Joint Spatial-Spectral Feature Space Clustering for Speech Activity Detection from ECoG Signals

Vasileios G. Kanas*, *Student Member*, *IEEE*, Iosif Mporas, Heather L. Benz, Kyriakos N. Sgarbas, Anastasios Bezerianos, *Senior Member*, *IEEE*, and Nathan E. Crone

***Abstract*—Brain machine interfaces for speech restoration have been extensively studied for more than two decades. The success of such a system will depend in part on selecting the best brain recording sites and signal features corresponding to speech production. The purpose of this study was to detect speech activity automatically from electrocorticographic signals based on joint spatial-frequency clustering of the ECoG feature space. For this study, the ECoG signals were recorded while a subject performed two different syllable repetition tasks. We found that the optimal frequency resolution to detect speech activity from ECoG signals was 8 Hz, achieving 98.8% accuracy by employing support vector machines (SVM) as a classifier. We also defined the cortical areas that held the most information about the discrimination of speech and non-speech time intervals. Additionally, the results shed light on the distinct cortical areas associated with the two syllable repetition tasks and may contribute to the development of portable ECoG-based communication.**

***Index Terms*—Brain machine interfaces, electrocorticography, feature space clustering, speech activity detection**

I. Introduction

Over the past two decades, rapid advances in electrophysiological recording technology [1], [2] and novel signal processing techniques have led to the dawn of brain machine interfaces (BMIs) for neurorestoration [3]-[5]. In addition to the rehabilitation of motor deficits [6]-[8], BMI systems could permit silent communication with disabled patients [9]-[27]. Such a speech prosthesis would completely replace the vocal mechanism of a locked-in individual [28] and enable the articulation of words through neural activity alone.

Several BMI communication systems have been proposed in the literature based on electroencephalography (EEG) [29], electrocorticography (ECoG) [30]-[32] or intracortical recordings [33]. Common approaches include letter or word selection using slow cortical potentials (SCP) [12]-[14], the P300 event-related potential (ERP) [15]-[18], steady state visual evoked potentials (SSVEP) [19], [20], sensorimotor rhythms (SMR) [21] and event-related (de)synchronization (ERD/ERS) [22], [23]. In response to the unfulfilled need for fast and natural artificial speech production, recent studies have proposed the prediction of words or phonemes directly from neural signals.

In [24], scalp-recorded EEG was used to discriminate between imagined spoken vowels /a/ and /u/ and a no-action state. Guenther et al. [10] used a wireless intracortical microelectrode array to obtain spike activity related to speech production, and were able to decode formant frequencies. The output of the decoder was used by a synthesizer to produce instantaneous artificial auditory feedback. Pei et al. [25] proposed the discrimination of vowels and consonants of overt and covert word production using ECoG recordings. In another study [26], authors proposed a scheme to classify a small set of spoken words using micro-ECoG arrays on the cortical surface. In a more recent study, Pasley et al. [27] focused on producing spoken words and sentences from ECoG recordings using a stimulus reconstruction model. Although these systems are effective, they are working in a highly controlled environment, such as a laboratory or a clinical setup. To employ speech prostheses in real, out of the lab conditions, several major challenges [34], [35] must be addressed.

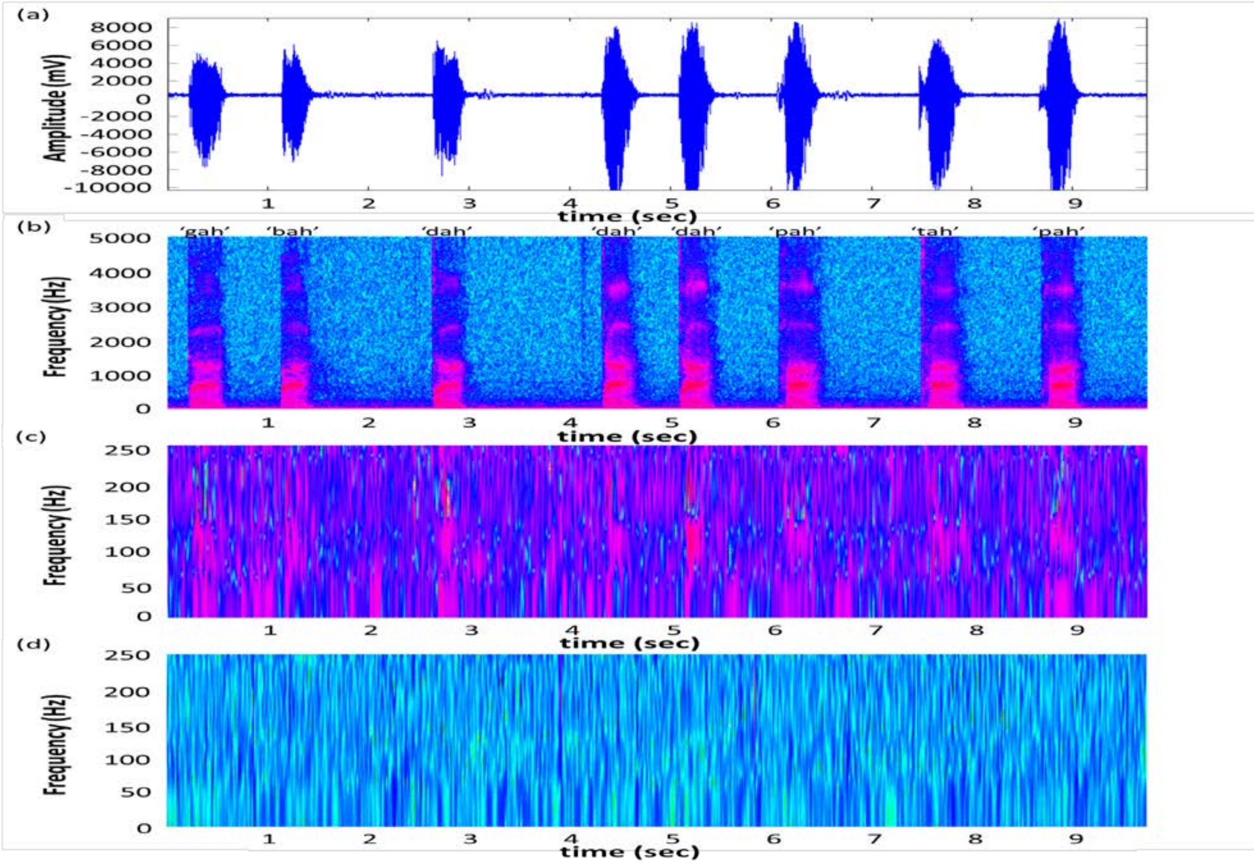

Fig. 1.Raw data and spectrograms of the experimental procedure. (a) Audio waveform of the speech response while the subject articulated several syllables (e.g., 'bah', 'gah', 'dah', 'pah', and 'tah'). (b) Normalized spectrogram of the speech response during the same time period as shown in (a). (c) and (d) Normalized spectrograms of neural data as recorded from the most and least discriminative electrodes 24 and 16, respectively. The electrode 24 was located over posterior superior temporal gyrus and the electrode 16 was located over parietal operculum (see Fig.3).

