Voice Activity Detection from Electrocorticographic Signals

V.G Kanas[1], I. Mporas[1],H.L. Benz[2], N. Huang[2], N.V. Thakor[3],K. Sgarbas[1], A. Bezerianos[3] and N.E. Crone[4]

[1]Department of Electrical and Computer Engineering, University of Patras, Patras, Greece
[2]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205USA
[3]Singapore Institute for Neurotechnology, National University of Singapore, Singapore
[4] Department of Neurology, Johns Hopkins University, Baltimore, MD 21205 USA

Abstract— **The purpose of this study was to explore voice activity detection (VAD) in a subject with implanted electrocorticographic (ECoG) electrodes. Accurate VAD is an important preliminary step before decoding and reconstructing speech from ECoG. For this study we used ECoG signals recorded while a subject performed a picture naming task. We extracted time-domain features from the raw ECoG and spectral features from the ECoG high gamma band (70-110Hz). The RelieF algorithm was used for selecting a subset of features to use with seven machine learning algorithms for classification. With this approach we were able to detect voice activity from ECoG signals, achieving a high accuracy using the 100 best features from all electrodes (96%) or only 12 features from the two best electrodes (94%) using the support vector machines or a linear regression classifier. These findings may contribute to the development of ECoG-based brain machine interface (BMI) systems for rehabilitating individuals with communication impairments.**

Keywords— **Voice activity detection, electrocorticography, brain machine interface, machine learning**

I. Introduction

Brain-machine interface (BMI) systems attempt to rehabilitate paralyzed individuals and simultaneously allow direct communication between the human brain and an external machine [1]. In addition to BMIs for cursor control [2] and limb prosthetic control [3][4][5][6], there is growing interest in BMIs for restoring speech function, for example in patients with profound articulatory impairment [7][8].

Detecting and decoding speech from cortical activity is more complex than decoding movement. Language involves large-scale cortical networks that are dynamically engaged in phonological analysis, speech articulation and other processes [9][10]. In [11], the authors aimed to discriminate three different tasks from EEG recordings (imagined speech of vowels /a/ and /u/ and a no action state). They designed spatial filters using the common spatial pattern (CSP) method to extract features and the support vector machine (SVM) algorithm to discriminate the three tasks. More recently, in [12] the discrimination of vowels and consonants of overt and covert word production using ECoG recordingswas proposed. They used spectral amplitudes in different frequency bands, estimated via an autoregressive (AR) model, and the local motor potential (LMP) as features and a Naïve Bayes model as classifier. ECoG recordings [13]have become recognized as an in vivo biocompatible option for a wireless chronically implantable BMI system. Such a system could be utilized as an assistive device to enable disabled individuals to produce speech through neural activity. This device could be feasible through speech reconstruction [14] from cortical brain activity. However, the first step before exploring decoding or reconstruction methodologies is to use neural activity to detect an individual's speech, i.e. to define the time intervals in which a subject speaks.

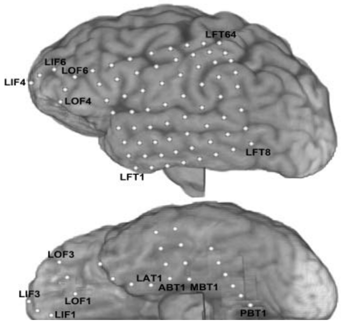

Fig. 1 The subject's electrode locations.

Voice activity detection (VAD) has been studied for more than ten years in the speech technology community [16]. The focus of VAD is to detect the time intervals in which vocal communication is occurring, especially when its acoustic spectrum overlaps that of other sources, such as music or noises. Current experimental protocols require human intervention to differentiate between such intervals; to be useful outside the laboratory, however, prosthetic systems will need to determine the epoch autonomously [15]. In this study, we propose a VAD framework using ECoG signals as a preliminary step before speech decoding methodologies. We intend to detect voice activity automatically by exploiting machine learning techniques, including feature selection and classification.

II. ECOG DATA COLLECTION

A. Subject and Data collection

This experiment was conducted with one male patient with intractable epilepsy, who was temporarily implanted with subdural electrode arrays (Ad-Tech, Racine, Wisconsin; 2.3 mm exposed diameter, with 1 cm spacing between electrode centers) 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 electrode placement was chosen based on the clinical requirements of the patient, guided by estimation of the seizure focus prior to surgery. Localization of the ECoG electrodes after surgery was performed by co-registration of pre-implantation volumetric MRI with post-implantation volumetric CT using Bioimage [17].

