International Journal of Bioelectromagnetism
Vol. 13, No. 4, pp. 201-206, 2011

EEG-Based Discrimination of Imagined Speech Phonemes

Xuemin Chi[a], John B. Hagedorn[a], Daniel Schoonover[a] and Michael D'Zmura[a]
[a]*Dept. Cognitive Sciences, University of California, Irvine, USA*
Correspondence: mdzmura@uci.edu, Dept. Cognitive Sciences, University of California, Irvine, USA, 92697-5100
E-mail: mdzmura@uci.edu, phone +1 949 824 4055, fax +1 949 824 2307

Abstract. This paper reports positive results for classifying imagined phonemes on the basis of EEG signals. Subjects generated in imagination five types of phonemes that differ in their primary manner of vocal articulation during overt speech production (jaw, tongue, nasal, lips and fricative). Naive Bayes and linear discriminant analysis classification methods were applied to EEG signals that were recorded during imagined phoneme production. Results show that signals from these classes can be differentiated from those generated during periods of no imagined speech and that the signals among the classes are discriminable, particularly in data collected on a single day. The simple linear classification methods are suited well to online use in BCI applications.

Keywords: EEG, classification, speech, imagery, BCI

1. Introduction

Using imagined speech in an EEG-based BCI potentially offers a natural means of expression consistent with mobility. During the production of imagined speech, one might expect to find in EEG traces of brain activity related to auditory imagery (the voice in one's head), motor imagery (imagined vocal articulation), and other aspects of speech production. Yet positive results are few. Suppes and colleagues reported over a decade ago some success in using EEG signals to discriminate among imagined sentences[Suppes et al, 1997, 1998], yet the result has not been replicated. Better substantiated are MEG and EEG results for heard speech, which show that traces of the acoustic speech waveform envelope can be extracted[Ahissar et al., 2001; Luo & Poeppel, 2007; Deng and Srinivasan, 2010]. This has recently been shown true also for imagined speech; the presumptive loudness envelopes of auditory imagery generated--the rhythm or pattern of stress--are discriminable[Deng et al., 2010]. The present study focuses more squarely on motor imagery; it seeks to determine whether phonemes that differ in pattern of vocal articulation may be discriminated in EEG. Simple classification methods applied to spectrograms of EEG activity recorded during production of the imagined phonemes provide discrimination performance of high statistical significance. The simplicity of the analysis lends itself well to future online use in BCI applications.

2. Experimental Methods

An experimental session included twelve types of trial, ten of which involved the production of a phoneme in imagination. The subjects' task was to generate in imagination the phoneme cued at trial onset. Each trial started with an auditory cue presented through a loudspeaker that either stated the phoneme to be imagined or instructed the subject to relax (see Figure 1). The cue was followed by two audible clicks, separated by an interval of duration 1 sec. The purpose of the clicks was to indicate, on imagined phoneme trials, the time at which the phoneme was to be produced in imagination. Subjects were instructed to generate as best as possible both clear auditory imagery and motor imagery while remaining completely still during imagined phoneme production. A camera was used to capture video of the subject's face, neck and shoulders during experimental sessions to help monitor muscular activity, which was limited almost exclusively to blinks and eye movements and was absent during the intermittent periods during which imagined speech was produced.

Five articulation classes of phonemes were used in the experiment; each class was represented by a pair of phonemes. The ten English-language phonemes used include two examples of sounds that involve notable movements of the jaw (-aa and –ae), of the tongue (-l and –r), of the velum (nasal -m and –n) and of the lips (rounded for -uu and –ow) as well as two fricatives (-s and –z); see Table 1. Two further types of trial appeared during which the subject was instructed to relax. Twelve instances

of each trial type were presented in block-randomized order during a single experimental session for a total of 144 trials per session. One, two, or three sessions were run on a single day, providing 24, 36 or 48 trials per phoneme, respectively, and 48, 72 or 96 trials per articulation class, respectively. The sessions were run in total for either four (subjects A,C,D,E) or three (subject B) days.

