{2}} \cos \pi \, n + 2 \sum_{k=1}^{\frac{N}{2}-1} \cos \frac{2\pi}{N} kn + \phi_k \right]. \qquad [4]$$

Eq. **4** shows that the finite discrete time-domain signal $x_n, n = 0, \ldots, N-1$ can be decomposed as the superposition of a series of discrete sinusoidal functions, with the kth function defined by three attributes: frequency ($2\pi k/N$), amplitude (A_k), and phase (ϕ_k). The complex number DFT coefficient X_k represents the sine wave component at frequency $2\pi k/N$. In this article, we used several vectors to represent all or part of these spectral attributes of the signal x_n. The DFT representation vector, denoted as X_{DFT}, was constructed from the real and imaginary parts of all of the DFT coefficients X_k. For the conjugate symmetric property of the DFT, X_{DFT} can be reduced to (Eq. **5**)

$$X_{\text{DFT}} = \left[X_0, \text{Re}\left(X_1, \ldots, X_{\frac{N}{2}-1} \right), \text{Im}\left(X_1, \ldots, X_{\frac{N}{2}-1} \right), X_{\frac{N}{2}} \right]. \qquad [5]$$

In the amplitude vector X_{AMP}, only the amplitudes of the sine wave components were retained (Eq. **6**):

$$X_{\text{AMP}} = \left[A_0, A_1, \ldots, A_{\frac{N}{2}} \right]. \qquad [6]$$

To build a vector that contained only the phase of each sine wave component, we eliminated the amplitude information in the DFT by normalizing the amplitudes of all nonzero X_k values and transforming them back to the time domain using the inverse DFT (IDFT). More specifically, for each element of $X_k, k = 0, \ldots, N-1$, the normalized DFT was calculated as (Eq. **7**)

$$\tilde{X}_k = \begin{cases} e^{i\phi_k} & \text{if } A_k \neq 0 \\ 0 & \text{if } A_k = 0 \end{cases}. \qquad [7]$$

Additionally, the phase vector was defined as (Eq. **8**)

$$X_{\text{PHS}} = \text{IDFT}\left\{ \tilde{X}_k \right\}. \qquad [8]$$

SVM with Bagging Recognition Methods. The main statistical scheme underlying our phase recognition approach, SVM with bagging, is a method to generate multiple versions of an SVM recognizer through the bootstrap sampling approach and use these versions to get an aggregate recognition rate (38). We used a soft-margin SVM as the basic classification unit in this recognition model (39). The original SVM is a binary classifier looking for a separation hyperplane that maximizes the empirical functional margin. The soft-margin SVM introduces a cost factor C to allow training samples with margins less than one or even negative. We built the multiclass SVM using $M(M-1)/2$ one-against-one binary SVMs, one for each pair of the M classes. A test sample was predicted as belonging to the class that won the maximum number of votes from the binary classifiers. The SVM methods can implement nonlinear classification by mapping the data into higher-dimensional feature space using the kernel trick. The results of recognizing the eight initial consonants using nonlinear kernels are given in *SI Text*, section 12.

We used MATLAB to implement these phoneme recognition methods and analyzed the resulting confusion matrices. EEGLAB (40) was used in preprocessing. LIBSVM, a library for support vector machines (41), was used to implement SVM training and predicting.

In the preprocessing stage, we passed the signal of each channel through a fourth-order Butterworth high-pass filter with the cutoff frequency at 1 Hz and then down-sampled it to 62.5 Hz. A low-pass filter with zero phase response was used for antialiasing before down-sampling. We used independent component analysis (42) to remove eye movement artifacts. Details of the preprocessing are given in *SI Text*, section 14.

We randomly divided the trials into a TR and an OOS as described. Before using TR to estimate the SVM with bagging recognizer parameters, we applied PCA to reduce the dimension of the attribute vectors. The PCA transformation matrix was estimated using all of the individual trials of TR and then applied to both TR and OOS. In each of the 35 bootstrap repetitions, we randomly drew 80% of the trials in the training set TR to create a bootstrap replication of TR, noted as $\text{TR}^{(i)}$. Each replication was drawn independently and used to train an N-class SVM recognizer. The remaining 20% of TR trials, noted as $\text{TE}^{(i)}$, were used to test the rate of the SVM and also determine the weight of the SVM in aggregating.

In each bootstrap repetition, we used a sample with replacement scheme to reuse the individual trials in an efficient way without bringing bias to the recognition rate. This method was similar to the bootstrap sampling that randomly generated the averaged training and test samples from $\text{TR}^{(i)}$ and $\text{TE}^{(i)}$, respectively. The optimal parameters of each SVM classifier, such as the cost factor C, were determined by nested fivefold cross-validations

using $TR^{(i)}$. (*SI Text*, section 15 and Fig. S8 have the diagram of the recognition model.)