One large eight-by-eight grid of electrodes was placed over the left hemisphere of the brain, covering portions of the frontal, temporal, and parietal lobes (LFT1-64). Additional electrode strips were placed on the frontal lobe and on the basal temporal surface (Fig. 1).Data was amplified and recorded through a NeuroPort System (Blackrock Microsystems, Salt Lake City, Utah) at a sampling rate of 10 kHz, low pass filtered with a cutoff frequency of 500 Hz, and then downsampled to 1 kHz. The patient's spoken responses were recorded by a Zoom H2 recorder (Samson Technologies, Hauppauge, New York), also at 10 kHz but without subsequent downsampling. Each dataset was visually inspected and all channels that did not contain clean ECoG signals were excluded. Channels 2, 12, 23, 44 and 53 were removed, which left 89 channels for our analyses.

B. Experimental Protocol

A picture naming task was performed by the patient during ECoG recording. The patient was seated in a hospital bed. During each trial, an image was presented to the patient for one second using E-Prime software (Psychology Software Tools, Inc., Sharpsburg, Pennsylvania). The patient was instructed to name the image as quickly as possible, or to say 'pass' if unable to name the image. Between trials, a fixation cross was presented for an average of seven seconds, with a jitter of up to a half second to avoid stimulus anticipation by the subject. Three hundred and fifty trials were conducted in one session lasting around one hour, with a short break in the middle.We used the open source Praat software [18] to segment and label the patient's spokenand ECoG responses as "silence", "speech" and "noise" to train our model. During our analysis the noisy epochs have been removed.

III. EXPERIMENTAL SETUP

A. Feature extraction

Recorded data from each ECoG electrode were re-referencedby subtracting the common average (CAR) of electrodes in the samearray, as defined by equation (1),

$$x[n]_{ch}^{CAR} = x[n]_{ch} - \frac{1}{N}\sum_{l=1}^{N} x[n]_l \qquad (1)$$

where $x[n]_{ch}$ and $x[n]_{ch}^{CAR}$ are the ECoG and CARreferenced ECoG amplitudes on the*ch*-th channel out of a total of*N* recorded channels.

We aimed to characterize the temporal and frequency information in ECoG channels, so we chose hybrid featuresfor investigation, which included channel-specific power spectra and autoregressive model coefficients. Each ECoG channel was segmented with a sliding Hamming window and two sets of features were extracted for each window. We used a window length of 256 samples and a step size of 128 samples.

First, the power spectra of these frames were calculated by applying a fast Fourier transform (FFT). Since the high gamma oscillations were previously reported to be highly correlated with speech-related cortical activation [9], we computed high gamma frequencies between 70 Hz and 110 Hz in 2Hz bins. Power estimates in this frequency range were log-transformed to approximate normal distributions. We then averaged the power spectra in the above frequency range to obtain the final spectral features. We separately trained an autoregressive model of order 5 using the Yule-Walker method [19] for each channel and time window. The model order was found as a tradeoff between computational cost and model prediction accuracy. The AR model coefficients were used as temporal features. Consequently, a total of 534 features were used for our VAD analysis. In this paper, the k-th AR coefficient of n-th channel is denoted as channel(n) – AR(k) and the average power as channel(n) – PSD.

B. Feature Selection and Classification

Feature selection methods are typically applied to select a small set of effective features in order to improve generalization ability and classification performance. Consequently, feature selection was performed to reduce the dimensionality of the feature vector from $\mathbb{R}^{534}$ to a lower dimensional space $\mathbb{R}^{N}$, with N<534. The RelieF algorithm was used [20] for the selection of the N most important features. The RelieF algorithm evaluates the worth of a feature by iteratively sampling an instance and considering the value of the given feature for the nearest instance of the same and different class (here speech or silence). Using the feature ranking results, we evaluated the N-best features for N= {10, 50, 100, 150, 200, 300, 534} for the VAD task, i.e. the binary classification problem between speech/silence.

For classification we tested seven classifiers used in literature [21] to examine the robustness of our method:support vector machines (SVM), multilayer perceptron (MP), K-nearest neighbors (KNN), J48 tree, decision stump tree (DS), regression tree (RT) and logistic regression (LR). The evaluation of results was estimated using 10-fold cross validation, in order to avoid overlapping between training and test subsets.

For the SVM kernel we used the radial basis function (RBF), with parameters C=20.0 and γ=0.5, which were found as optimal values after a grid search at C= {1.0, 10.0, 20.0, 30.0} and γ= {0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 5.0}.Additionally, for the KNN classifier K=20 was found as the optimal value after searching at K= {1, 10, 15, 20, 25, 30, 40, 50}.

IV. Experimental results

The ECoG features described above were evaluated for their appropriateness for the VAD task. The RelieF algorithm feature ranking results for the 10 most discriminative ECoG features are shown in Table 1. The most discriminative ECoG features are both spectral and temporal features. The most informative feature is the average high gamma power spectrum of channel 20, with a ranking score of 0.074.