Prosthetic systems need to interpret the user's current behavioral context (e.g., awake versus sleeping) to minimize power consumption [34]. The proposed experimental protocols in current literature require human intervention to distinguish between speech modes (speech versus silence), resulting in non-autonomous speech prosthetic systems. Neuroprosthetic devices will need to identify these modes autonomously and continuously over time to be viable and acceptable to a patient population in everyday life. Meeting power constraints through the efficient usage of the available resources is a crucial concern for clinically permanently-implantable speech prosthetic systems, and thus, the detection of individual's speech activity (i.e., the time interval in which an individual speaks) is essential for their operation.

In this article, we study for the first time the detection of speech activity from ECoG signals using spectral characteristics extracted from the entire frequency bandwidth. ECoG measures brain potentials without penetrating into cerebral cortical layers, providing an equilibrium between invasiveness and signal fidelity [32], [36].The proposed scheme is based on joint spatial-frequency clustering of the ECoG feature space and exploitation of those clusters of features that most contribute to the discrimination of speech activity time intervals from non-speech intervals. In contrast to similar studies [4], [7], [25], where specific frequency bands are used, here we examine the underlying spectral information from the entire frequency bandwidth. Moreover, we propose a data-driven unsupervised scheme for clustering the feature space to sub-spaces. With this approach we aim to extract the most discriminative features, rather than setting a threshold as in previous studies [3], [4], [9], [26]. Furthermore, the speech activity is jointly studied in the spatial and spectral domains to reveal how the speech activity is organized within different cortical areas and frequency bands.

The remainder of the paper is organized as follows. First the proposed system for speech activity detection is described in section II. Then in section III we present the data used in our analysis, the parameterization of ECoG signals, and feature space clustering and classification methods. In section IV, the experimental results are presented, and the final section is devoted to some discussion and concluding remarks.

II. Speech Activity Detection From ECoG

We assume the speech activity of a subject is encoded in the ECoG signal activity disparately distributed over the cortical area covered by electrodes and over the frequency domain. The speech activity captured by the subdural ECoG electrodes might appear in different frequency bands for each electrode. The present scheme for speech activity detection jointly exploits spatial and spectral information, as captured from the electrodes, without any *a priori* knowledge about the dominant frequency bands in each electrode channel. This approach leverages the underlying network of neural activity responsible for speech production, which is broadly distributed spatially, and the several mechanisms underlying neural oscillations, which include neural spiking [37], suppressing movement when motor activity is not desired [39], and synchronizing distant cortical areas [40].

Our assumption is supported by the ECoG spectrum. Fig. 1(c) and (d) show an example of normalized ECoG time–frequency spectrograms recorded from channels24, superior temporal gyrus (STG), and16, parietal operculum. Fig. 1(a) and (b) illustrate the audio signal and spectrum of subject's voice, respectively. The articulated syllables are also presented. The spectral representation of the speech activity is different in these two channels, resulting in different dominant frequency bands. These differences reflect the different roles of STG and parietal operculum in the speech production and feedback pathway.

The block diagram of the proposed speech activity detection scheme is shown in Fig. 2.During the training phase a set of multichannel ECoG data with known time annotations (i.e., speech/non-speech intervals) is used to train the detector, exploiting those parametric channels (spatial domain) and feature space dimensions (spectral domain) that significantly discriminate the speech intervals from rest. In the test phase the unknown multichannel ECoG signal is processed by the speech activity models and time intervals that correspond to speech activity are detected.

Let us denote the multidimensional ECoG signal, $X = \{x_i\}$, $1 \le i \le I$, with I the number of samples per channel and $x_i \in \mathbb{R}^N$, where N is the number of electrodes. The ECoG signal is initially processed by the Feature Extraction block, in which it is decomposed to a sequence of parametric vectors. Specifically, the Feature Extraction block includes preprocessing of the signals, i.e., frame blocking the signal (separately for each dimension) to overlapping frames of w samples with a time shift between two successive frames equal to s samples, and Hamming windowing each frame. For each of the N electrodes and for each windowed frame, Q spatial-spectral ECoG features are estimated, constructing a feature vector $V \in \mathbb{R}^{N\times Q}$ .The aforementioned parameterization of the ECoG data is presented in Section III with more details.

The short-time parametric representation $V \in \mathbb{R}^{N\times Q}$ of the ECoG training signal is forwarded to the Feature Evaluation block. At this stage the discriminative ability of each feature Q of each of the N dimensions is evaluated with a ranking score $S = \{s_j\}$, $1 \le j \le N\times Q$, indicating for each of the $N\times Q$ parameters the most discriminative to the least discriminative. Evaluation of the parameters is performed to select that subset of features per electrode that most contribute to the accurate detection of speech activity and reject those features that will reduce the overall performance, either because they increase noise or do not add much new information, which can result in a drop in performance [42]. The evaluation of the ECoG parameters is jointly based on spatial (i.e., the selected electrodes) as well as frequency characteristics. In order to select a subset of ECoG parameters with an unsupervised method (instead of, for example, selecting the K -best parameters as in [3], [4], [9], and [26]), we cluster the parameters to C clusters, based on the ranking scores S . The number of clusters C is defined by the user(see Section III). The resulting feature clusters will group the ECoG features per electrode according to their discriminative ability.