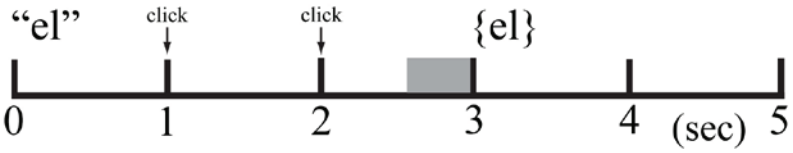


Figure 1. *Trial time course illustrated for the -l phoneme. The phoneme to be imagined was cued using a heard prompt presented at trial onset. This was followed by two audible clicks. The subject generated the cued phoneme in imagination during the second-long interval with onset three seconds into the trial. Data drawn from the 400 msec period indicated by the gray rectangle and starting at time 2.6 sec were subject to classification analysis.*

An ElectrodeArrays EA136 system was used to collect EEG data from 52 electrodes with scalp positions drawn from the 10/20 system but omitting occipital (O) and far frontal (AF) positions. Signals were collected at a sampling rate of 1000Hz, low-pass filtered to remove line noise, online-average referenced, and convolved with truncated Gabor functions in sine and cosine phase to generate causal amplitude spectrograms at nine logarithmically-spaced frequencies in the range 4-28 Hz. Both waveform data (decimated to 200 samples/sec) and spectrogram data (20 samples/sec) were collected for each trial. The onset of each trial's heard cue was accompanied by a marker signal fed directly into the EEG amplifier to mark in the EEG records the onset times of cue and imagined speech production periods.

Table 1. *The ten indicated phonemes were chosen to contrast five patterns of vocal articulation (jaw, tongue, nasal, lips and fricative)[Stevens, 1998]. Two further trial types instructed the subject to "relax" rather than generate imagined speech to produce twelve total trial types.*

Articulation class	*Phoneme*	*Examples*	*Manner of Production*
Jaw	-aa	s**aw**, j**aw**	"ah"
	-ae	h**a**t, c**a**t	"aeh"
Tongue	-l	**l**ight, **l**ed	"el"
	-r	**r**ight, **r**ed	"are"
Nasal	-m	**m**at, **m**ice	"em"
	-n	**n**et, **n**ice	"en"
Lips	-uu	wh**o**, dr**ew**	"oo"
	-ow	b**oa**t, **o**ver	"oh"
Fricative	-s	**s**ame, hi**ss**	s - hissing
	-z	**z**oo, hi**s**	z - buzzing

3. Classification Methods

The spectrographic data generated online during the experimental sessions were analyzed using two classification methods: naive Bayes classification (with a diagonal covariance matrix) and linear discriminant analysis (LDA). The need to invert the empirical covariance matrices for the LDA led us to use data from only five available frequencies (4, 6.4, 10.3, 16.4 and 26.3 Hz) and during the 400 msec period that started at the time 2.6 sec into the trial: the 400 msec preceding the nominal time at which imagined speech production commenced; see Figure 2. Including further frequencies or

extending the time period of analysis caused covariance matrices of less than full rank to be generated by Matlab's LDA routine.