Computation Using Electric Field and Surface Laplacian Operators for Improving Recognition Rate. At each sampling time, the components of the electric field and the surface Laplacian are linear transformations of the EEG-recorded potential signal. The transformation matrices depend only on the electrode locations and regularization parameters. In our study, we first computed the transformation matrices, as described in *SI Text*, section 16, with the regularization parameter $\lambda = 2$, interpolation order $m = 3$ and head radius $r = 9.2$cm of the assumed spherical head model. After preprocessing the signal, which included high-pass filtering, independent component analysis ocular artifacts removal, and down-sampling, each monopolar channel signal was transformed to three waveforms corresponding to the surface Laplacian and the polar and azimuthal components of the scalp electric field. When the spectral attribute vectors were tested, we converted each component of a trial to its spectral representation individually.

Analyzing the Confusion Matrices. After we recognized the brain waves of phonemes, we calculated the number of test samples of phoneme o_j that were recognized as belonging to phoneme o_i. Normalizing this number by dividing it by the total number of test samples of phoneme o_j, we obtained an estimate of the conditional probability $\mathrm{p}(o_j^+|o_i^-)$, where + and − denote the prototypes and test samples, respectively. By repeating this estimation for each pair (i, j), a conditional probability matrix was constructed. The normalized confusion matrix of recognition results provided empirical evidence by which to order the similarity–differences of the brain wave representations of phonemes. Our previous work (20) proposed a method to derive invariant partial orders of similarity–differences of brain and perceptual representations and corresponding similarity trees using the estimated conditional probability matrices. The same technique was used to analyze the recognition results in this article. It is natural to think that the phoneme o_i^- is more similar to the prototype o_j^+ than the phoneme $o_{i'}^-$ is similar to the prototype $o_{j'}^+$ if and only if $\mathrm{p}(o_j^+|o_i^-) > p(o_{j'}^+|o_{i'}^-)$. We write the corresponding qualitative similarity relation as $o_j^+|o_i^- \succ o_{j'}^+|o_{i'}^-$. Because the confusion matrices are generally not symmetric, the similarity–differences are not necessarily symmetric. In practice, the difference between the similarity of o_j^+ and o_i^- and the similarity of $o_{j'}^+$ and $o_{i'}^-$ is considered statistically insignificant if $\mathrm{p}(o_j^+|o_i^-)$ and $\mathrm{p}(o_{j'}^+|o_{i'}^-)$ are close enough. For this purpose, we introduce the numerical threshold ε, and therefore (**9**),

$$o_j^+\left|o_i^- \approx o_{j'}^+\right|o_{i'}^- \ \text{ iff} \left|\mathrm{p}\left(o_j^+\middle|o_i^-\right) - \mathrm{p}\left(o_{j'}^+\middle|o_{i'}^-\right)\right| < \varepsilon. \qquad [9]$$

The similarity–difference ordering is the basis of generating a qualitative similarity tree that gives a hierarchy partition of the combined set of test samples and prototypes $\mathcal{O} = \{o_1^+, \ldots, o_N^+, o_1^-, \ldots, o_N^-\}$ in a binary tree structure. We defined the merged product of two subsets of $\mathcal{O}$ ($\mathcal{O}_I$ and $\mathcal{O}_J$) as (Eq. **10**)

$$\mathcal{O}_I\mathcal{O}_J = \left\{o_j^+\middle|o_i^- : o_j^+ \in \mathcal{O}_I \ \&\ o_i^- \in \mathcal{O}_J \ \text{ or } \ o_j^+ \in \mathcal{O}_J \ \&\ o_i^- \in \mathcal{O}_I\right\}. \qquad [10]$$

The inductive procedure started from a partition P_0, which includes $2N$ singleton sets of elements of $\mathcal{O}$. In the k^{th} inductive step, two subsets in the partition P_{k-1} are chosen to be merged, and therefore, the least pair of their merged product under the similarity–difference ordering $\succ$ is maximized among all of the possible merges. Consequently, the subsets with greater similarity are merged earlier than the subsets with smaller similarity in the inductive steps. For instance, in the brain similarity tree of the four vowels, the merge point of the branches of +i and −i was to the right of the merge point of those branches of +u and −u. This finding means that the similarity–difference of +u and −u is greater than the similarity–difference of +i and −i. The similarity tree provides a fairly intuitive approach to summarize the similarity of the test samples and prototypes in a matrix of conditional probability densities. More details of these methods can be found in ref. 20.

ACKNOWLEDGMENTS. We thank Blair Kaneshiro and Duc Nguyen for their assistance in conducting the EEG experiments. The work was supported by an anonymous private charitable fund.