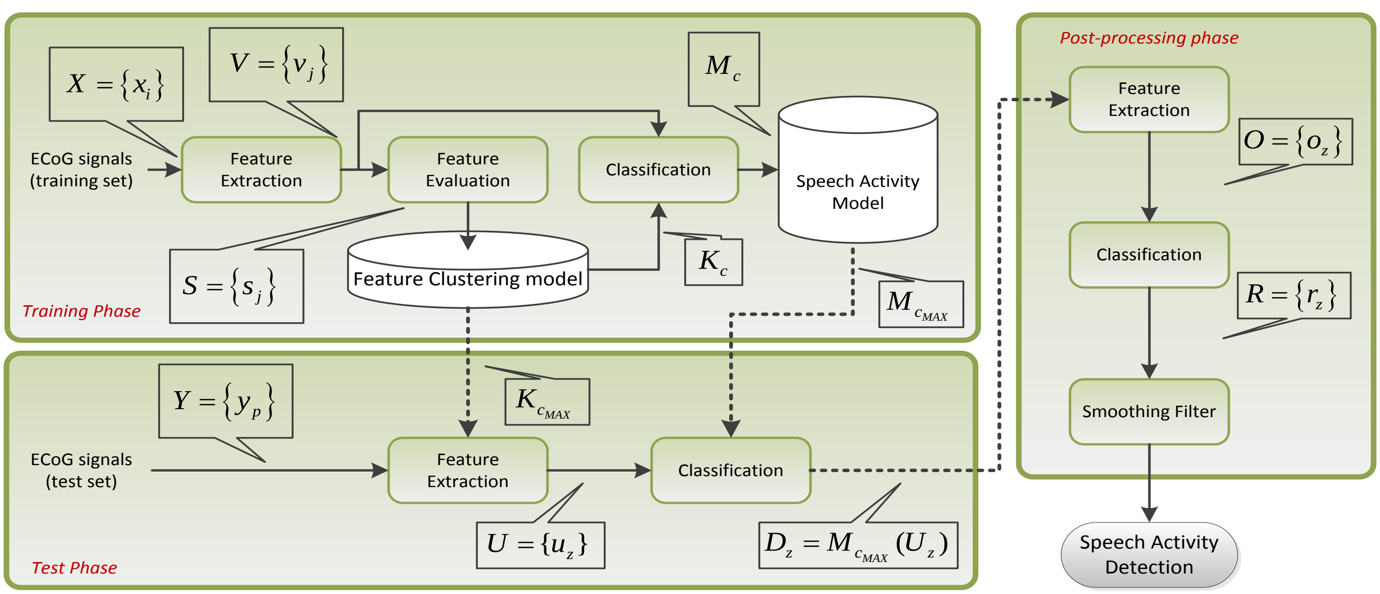

Fig. 2.Block diagram of the proposed scheme for speech activity detection from ECoG signals. The proposed scheme consists of training, test, and post-processing phases.

TABLE I
SYLLABLES PRESENTED TO THE PATIENT DURING THE AUDITORY AND VISUAL PHONEME TASK

Labial		Coronal		Guttural	
Voiced	**Voiceless**	**Voiced**	**Voiceless**	**Voiced**	**Voiceless**
bah	pah	dah	tah	gah	kah
bee	pee	dee	tee	gee	kee

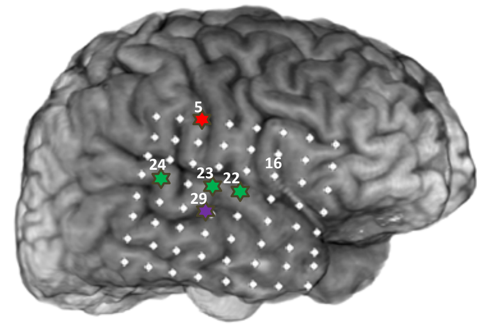

Fig. 3. Electrode locations in the subject. The color coded stars show the five best ECoG electrodes for VAD as calculated by the proposed scheme (see Experimental Results). Color corresponds to cortical area (red: ventral sensorimotor cortex, green: superior temporal gyrus, purple: superior temporal sulcus)

For each combination of feature clusters, a speech activity detection model, $M_c, 1 \le c \le C$, is trained with C total detectors. Specifically, the first detector is trained with the features of the most discriminative feature cluster, the second detector with two most discriminative clusters, and the C -th detector with all clusters, i.e., the full parametric set. The detector with the maximum speech activity detection performance, $M_{c_{MAX}}$, is selected for the test phase.

During the test phase, let the unknown ECoG signal be denoted as $Y = \{y_p\}$, $1 \le p \le P$, with P the number of samples per channel. The test and training signals may be of different lengths, so P is not constrained to be the same as I. Then, $y_p \in \mathbb{R}^N$ is processed by the Feature Extraction block, and the spatio-spectral clustering c_{MAX} from the training phase is used to decompose it to the corresponding feature vector sequence, $U = \{u_z\}$ .Here $u_z \in \mathbb{R}^L, 1 \le z \le Z$ contains only the features that belong to the desired clusters, where Z is the number of test feature vectors (i.e., windowed frames) and L is the number of features in the first $K_{c_{MAX}}$ clusters. The test feature vector sequence is then processed by the corresponding speech activity detection model, $M_{c_{MAX}}$ .Given $D_z = M_{c_{MAX}}(U_z)$ $,1 \le z \le Z$,the probability ($D_z \in [0,1]$) of the z-th frame to include speech activity, a binary decision(speech/non-speech) on frame-level is made.

As a final stage, two-step post-processing over the sequence of frame-based decisions is applied. At the first step, the speech activity probabilities of the current frame as well as the probabilities of the $T \ge 0$ preceding and the $T \ge 0$ successive ECoG frames are concatenated, resulting in a sequence of decision vectors $O = \{o_z\}$, $o_z \in \mathbb{R}^{2 \cdot T+1}$ and $1 \le z \le Z$ test feature vectors. The sequence of decision vectors is used by a post-processing classifier f to produce the final decision $R = \{r_z\}, 1 \le z \le Z$, where $r_z = f(o_z, T)$ .At the second step of post-processing, to eliminate sporadic erroneous labeling of the current ECoG frame, e.g., due to momentary bursts of interference, we smooth each decision r_z with respect to its closest neighbors. In particular, if the $L \ge 0$ preceding and the $L \ge 0$ successive ECoG frames are classified as one label (speech or non-speech), then the current frame is also relabeled as this label.

III. Experimental Setup

The architecture for speech activity detection described in Section II jointly examines the spatial and the spectral information of the multidimensional ECoG signal. We investigate the optimum number of frequency bands that should be used to accurately detect speech activity intervals. In addition to evaluating the overall accuracy of the proposed scheme, we examine the spectral content and the channels that offer the most discriminative information. The location of the electrodes that significantly contribute to the discrimination of speech activity from silence will provide practical information about the cortical areas that contribute to speech processing.

A. Data Description

One male patient with intractable epilepsy participated in this study. ECoG electrodes were implanted for one week to localize his seizure focus for resection. The experimental protocol was approved by the Johns Hopkins Medicine Institutional Review Board, and the patient gave informed consent for this research. The subdural array contained 64 electrodes (Ad-Tech, Racine, Wisconsin; 2.3 mm exposed diameter, with 1 cm spacing between electrode centers) and was placed according to clinical requirements. Electrodes in the array, shown in Fig. 3, covered portions of the frontal, temporal, and parietal lobes of the right hemisphere. Localization of the ECoG electrodes after surgery was performed using Bioimage by co-registration of pre-implantation volumetric MRI with post-implantation volumetric CT [43].