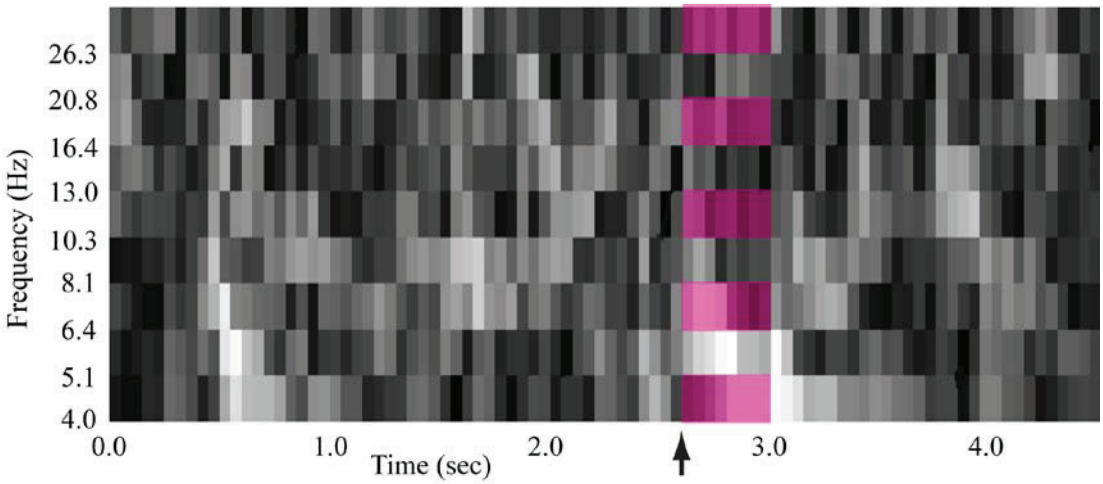


Figure 2. *Spectrogram generated online by Subject A at electrode FT7 during a trial for phoneme -l. Data within the colored rectangle were used for classification; these are drawn from the five frequencies 4.0, 6.4, 10.3, 16.4 and 26.3Hz during a 400 msec period starting at the bin indicated by the arrow.*

Both the naive Bayes and the LDA classifications were performed on data from individual electrodes. The naive Bayes calculation assumes that the covariance matrix determined by trial spectrogram features is a diagonal one; the LDA does not. Leave-one-out cross-validation was performed using the two classification methods on partial data sets corresponding to single days as well as on the full data sets. With this cross-validation scheme, classification training is performed using data from all trials but one; classification of the single trial remaining is then tested. This process is performed for each trial to arrive at an overall evaluation of classification performance. In two-way classifications, the results from single electrodes were aggregated by using the nine best electrodes in a simple voting scheme. The classification performance determined by the leave-one-out procedure was used to rank order the electrodes; votes by the top nine electrodes were then calculated to determine the overall classification.

4. Results

Classification performance levels reached 80% correct or greater in two-way classifications concerning articulation classes. As expected, the overall performance of LDA was superior to that of the naive Bayes classifier; results from the former method are detailed in the results below. Performance levels found for trials drawn from a single day were consistently higher than those found when examining data across all days.

Table 2 shows LDA performance levels for the five subjects found when classifying articulation class using the top-nine-electrode voting scheme. Each cell in Table 2 shows results for a two-way classification of articulation class indicated by row and column labels. The rows in each cell present results for the subject labeled in the second column from the left. The first number in each row shows classification performance rates found by averaging arithmetically across the rates found for trials recorded on single days. The second number in each row shows classification performance found when using the entire data set in a single classification. Values are italicized if they reach a criterion statistical significance: a p value of less than $1x10^{-4}$ for at least three of the single-day results and a p value of less than $1x10^{-5}$ for the result using the total dataset. The p values are likelihoods returned by a binomial test. The guessing rate in a two-way classification is 50%, and it is against this figure that these classification performances are compared.

Two-way classification of trials from single days produced average performance levels varying from 66.4% through 76.0% for the five subjects (see Table 2). Each subject reached a level of 80% or greater on a single day in a particular classification. Performance levels found using all trials varied from 56.5% through 70.3%. Performance levels were similar across the 15 two-way classifications,

although those single-day classifications involving the fricatives -s and -z are marginally superior, as are the total-data discriminations involving relax.

Table 2. *LDA classification performance in percent for 15 pairwise classifications of articulation class for subjects A, B, C, D and E using the top-nine-electrode voting scheme. The first number in each row is the arithmetic average of the single-day classification performances. The second number is the performance found when classifying the entire dataset. The guessing rate is 50%; italicized performance rates are highly statistically significant (see text for details).*