1. Well FL (1906) *Linguistic Lapses. With Especial Reference to the Perception of Linguistic Sounds*, Columbia University Contributions to Philosophy and Psychology (Science Press, New York).
2. Miller GA, Nicely PE (1955) An analysis of perceptual confusions among some English consonants. *J Acoust Soc Am* 27:338–352.
3. Pickett JM (1957) Perception of vowels heard in noises of various spectra. *J Acoust Soc Am* 29:613–620.
4. Wang MD, Bilger RC (1973) Consonant confusions in noise: A study of perceptual features. *J Acoust Soc Am* 54(5):1248–1266.
5. Phatak SA, Lovitt A, Allen JB (2008) Consonant confusions in white noise. *J Acoust Soc Am* 124(2):1220–1233.
6. Näätänen R, Gaillard AWK, Mäntysalo S (1978) Early selective-attention effect on evoked potential reinterpreted. *Acta Psychol (Amst)* 42(4):313–329.
7. Näätänen R (2001) The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology* 38(1):1–21.
8. Näätänen R, et al. (1997) Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385(6615):432–434.
9. Obleser J, Lahiri A, Eulitz C (2004) Magnetic brain response mirrors extraction of phonological features from spoken vowels. *J Cogn Neurosci* 16(1):31–39.
10. Eulitz C (2007) Representation of phonological features in the brain: Evidence from mismatch negativity. *Proceedings of the 16th International Congress of Phonetic Science*, eds Trouvain J, Barry WJ (Saarland University, Saarbrücken) pp 113–116.
11. Frye RE, et al. (2007) Linear coding of voice onset time. *J Cogn Neurosci* 19(9):1476–1487.
12. Steinschneider M, Reser D, Schroeder CE, Arezzo JC (1995) Tonotopic organization of responses reflecting stop consonant place of articulation in primary auditory cortex (A1) of the monkey. *Brain Res* 674(1):147–152.
13. Steinschneider M, Fishman YI, Arezzo JC (2003) Representation of the voice onset time (VOT) speech parameter in population responses within primary auditory cortex of the awake monkey. *J Acoust Soc Am* 114(1):307–321.
14. Engineer CT, et al. (2008) Cortical activity patterns predict speech discrimination ability. *Nat Neurosci* 11(5):603–608.
15. Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am* 123(2):899–909.
16. Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural substrates of phonemic perception. *Cereb Cortex* 15(10):1621–1631.
17. Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322(5903):970–973.
18. Pfurtscheller G, Lopes da Silva FH (1999) Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clin Neurophysiol* 110(11):1842–1857.
19. Tallon-Baudry C, Bertrand O (1999) Oscillatory gamma activity in humans and its role in object representation. *Trends Cogn Sci* 3(4):151–162.
20. Suppes P, Perreau-Guimaraes M, Wong DK (2009) Partial orders of similarity differences invariant between EEG-recorded brain and perceptual representations of language. *Neural Comput* 21(11):3228–3269.
21. Fell J, et al. (2003) Rhinal-hippocampal theta coherence during declarative memory formation: Interaction with gamma synchronization? *Eur J Neurosci* 17(5):1082–1088.
22. Sederberg PB, Kahana MJ, Howard MW, Donner EJ, Madsen JR (2003) Theta and gamma oscillations during encoding predict subsequent recall. *J Neurosci* 23(34):10809–10814.
23. Vassilieva E, Pinto G, de Barros JA, Suppes P (2011) Learning pattern recognition through quasi-synchronization of phase oscillators. *IEEE Trans Neural Netw* 22(1):84–95.
24. Suppes P, de Barros JA, Oas G (2012) Phase-oscillator computations as neural models of stimulus-response conditioning and response selection. *J Math Psychol* 56:95–117.
25. Bastiaansen M, Hagoort P (2006) Oscillatory neuronal dynamics during language comprehension. *Prog Brain Res* 159:179–196.
26. Sauseng P, Klimesch W (2008) What does phase information of oscillatory brain activity tell us about cognitive processes? *Neurosci Biobehav Rev* 32(5):1001–1013.
27. Fell J, Axmacher N (2011) The role of phase synchronization in memory processes. *Nat Rev Neurosci* 12(2):105–118.
28. Penny WD, Kiebel SJ, Kilner JM, Rugg MD (2002) Event-related brain dynamics. *Trends Neurosci* 25(8):387–389.
29. Carvalhaes CG, Suppes P (2011) A spline framework for estimating the EEG surface laplacian using the Euclidean metric. *Neural Comput* 23(11):2974–3000.
30. Lisker L, Abramson A (1964) A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20:384–422.
31. Fujimura O, Erickson D (1997) Acoustic phonetics. *The Handbook of Phonetic Sciences*, eds Hardcastle WJ, Laver J (Blackwell, Oxford), pp 65–115.
32. Wickelgren WA (1966) Distinctive features and errors in short-term memory for English consonants. *J Acoust Soc Am* 39(2):388–398.
33. Silla CN, Freitas AA (2011) A survey of hierarchical classification across different application domains. *Data Min Knowl Discov* 22:31–72.
34. Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54(6):1001–1010.
35. Gillund G, Shiffrin RM (1984) A retrieval model for both recognition and recall. *Psychol Rev* 91(1):1–67.
36. Shiffrin RM, Steyvers M (1997) A model for recognition memory: REM-retrieving effectively from memory. *Psychon Bull Rev* 4(2):145–166.
37. Tucker DM (1993) Spatial sampling of head electrical fields: The geodesic sensor net. *Electroencephalogr Clin Neurophysiol* 87(3):154–163.
38. Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140.
39. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297.
40. Delorme A, Makeig S (2004) EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1):9–21.
41. Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol*, 2:27:1–27:27. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
42. Jung TP, et al. (2000) Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37(2):163–178.