TABLE II
SYSTEM PERFORMANCE (%) USING THE K-BEST FEATURE SUBSPACE CLUSTERS FOR EACH EXPERIMENTAL SETUP DURING THE TEST PHASE

Frequency resolution	***Number of best clusters (K)***				
	M_1	M_2	M_3	M_4	M_5
256 Hz (q = 0)	82.5	83.54	84.97	85.51	85.99
128 Hz (q = 1)	87.12	88.01	88.49	89.27	88.25
64 Hz (q = 2)	88.37	88.31	89.21	89.86	89.92
32 Hz (q = 3)	88.91	88.79	89.39	89.45	89.39
16 Hz (q = 4)	88.85	90.22	90.7	90.22	90.28
8 Hz (q = 5)	**95.25**	93.88	93.45	93.2	92.88
4 Hz (q = 6)	94.4	94.28	92.31	92.19	92.25
2 Hz (q = 7)	93.68	93.38	90.94	90.52	90.46
1 Hz (q = 8)	86.76	86.82	86.82	86.82	86.76

TABLE III
NUMBER OF FEATURES USING THE K-BEST CLUSTERS FOR EACH EXPERIMENTAL SETUP

Frequency resolution	***Number of best clusters (K)***				
	M_1	M_2	M_3	M_4	M_5
256 Hz (q = 0)	15	16	20	36	55
128 Hz (q = 1)	11	52	92	109	110
64 Hz (q = 2)	63	75	132	219	220
32 Hz (q = 3)	207	229	304	306	440
16 Hz (q = 4)	48	180	184	601	880
8 Hz (q = 5)	**380**	986	1650	1657	1760
4 Hz (q = 6)	1154	1198	3204	3506	3520
2 Hz (q = 7)	2462	2535	6378	7019	7040
1 Hz (q = 8)	13840	13900	13959	14016	14080

Data was amplified and recorded through a NeuroPort System (Blackrock Microsystems, Salt Lake City, Utah) at a sampling rate of 10 kHz, and low pass filtered with a cutoff frequency of 500 Hz. The patient's spoken responses were recorded by a Zoom H2 recorder (Samson Technologies, Hauppauge, New York), also at 10 kHz and time-aligned with ECoG recordings.

Two syllable tasks were performed by the patient during ECoG recording. The patient was seated in a hospital bed with a computer screen in front of him on a hospital table. Syllable stimuli were presented to the patient using E-Prime software (Psychology Software Tools, Inc., Sharpsburg, Pennsylvania). The patient was instructed to speak each syllable as it was presented. The syllables were constructed from two vowels ("ah" and "ee") and six consonants, which varied by place of articulation and voiced or voiceless manner of articulation ("p", "b", "t", "d", "k", hard "g"). Table I summarizes the syllables presented to the patient. Each of the 12 syllables was presented 10 times, for a total of 120 trials in each task. Between trials a fixation cross was displayed on the screen for 1,024 ms. In one version of the syllable repetition task, in each trial the patient was presented with written syllables, spelled phonetically, on a computer screen. Each syllable was presented for 3,072 ms. In the auditory version of the task, a recording of each syllable, spoken by an native English speaker, was presented by speakers to the patient, after which the patient repeated the syllable. Each trial was 4,000 ms long.

B. Preprocessing and Parameterization of ECoG data

Prior to any other processing, each recorded dataset is visually inspected and all channels that do not contain clean ECoG signals are excluded, leaving $N = 55$ channels for our analysis. To eliminate any noise common to all channels, recorded data from each ECoG electrode are re-referenced by subtracting the common average (CAR) [44]of electrodes in the same array, as follows,

$$X_{CAR}^{ch} = X^{ch} - \frac{1}{N}\sum_{m=1}^{N} X^{m} \quad (1)$$

where X^{ch} and X_{CAR}^{ch} are the ECoG and CAR referenced ECoG amplitudes on the *ch-th* channel out of a total of N recorded channels. The ECoG signals of each channel are also normalized by subtracting the average value and dividing by the standard deviation. In addition to preprocessing of ECoG recordings, the open source Praat software [45] is used to manually segment the patient's spoken response and label the epochs as *silence*, *speech* and *noise* to train the corresponding models. The noisy intervals are excluded from the evaluation.

The parameterization of the ECoG signals is based on the spectral information in the signals, and the frequency bands that provide the highest performance for the speech activity detection task are examined. In the literature, a variety of ECoG studies have demonstrated that functional activation of cortex is consistently associated with a broadband increase in signal power at high frequencies. Specifically, in [37], [47] the authors examined the 80-100 Hz frequency range, while Canolty et al. [48] used 80-200 Hz high gamma activity to track the spatiotemporal dynamics of word processing. However, to the authors' best knowledge, no previous work has extensively considered the problem of speech activity detection. Therefore, in this study, we also examine the spectral information included in the low frequency bands. To extract the spectral features, each ECoG channel is segmented by applying a sliding Hamming window with length w = 256 samples and shifting steps s = 128 samples. For each of the overlapping frames the power spectral density (PSD) is estimated with the fast Fourier transform (FFT) [46]. Power estimates in the whole frequency range are log-transformed to approximate normal distributions. Each frame is decomposed to a feature vector of dimension 257, consisting of the PSD values estimated every 1 Hz from 0 Hz to 256 Hz, for each ECoG channel. Subsequently, the PSD values are averaged in $Q = 2^q$ frequency bands to obtain the final spectral features per ECoG channel, resulting in different sets of feature vectors $V \in \mathbb{R}^{N\times Q}, q = 0,1,...,8$ .Then a total of 55-14080 features (depending on the number of averaged frequency bands) are used in our analysis. In the remainder of this paper, the average power in the frequency range $[f_1, f_2]$ of the ch-th channel is denoted as $channel(ch)-PSD[f_1, f_2]$. Here we use the term *frequency resolution* to denote the spectral components, with 1 Hz distance between them, contained in each of the $Q = 2^q$ frequency bands.

C. Feature Space Clustering

The discriminative ability of each feature for the speech activity detection task is evaluated to investigate the performance of subsets of features. The PSD features $V \in \mathbb{R}^{N \times Q}$ are ranked using the RelieF algorithm [49] separately for each of the feature vector sets (i.e., for $q = 0,1,...,8$).The k-means algorithm is applied to the ranking scores of each feature, as described in Section II, to group the PSD features into $C = 5$ clusters. The value of C is manually selected. We also tested different values of the C parameter without resulting in better performance. The resulting clusters of the PSD-based feature space are used as inputs to the classification model.

As described in Section II, the cluster $C = 1$ is the group of the most discriminative features and the cluster $C = 5$ is the group of the least discriminative features. Initially, the features of cluster $C = 1$ are used to train the classification model M_1. The classification model M_2 is trained using the feature sub-space from the clusters $C = 1$ and $C = 2$, and so on. The final classification model M_5 is trained using the whole feature space.

D. Classification

Fig. 4. Feature ranking maps as calculated by the RelieF algorithm. From top to bottom and from left to right the feature ranking maps represent the ranking scores per channel and frequency band for frequency resolutions 32, 16, 8, and 4 Hz.

Fig. 5. (a) The average ranking scores per frequency for q = 5. The spectral information is located in low (0-48 Hz) and high (168-208 Hz) frequencies. (b)The average ranking scores per channel for q = 5. The five best channels are 24, 29, 23, 22 and 5.

For the classification block we rely on five well-known machine learning algorithms that have been used in similar tasks in the literature [24], [50]-[53]. These algorithms are: support vector machines (SVMs) using the sequential minimal optimization algorithm [54], multilayer perceptron neural networks (MLP) using a 3-layered structure [55], the k-nearest neighbors (kNN) algorithm [56], the C4.5 decision tree (J48) [57], and linear logistic regression [58]. SVMs are found to outperform the other classification algorithms and achieved classification accuracy of 95.25%, while the second best classifier, MLP, achieved 92.90%. We use the radial basis function (RBF) for the SVM kernel. The RBF values C =10.0 and γ =0.01 are found to offer optimal classification performance after a grid search at all combinations of C ={1.0, 5.0, 10.0, 20.0} and γ = {0.001, 0.01, 0.1, 0.5, 1.0, 2.0}.The evaluation of the results is performed using 10-fold cross validation and the accuracy was computed as a fraction of the number of correctly identified speech windows to the total number of actual speech windows.