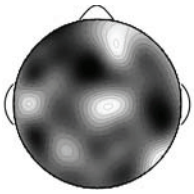
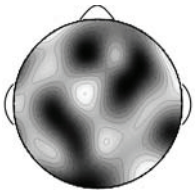
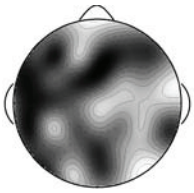
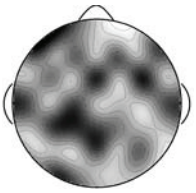
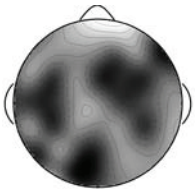
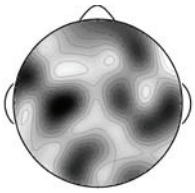
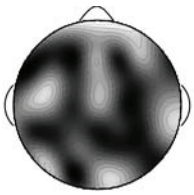
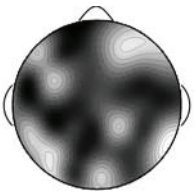
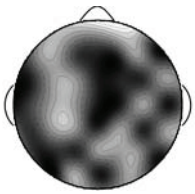
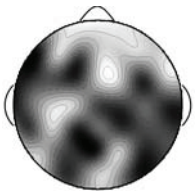
	Subject	Tongue	Nasal	Lips	Fricative	Relax
Jaw	A	68.0, *61.5*	*73.2*, 59.4	*72.4*, *62.0*	*73.7*, *63.0*	*71.9*, *68.8*
	B	66.9, 58.6	*70.1*, *62.2*	*69.6*, 60.7	*71.8*, 60.4	*73.7*, *68.8*
	C	*73.2*, *63.7*	71.3, *61.6*	*74.5*, 59.7	*75.2*, 60.5	*73.9*, *62.4*
	D	*71.7*, *60.7*	*69.8*, *60.7*	*73.0*, *60.7*	*74.8*, *61.2*	*70.8*, *61.9*
	E	*70.9*, *62.9*	*71.6*, *65.4*	*72.4*, *62.4*	69.8, *66.4*	*72.1*, *70.3*
Tongue	A		67.7, *65.6*	*75.3*, *62.2*	*74.0*, *66.4*	*72.1*, *69.5*
	B		71.2, *62.8*	*70.9*, 56.5	67.1, 61.3	*70.5*, *63.7*
	C	-	*74.0*, *64.7*	71.3, 58.9	*75.8*, *62.1*	*70.6*, *61.8*
	D		70.2, *62.6*	*72.2*, 57.7	*74.6*, 57.0	69.1, *60.5*
	E		*76.0*, *66.1*	69.7, *65.0*	70.6, *68.2*	*72.3*, *63.6*
Nasal	A			*72.1*, 58.3	*72.1*, *68.0*	*72.1*, *65.1*
	B			69.3, 58.6	72.1, *64.0*	70.6, *67.0*
	C	-	-	*72.6*, *61.1*	*73.6*, 58.4	*72.1*, *63.2*
	D			66.4, 59.1	*71.5*, 59.8	*70.3*, 59.1
	E			*73.7*, *66.1*	*71.0*, *67.8*	*72.9*, *68.7*
Lips	A				68.8, *62.8*	*75.5*, *69.0*
	B				*69.6*, *62.2*	68.1, *62.8*
	C	-	-	-	*74.8*, *64.5*	*69.0*, 60.0
	D				*73.1*, *60.3*	69.5, 57.7
	E				*72.5*, *68.0*	70.5, *68.9*
Fricative	A					*75.5*, *69.0*
	B					*74.3*, *65.8*
	C	-	-	-	-	*71.0*, *64.5*
	D					*72.8*, 59.8
	E					*71.5*, *66.6*

Performance levels found in two-way classifications of individual phonemes were higher than those found when classifying articulation classes (each comprising a pair of phonemes). Individual electrode performance in pair-wise phoneme classifications in single days exceeded 90% in multiple cases for each subject with one exception: subject D. The results of the top-nine-electrode voting scheme also exceeded 90% in many two-way classifications of phonemes when applied to data from single days; this is true for all five subjects. Performance levels found for trials recorded on all days exceeded 70% in many cases; the maximum such levels were 75%, 73.2%, 73.2%, 70% and 74.8% for Subjects A through E, respectively. While this suggests that variation in the two phonemes chosen to represent an articulation class may hinder discriminability, the phoneme discriminations involved half the number of trials used when discriminating between articulation classes. In fact, the levels of statistical significance associated with the phoneme discriminations are similar to those found for the articulation class discriminations.

Topographies generated by plotting single electrode classification performance suggest that vocal articulation class discrimination depends on activity in areas that include known speech and motor areas, although high variability limits the conclusions that one can draw concerning localization (see Table 3). Scalp topographies for the 15 two-way classifications of articulation class were generated by plotting single electrode classification performance as a function of electrode position. All topographies show considerable variation across articulation class, days and subjects. Some indication of this variability is shown in Table 3, which displays the topographies generated using the entire dataset for the fricative *vs*. relax discrimination in the top row and, in the middle row, the corresponding topographies generated using a single representative day. While some correlation is evident for some

subjects, there is clearly significant variation in single-day topographies. The bottom row of Table 3 shows the topographies for single-electrode performance in the fricative *vs*. nasal discrimination. Note that there is no particular reason for these to resemble the topographies found in the fricative *vs*. relax discrimination (top row) closely. Day-to-day variation and variation among subjects likely contribute to variation among these topographies.