IV. Experimental Results

The architecture for speech activity detection presented in Section II is evaluated according to the experimental setup and protocol presented in Section III. In the following we present the experimental results for the evaluated ECoG data using the SVM classification algorithm.

A. *Speech Activity Detection Performance*

The speech activity detection performance for the nine frequency vector sets using the K -best feature subspace clusters (i.e., classification model $M_c, 1 \le c \le C$) is shown in Table II. The best classification accuracy (95.25%) is achieved for q =5 and K =1,

TABLE IV
SYSTEM PERFORMANCE (%) AFTER THE IMPLEMENTATION OF THE TWO-STEP POST-PROCESSING STEP

Number of frames smoothed	Number of adjacent frames needed to classify a label T=0	T=1	T=2	T=3
L=0	95.25	96.17	96.05	95.80
L =1	97.13	97.68	97.39	97.26
L =2	97.62	**98.84**	98.31	98.01
L =3	96.56	97.22	97.09	96.98

TABLE V
RANKING SCORES OF THE 10-BEST ECOG FEATURES AS EVALUATED BY THE RELIEF ALGORITHM ACHIEVING THE HIGHEST PERFORMANCE

Ranking	*ECoG features (V)*	*Ranking Scores (S)*
1	channel(24)-PSD(120-128 Hz)	0.029
2	channel(24)-PSD(112-120 Hz)	0.026
3	channel(24)-PSD(104-112 Hz)	0.023
4	channel(24)-PSD(128-136 Hz)	0.023
5	channel(22)-PSD(176-184 Hz)	0.018
6	channel(22)-PSD(184-192 Hz)	0.018
7	channel(24)-PSD(96-104 Hz)	0.017
8	channel(23)-PSD(184-192 Hz)	0.014
9	channel(5)-PSD(120-128 Hz)	0.013
10	channel(24)-PSD(136-144 Hz)	0.013

which represents averaged PSD values equally distributed at 32 frequency bands (each of the 32 bands corresponds to resolution of 8 Hz) and for the single best feature subspace cluster (a feature vector with 380 elements) (Table III).

The use of other than 32 bands, or the parameterization of the ECoG signals at a resolution higher or lower than 8 Hz, results in a drop of the speech activity detection performance. Moreover, the use of more clusters of the features than the best ranked one not only does not offer improvement to the overall speech activity detection performance, but also results in a significant reduction of it, especially when using all subspace clusters. Since the clustering is performed using joint spatial-spectral criteria, this drop is an indication that some of the channels and frequency bands do not carry useful information for the speech discrimination task and thus overtrain the classification model with useless and noisy information. Further analysis on this effect appears in the following section.

The effect of the post-processing stage is evaluated for different values of the parameters T, related to the number of adjacent frames required to reclassify a label, and L, related to the number of frames smoothed after the classifier decision. The performance results after the application of the post-processing stage are shown in Table IV. The best performance, 98.84%, is achieved for $L = 2$ and $T = 1$, which corresponds to the fusion of the two preceding and two succeeding speech probabilities and the smoothing of decisions within a window of three frames length. This accuracy indicates the efficiency of the post-processing stage, which improved the speech activity detection by 3.59% in absolute performance.

B. Feature Ranking Maps

The extracted features describe the ECoG activity during speech in the spatial and spectral domains. To investigate which cortical areas and frequency bands contribute to speech activity detection we performed a feature ranking evaluation using the ReliefF algorithm, as shown in Fig.4. The ranking maps depict the ranking scores per channel and frequency band. These figures point out which of the channels and frequency bands hold most of the information about the speech activity (intensity denotes the ranking scores, and a darker color corresponds to a more discriminative spatio-spectral feature).For frequency resolution 8 Hz (q = 5), the most information is present in very high frequencies on most channels, while channel 24, located over posterior STG, held information in the high gamma range between 88-144 Hz. The most informative feature is the average 120-128 Hz high gamma power of channel 24, with a ranking score of 0.029 as calculated by the ReliefF algorithm. Channel 24's utility in discrimination is also apparent in Table V, which shows the 10best features for speech discrimination.

To reveal which channels and frequency bands are most informative about speech activity, we average the feature ranking map, corresponding to the optimal accuracy (q=5), across each ECoG channel and frequency band separately. Fig. 5(a) illustrates the average ranking scores per frequency band. There are two distinct informative regions. The first region, having the highest

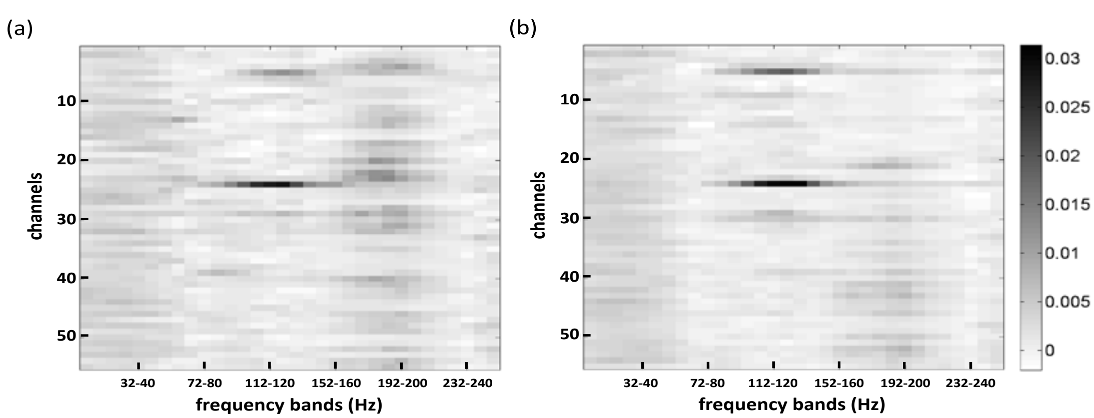

Fig. 6. Feature ranking maps as calculated by the ReliefF algorithm for the a) auditory and b) visual phoneme tasks separately using the optimal frequency resolution of 8 Hz.

Fig. 7. a) The average ranking scores by frequency, b) the average ranking scores by channel, and c) the five dominant channels corresponding to the visual (column 1) and auditory (column 2) phoneme task. The channels 22 and 32 are equally ranked. Color corresponds to cortical areas (red: ventral sensorimotor cortex, green: superior temporal gyrus, purple: superior temporal sulcus, yellow: inferior temporal gyrus, light blue: middle temporal gyrus). The frequency resolution is 8 Hz.

ranking scores, is in very high frequency bands (168-216 Hz), while the second one is in lower frequencies (0-48 Hz).Fig. 5(b) shows the average ranking scores per channel. The five most important electrodes are the 24, 29, 23, 22 and 5,which are also marked in Fig. 3.These electrodes are located in cortical areas typically involved in speech and language processing, although they are on the right (non-dominant) hemisphere. Channel 5 is located over ventral sensorimotor cortex, which is involved in the motor production of speech and somatosensory feedback. Channels 24, 23 and 22 are located over the posterior STG, which contains auditory association cortex and is part of Wernicke's area in the left hemisphere, typically important for speech perception. Channel 29 is located over superior temporal sulcus, which is used in speech processing.

Finally, we examine the detection of speech activity separately for each syllable task. Our analysis shows that, once again, the use of 32 frequency bands resulted in the best detection performance using the single best feature subspace cluster (95.35% for the auditory phoneme task and 90.34% for the visual phoneme task). The feature ranking maps are shown in Fig. 6. In the visual phoneme task (Fig. 6b, 7), the high frequencies between 152-200 Hz carry less information than they do for speech activity detection during the auditory phoneme task (Fig. 6a, 7). Additionally, the lower frequencies (0-40 Hz) hold more discriminative information than high gamma frequencies, in contrast to the auditory phoneme task, where high frequencies (176-200 Hz) are more informative than the low frequencies. The lower frequencies (0-40 Hz) are similarly informative for the two tasks. The most informative channels for the auditory task are 24, 23, 29, 22and 5, all of which are discussed above, and 32, located over middle temporal gyrus (MTG), which is involved in auditory and language processing. The channels 22 and 32 are ranked equally. For the visual task the most informative channels are 24 and 5, discussed above; 30, located over MTG; 52, located over inferior temporal gyrus (ITG); and 39, located over middle temporal gyrus (MTG) (Fig. 7). ITG is a component in the visual processing stream. The involvement of MTG is consistent with the language processing necessary for both tasks [61]. The involvement of ITG in the visual task, but not in the auditory repetition task, is also to be expected [62].

In both tasks, channels 24 (posterior STG) and 5 (ventral sensorimotor cortex) are highly informative at frequencies in the 88-144 Hz range (high gamma oscillations). It is likely that these channels contribute so significantly to the decoding accuracy because they are located in cortical areas that are related to the production of speech and auditory and sensory feedback. Pasley et al. have demonstrated that in the left hemisphere posterior STG encodes the acoustic information in speech [27], and left sensorimotor cortex has previously been used to decode three vowels [10]. High gamma oscillations reflect local population firing [37] and are an index of cortical processing in these key speech production and feedback areas.

V. Discussion and Conclusions

In this study, we propose a framework for speech activity detection from ECoG signals with high accuracy, using unsupervised feature space clustering. We demonstrate that speech-related activity is represented in a variety of frequency bands in electrodes in relevant cortical areas. We explore the spectral information in the ECoG channels, examining the frequency bands that provided the highest performance for the speech activity task. Our results give evidence that 32 frequency bands are optimal for detecting human articulation. At the same time the fact that distributed locations hold information about speech activity, suggests that language processing involves large-scale cortical networks that are engaged in phonological analysis, speech articulation and other processes [36].Moreover, our results show that in addition to high gamma frequencies, lower frequencies are useful for speech activity detection.

The electrodes that most contribute to the high classification accuracy are located over cortical areas relevant to speech in the right hemisphere: posterior STG (3 electrodes) [59], superior temporal sulcus (1 electrode) [60], and ventral sensorimotor cortex (1 electrode) [47]. The spatial distribution of these electrodes highlights the importance of large-scale cortical networks in speech production, and therefore in speech detection. The importance of the high gamma contributions to speech detection, especially from the posterior STG electrode, is consistent with the view that high gamma ECoG activity is related to the underlying population spiking activity [37]. The robustness of low frequency contributions to speech detection may reflect the role of beta oscillations in gating motor activity [39] and theta oscillations in synchronizing distant cortical areas involved in processing for a task [40].In conclusion, to our best knowledge, this study has validated for the first time the feasibility of speech activity detection from ECoG signals. Thus, no direct comparison with other approaches is feasible. Instead of detecting speech activity, several approaches have been proposed to decode semantic information [63], control a one-dimensional computer cursor using phoneme articulation [64], discriminate between different phonemes [24], [25] and words [26], and reconstruct speech [27]. Further research is needed to extend our results to word articulation. In particular, the use of information acquired from causal interactions between cortical areas should prove useful. The approach described here for selecting optimal features and applying classifiers to labeled epochs of speech data may be applied to other decoding problems beyond speech detection. For example, if labels reflected spoken or imagined phonemes rather than speech and non-speech epochs, a classifier could be trained to discriminate different phonemes using this method. Such a decoder would require ECoG signals from speech motor cortex or language areas in the frontal and temporal lobes. These results support constructing a speech BCI in a hierarchical fashion, with the speech detector described here segmenting data during classifier training and online operation, and a decoder processing only ECoG data related to speech epochs.

References

[1] L.-D. Liao, C.-T. Lin, K. McDowell, A.E. Wickenden, K. Gramann, T.-P. Jung, L.-W. Ko, and J.-Y. Chang, "Biosensor technologies for augmented brain computer interfaces in the next decades," *Proc. IEEE*, vol. 100,no. 5, pp. 1553-1566, May 2012,.

[2] K. Mcdowell, C.-T. Lin, K.S. Oie, T.-P. Jung, S. Gordon, K.W. Whitaker, S.-Y. Li, S.-W. Lu, W.D. Hairston, "Real-World Neuroimaging Technologies," *Access, IEEE*, vol. 1, no., pp.131-149, May 2013.

[3] H. Benz, H. Zhang, A. Bezerianos, S. Acharya, N.E. Crone, X. Zheng, and N.V. Thakor, "Connectivity analysis as a novel approach to motor decoding for prosthesis control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, pp. 143-152, Mar. 2012.

[4] G. Schalk, J. Kubanek, K.J. Miller, N.R. Anderson, E.C. Leuthardt, J.G. Ojemann, D. Limbricj, D. Moran, L.A. Gerhardt, and J.W. Wolpaw, "Decoding two-dimensional movement trajectories using electrocorticographic signals in humans," *J. Neural Eng.*, vol.4, pp. 264-275, Sem. 2007.

[5] M.S. Fifer, S. Acharya, H.L. Benz, M. Mollazadeh, N.E. Crone, and N.V. Thakor, "Toward Electrocorticographic Control of a Dexterous Upper Limb Prosthesis: Building Brain-Machine Interfaces," *IEEE Pulse*, vol. 3, pp. 38-42, Jan. 2012.

[6] M.L. Stavrinou, L. Moraru, L. Cimponeriu, S. Della Penna, and A. Bezerianos, "Evaluation of Cortical Connectivity During Real and Imagined Rhythmic Finger Tapping," *Brain topography*, vol. 19, pp. 137-145, Mar. 2007.