Table 3. *Topographies of LDA classification performance for single electrodes. The top row shows the results found using trials from all days in the fricative vs. relax discrimination for subjects A, B, C, D and E (columns). The middle row shows representative results found using trials recorded on a single day for the fricative vs. relax discrimination. The bottom row shows the results found using trials from all days in the fricative vs. nasal discrimination. The maximum classification rate for a single electrode in the fricative vs. relax comparison is 69.8%; this value is shown as black in all topographies. The maximum classification rate for a single electrode in the fricative vs. nasal comparison is 66.7%. The rate that corresponds to guessing (50%) is shown as white.*

A	B	C	D	E
Fricative vs. Relax, All Trials				
				
Fricative vs. Relax, Trials from a Single Day				
				
Fricative vs. Nasal, All Trials				
				

5. Discussion

Performance levels found with data generated on single days reached as high as 80% in two-way classifications of articulation class, while performance levels found for data generated on all days reached as high as 70%. Classification rates found when discriminating between individual phonemes were higher; indeed, classification using a single electrode provided rates higher than 90% in many single-day phoneme discriminations.

Finding highly significant results in such an experiment is new. We are aware of no similar result for EEG decoding of imagined speech. Indeed, there is every reason to believe that these values may be improved through further analysis of existing data. For one, the present use of spectrogram samples at just five frequencies in just eight time bins is limiting; extending the analysis window in both frequency and time is likely to provide further information useful in classification. Furthermore, the classification performance found with data generated on single days is substantially higher than that

found using data aggregated across several days; the maximum performance rate found using all available data was only 70%. This inter-session variability is common in BCI applications; its practical import is minimized by training at the start of each session.

Following on earlier work by Guenther and colleagues[Guenther et al., 1998, 2006], Tian and Poeppel recently described a model of speech production that includes an efferent signal that provides feedback to motor and auditory areas[Tian and Poeppel, 2010]. It is this efferent signal, in combination with inhibition of motor activity, that is thought to produce auditory and motor imagery during imagined speech. Subjects in the present experiment are instructed to generate both forms of imagery, and we speculate that it is the combination of generated motor imagery and an experimental design involving articulation class that is responsible for the positive results for classifying these classes of phonemes using EEG.

Our immediate goals are to substantiate these findings in further subjects and to conduct further experiments in which a BCI provides feedback concerning the imagined phoneme. Closing the loop in this fashion is likely to enhance classification performance.

Acknowledgements

We thank Robert Coleman, Siyi Deng, Cort Horton, Tom Lappas, and Alvin Li for their help. This work was funded by ARO LS-54228-MUR.

References

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences USA* 98(23), 13367-13372, 2001.

Deng S, Srinivasan R, Lappas T, D'Zmura M. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of Neural Engineering* 7, 046006, 2010.

Deng S, Srinivasan R. Semantic and acoustic analysis of speech by functional networks with distinct time scales. *Brain Research* 1346, 132-144, 2010.

Luo H, Poeppel D. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron* 54, 1001-1010, 2007.

Guenther FH, Ghosh SS, Tourville JA. Neural modeling and the imaging of the cortical interactions underlying syllable production. *Brain and Language* 96, 280-301, 2006.

Guenther FH, Hampson M, Johnson D. A theoretical investigation of reference frames for the planning of speech movements. *Psychological Review* 105, 611-633, 1998.

Stevens KN. *Acoustics Phonetics.* The MIT Press, Cambridge, MA. 243-255, 1998.

Suppes P, Lu ZL, Han B. Brain wave recognition of words. *Proceedings of the National Academy of Science USA* 94, 14965-14969, 1997.

Suppes P, Han B, Lu ZL. Brain wave recognition of sentences. *Proceedings of the National Academy of Science USA* 95, 15861-15866, 1998.

Tian X, Poeppel D. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1, 1-23, 2010.