[7] J. Kubanek, K.J. Miller, J.G. Ojemann, J.R. Wolpaw, and G. Schalk "Decoding flexion of individual fingers using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 6, 066001, pp. 14, Dec. 2009.

[8] J.M. Carmena, M.A. Lebedev, R.E. Crist, J.E. O'Doherty, D.M. Santucci, D.F Dimitrov, P.G Patil, C.S Henriquez, and M.A. Nicolelis. "Learning to Control a Brain–Machine Interface for Reaching and Grasping by Primates," *PLoS Biology*, vol. 1, e42, pp. 16, Oct. 2003.

[9] X. Pei, J. Hill, and G. Schalk, "Silent communication: Toward using brain signals," *IEEE Pulse*, vol. 3, pp. 43-46, Jan. 2012.

[10] F.H. Guenther, J.S. Brumberg, E.J. Wright, A. Nieto-Castanon, J.A. Tourville et al., "A wireless brain–machine interface for real-time speech synthesis," *PloS Biology*, vol. 4, e8218, pp. 11, Dec. 2009.

[11] T. Hinterberger, A. Kubler, J. Kaiser, N. Neumann, and N. Birbaumer, "A brain-computer interface (BCI) for the locked-in: comparison of different EEG classifications for the thought translation device," *Clin Neurophysiol*, vol. 114, pp. 416–425, Mar. 2003.

[12] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor, "A spelling device for the paralysed," *Nature*, vol. 398, pp.297–298, Mar.1999.

[13] N. Birbaumer, A. Kubler, N. Ghanayim, T. Hinterberger, J. Perelmouter, J. Kaiser, I. Iversen, B. Kotchoubey, N. Neumann, and H. Flor, "The thought translation device (TTD) for completely paralyzed patients," *IEEE Trans. Neural. Syst. Rehabil. Eng*, vol. 8, pp. 190–193, Jun. 2000.

[14] N. Birbaumer, T. Hinterberger, A. Kübler, and N. Neumann, "The thought-translation device (TTD): neurobehavioral mechanisms and clinical outcome," *IEEE Trans. Neural. Syst. Rehabil. Eng*, vol. 11, pp. 120–123, Jun. 2003.

[15] E. Donchin, K. Spencer, and R. Wijesinghe, "The mental prosthesis: assessing the speed of a P300-based brain-computer interface," *IEEE Trans. Neural. Syst. Rehabil. Eng*, vol.8, pp. 174–179, Jun. 2000.

[16] F. Nijboer, E. Sellers, J. Mellinger, M.A. Jordan, T. Matuz, A. Furdea, S. Halder, U. Mochty, D.J. Krusienski, T.M. Vaughan, J.R. Wolpaw, N. Birbaumer, and A. Kübler, "A P300-based brain-computer interface for people with amyotrophic lateral sclerosis," *Clin. Neurophysiol*, vol. 119, pp. 1909–1916, Aug. 2008.

[17] D.J. Krusienski, E.W. Sellers, D.J McFarland, T.M. Vaughan, and J.R. Wolpaw, "Toward enhanced P300 speller performance," *J. Neurosci. Methods*, vol. 167, pp. 15–21, Jan. 2008.

[18] E.W Sellers, and E. Donchin, "A P300-based brain-computer interface: initial tests by patients ALS," *Clin. Neurophysiol*, vol. 117, pp. 538–548, Mar. 2006.

[19] M. Cheng, X. Gao, S. Gao, and D. Xu, "Design and implementation of a brain-computer interface with high transfer rates," *IEEE Trans. Biomed. Eng*, vol. 49, pp. 1181–1186, Oct. 2002.

[20] O. Friman,T. Luth, I. Volosyak, and A. Graser, "Spelling with steady-state visual evoked potentials," *3rd International IEEE/EMBS Conference on Neural Engineering*, Kohala Coast, May 2007 pp. 354-357.

[21] T. Vaughan, D. McFarland, G. Schalk, W.A. Sarnacki, D.J. Krusienski, E.W. Sellers, and J.R. Wolpaw, "The wadsworth BCI research and development program at home with BCI," *IEEE Trans. Neural. Syst. Rehabil. Eng*, vol. 14, pp. 229–233, Jun. 2006.

[22] C. Neuper, G.R. Müller-Putz, R. Scherer, and G. Pfurtscheller, "Motor imagery and EEG-based control of spelling devices and neuroprostheses", *In Progress in Brain Research*, Christa Neuper and Wolfgang Klimesch, Ed. Elsevier, vol.159, pp. 393-409, 2006.

[23] R. Scherer, G. Muller, C. Neuper, B. Graimann, and G. Pfurtscheller, "An asynchronously controlled EEG based virtual keyboard: improvement of the spelling rate," *IEEE Trans. Biomed. Eng*, vol. 51, pp. 979–984, Jun. 2004.

[24] C.S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," *Neural Networks*, vol. 22, pp. 1334-1339, Nov. 2009.

[25] X. Pei, D.L Barbour, E.C. Leuthardt, and G. Schalk, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans," *J. Neural Eng.*, vol. 8, 046028, pp. 11, Aug. 2011.

[26] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House and B. Greger. "Decoding spoken words using local field potentials recorded from the cortical surface," *J. Neural Eng.*, vol. 7, 056007, pp. 10, Oct. 2010.

[27] B.N. Pasley, S.V. David, N. Mesgarani, A Flinker, S.A. Shamma, N.E. Crone, R.T. Knight, and E.F. Chang, "Reconstructing Speech from Human Auditory Cortex," PloS Biology, vol. 10, e1001251, pp. 13, Jan. 2012.

[28] E. Smith and M. Delargy, "Locked-in syndrome," *Br. Med. Journal*, vol 330, pp.406–409, Feb. 2005.

[29] F. Niedermeyer, L. Da Silva, *Electroencephalography, Basic Principles and related fields*, Lippinkott Williams and Wilkins, 2011.

[30] C. Henle, M. Schuettler, J. Rickert, and T. Stieglitz, "Towards Electrocorticographic Electrodes for Chronic Use in BCI Applications," *in Towards Practical Brain-Computer Interfaces*, Springer Berlin Heidelberg,pp.85–103, 2013.

[31] G. Schalk, and E.C. Leuthardt, "Brain-Computer Interfaces Using Electrocorticographic Signals," *IEEE Reviews in Biomedical Engineering*, vol. 4, no., pp. 140–154, Oct. 2011.

[32] Z.C. Chao, Y. Nagasaka, and N. Fujii, "Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys," *Front. Neuroeng*., vol. 3, Article 3, pp. 10, Mar. 2010.

[33] S. Gibson, J.W. Judy, D. Markovic, "Spike Sorting: The First Step in Decoding the Brain," *Signal Processing Magazine IEEE*, vol.29, no.1, Jan. 2012, pp. 124-143.

[34] M.D. Linderman, G. Santhanam, C.T. Kemere, V. Gilja, S. O'Driscoll, B.M. Yu, A. Afshar, S.I. Ryu, K.V. Shenoy, and T.H. Meng, "Signal Processing Challenges for Neural Prostheses," *Signal Processing Magazine IEEE*, vol.25, no.1, pp.18-28, 2008.

[35] N. Achtman, A. Afshar, G. Santhanam, B.M. Yu, S.I. Ryu, and K.V. Shenoy, "Free-paced high-performance brain computer interfaces," *J. Neural Eng.*, vol. 4, no. 3, pp. 336–347, Sep. 2007.

[36] A. Korzeniewska, P.J. Franaszczuk, C.M. Crainiceanu, R. Kuś, and N.E. Crone, "Dynamics of large-scale cortical interactions at high gamma frequencies during word production: Event related causality (ERC) analysis of human electrocorticography (ECoG)," *NeuroImage*, vol. 56, pp. 2218–2237, Jun. 2011.

[37] N.E. Crone, D. Boatman, B. Gordon, and L. Hao, "Induced electrocorticographic gamma activity during auditory perception," Brazier Award-winning article, *Clin. Neurophysiol*., vol. 112, pp. 565–582, Apr. 2001.

[38] S. Ray, N.E. Crone, E. Niebur, P.J. Franaszczuk, and S.S. Hsiao, "Neural correlates of high-gamma oscillations (60–200 Hz) in macaque local field potentials and their potential implications in electrocorticography," *The Journal of Neuroscience*, vol. 28, pp. 11526-11536, Nov. 2008.

[39] K.J. Miller, D. Hermes, C.J. Honey, A.O. Hebb, N.F. Ramsey, R.T. Knight, J.G. Ojemann, and E.E Fetz, "Human motor cortical activity is selectively phase-entrained on underlying rhythms," *PLoS computational biology*, vol. 8, e1002655, pp. 21, Sep. 2012.

[40] J.E. Lisman, and O. Jensen, "The θ-γ neural code," *Neuron*, vol. 77, pp. 1002-1016, Mar. 2013.

[41] A.V. Oppenheim, R.W. Schafer, and J.R. Buck, *Discrete-time signal processing*. Englewood Cliffs: Prentice-hall, 1989.

[42] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification*. Wiley-interscience, 2012.

[43] J.S. Duncan, X. Papademetris, J. Yang, M. Jackowski, X. Zeng, L.H. Staib, "Geometric strategies for neuroanatomic analysis from MRI," *Neuroimage*, vol. 23,Suppl. 1, pp. S34-45, 2004.

[44] D. Goldman, "The clinical use of the 'average' reference electrode in monopolar recording," *Electroencephalogr. Clin. Neurophysiol*., vol. 2, pp. 209-212, May 1950.

[45] P. Boersma, D. Weeninck, "Praat, a system for doing phonetics by computer," *Glot. International*, vol. 5, no. 9/10, pp. 341-345, 2001.

[46] R.N. Bracewell, "The Fourier transform," *Sci. Am*., vol. 260, pp. 86-9, 92-5, Jun. 1989.

[47] N.E. Crone, L. Hao, J. Hart, D. Boatman, R.P. Lesser, R. Irizarry, and B. Gordon, "Electrocorticographic gamma activity during word production in spoken and sign language," *Neurology*, vol. 57, pp. 2045-2053, Dec. 2001.

[48] R.T. Canolty, M. Soltani, S.S. Dalal, E. Edwards, N.F. Dronkers, S.S. Nagarajan, H.E. Kirsch, N.M. Barbaro, and R.T. Knight "Spatiotemporal dynamics of word processing in the human brain," *Front Neurosci*, vol. 1, pp.1185–1196, Oct. 2007.

[49] I. Kononenko. "Estimating Attributes: Analysis and Extensions of RELIEF, "*In Proc. of the European Conference on Machine Learning*, pp. 171-182, 1994.

[50] A. Bashashati, M. Fatourechi, K.W. Rabab and G.E. Birch, "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals," *J. Neural Eng.*, vol. 4, pp. 32-57, Jun. 2007.

[51] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol 4, pp. R1-R13, Jun. 2007.

[52] E. Haselsteiner, and G. Pfurtscheller ,"Using time-dependant neural networks for EEG classification," *IEEE Trans. Rehabil. Eng*., vol. 8, pp. 457–63, Dec. 2000.

[53] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical pattern recognition: a review," *IEEE Trans Pattern Anal Mach Intell*, vol. 22, pp. 4–37, Jan. 2000.

[54] J. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," *In Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf and C. Burges and A. Smola, 1998.

[55] A.K. Jain,M. Jianchang, and K.M. Mohiuddin, "Artificial neural networks: a tutorial," *Computer*, vol.29,pp.31-44, Mar 1996.

[56] D. Aha, and D. Kibler, "Instance-based learning algorithms," *Machine Learning*, vol. 6, pp. 37-66, Jan. 1991.

[57] Ross Quinlan, *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[58] N. Landwehr, M. Hall, and E. Frank, "Logistic Model Trees," *Machine Learning*, vol. 59, pp. 161-205, May 2005.

[59] E. Edwards, S.S. Nagarajan, S.S. Dalal, R.T. Canolty, H.E. Kirsch N.M. Barbaro, and R.T. Knight, "Spatiotemporal imaging of cortical activation during verb generation and picture naming." *Neuroimage*, vol. 50, pp. 291-301, Mar. 2010.

[60] V.L. Towle, H.A. Yoon, M. Castelle, J.C. Edgar, N.M. Biassou, D.M., Frim, J.-P. Spire, and M.H. Kohrman, "ECoG gamma activity during a language task: differentiating expressive and receptive speech areas," *Brain*, vol. 131, pp. 2013-2027, Jul. 2008.

[61] I. Laufer, M. Negishi, C.M. Lacadie, X. Papademetris, and R.T. Constable, "Dissociation between the activity of the right middle frontal gyrus and the middle temporal gyrus in processing semantic priming," *PloS one*, vol. 6, e22368, pp. 10, Aug. 2011.

[62] E.J. Jensen, I. Hargreaves, A. Bass, P. Pexman, B.G. Goodyear, and P. Federico, "Cortical reorganization and reduced efficiency of visual word recognition in right temporal lobe epilepsy: A functional MRI study," *Epilepsy research*, vol. 93,pp. 155-163, Feb. 2011.

[63] W. Wang, A.D. Degenhart, G.P. Sudre, D.A. Pomerleau, E.C. Tyler-Kabara, "Decoding semantic information from human electrocorticographic (ECoG) signals," *Engineering in Medicine and Biology Society, Annual International Conference of the IEEE*, Boston, Aug. 2011, pp.6294-6298.

[64] E.C Leuthardt, C. Gaona, M. Sharma, N. Szrama, J. Roland, Z. Freudenberg, J. Solis, J. Breshears, G. Schalk, "Using the electrocorticographic speech network to control a brain–computer interface in humans." *J. Neural Eng.*, vol. 8, 036004, pp. 11, Apr.2011.