Table 1 Ranking of the 10-best ECoG features for the Voice Activity Detection task as evaluated by the RelieF algorithm

Ranking	***ECoG features***	***Ranking Score***
1	channel(20)-PSD	0.074
2	channel(33)-AR(2)	0.063
3	channel(20)-AR(5)	0.061
4	channel(33)-PSD	0.058
5	channel(20)-AR(2)	0.049
6	channel(21)-PSD	0.047
7	channel(32)-PSD	0.046
8	channel(20)-AR(1)	0.041
9	channel(89)-AR(5)	0.035
10	channel(41)-AR(5)	0.034

We next evaluated subsets of ECoG features, in the order of their RelieF ranking outcome, to examine the optimal parametric subset. The VAD accuracy, in percentage, for the N-best ECoG features for the tested classifiers is shown in Table 2. As can be seen from Table 2, the optimal performance (average 96%) was achieved when using the 100 to 150 most discriminative ECoG features for SVM and LR classifiers. The use of fewer than 100 or more than 150 of the best ECoG features did not improve the speech/silence discrimination accuracy. This was due to the fact that the discriminative information of the best 100 ECoG features was not complementary to the information carried by the rest of the features.

Table 2 Voice Activity Detection accuracy (%) using seven classifiers for a varying number of ECoG features

Nf	SVM	KNN	LR	MP	DS	J48	RT
10	94.36	93.19	93.79	93.29	92.02	91.13	92.99
50	90.65	93.39	95.53	95.24	92.18	90.23	92.95
100	96.03	92.74	96.09	95.55	92.04	89.88	92.95
150	95.61	92.08	95.53	95.45	92.04	89.16	92.89
200	95.43	90.91	95.77	95.63	92.04	89.04	92.97
300	94.50	88.15	94.08	95.63	92.04	89.02	93.01
534	84.40	85.70	93.67	95.70	92.02	89.02	92.69

Nf: the number of N- best ECoG features

Since we are interested in the development of minimally invasive BMI systems, temporal and spectral features extracted from the "best" channels, as illustrated in Table 3, using the two "best" classifiers. Features extracted from the two "best" electrodes (channels 20 and 33) resulted in a 94% classification accuracy using SVM or LR classifier. Overall the feature selection and ranking showed that features extracted from channels 20 and 33 were more informative than features extracted from other channels. Additionally, the above step is crucial to investigate the channels' significance in relationship to their location on the brain. The two electrodes that were ranked highest by the RelieF algorithm, shown in Fig. 2 with enlarged circles, were located in cortical areas typically involved in speech and language processing. Channel 20 was located over the posterior superior temporal gyrus (STG), which contains auditory association cortex and is part of Wernicke's area, typically important for speech perception. Channel 33 was located over or near tongue motor cortex, which is important for articulation. Many of the other most highly ranked electrodes were located near these electrodes or in cortical areas also relevant to speech processing, including Broca's area.

V. Conclusions

While previous ECoG studies on speech decoding and reconstruction focused on phoneme or word level processing [12][14], they did not consider the problem of voice activity detection (VAD). Here, VAD from ECoG signals was studied. Temporal features were extracted from the raw ECoG signal, and spectral features were extracted from the high gamma (70 Hz – 110Hz) response. Then the machine learning RelieF algorithm was implemented to reduce the dimensionality of the feature space and several classification algorithms were tested. We achieved the highest average classification accuracy, 96%, using the 100- best ECoG features. Moreover, we found that channels 20 and 33, which were located in cortex typically important for speech production, were the most informative.

While results from the current study are encouraging, more extensive training using larger datasetsis expected to further improve the generalization ability and increase the performance of our classificationsystem. In the future, we aim to extract more complex features and investigate several machine learning techniques for feature selection and classification to detect voice activity, not only from overt but also from covert articulation. This may assist in fully automated natural speech BMIs, which will enable people to communicate silently using related brain activity.

Table 3 The 10- best ranked channels evaluated by the RelieF algorithm

Ranking	*Channel*	*Ranking Score*
1	channel(20)	0.041
2	channel(33)	0.034
3	channel(89)	0.024
4	channel(32)	0.019
5	channel(49)	0.016
6	channel(25)	0.016
7	channel(12)	0.015
8	channel(41)	0.013
9	channel(18)	0.012
10	channel(21)	0.011

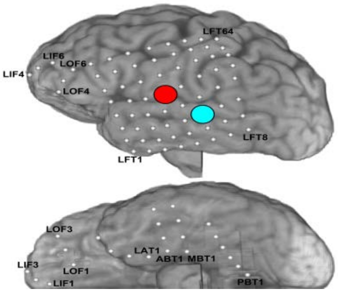

Fig. 2 The two best-ranked electrodes, channel 20 (blue) and channel 33 (red), were located in cortical areas relevant to the speech task.

Acknowledgment

This research was funded by the John S. Latsis Public Benefit Foundation, the European Social Fund, and the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework, Research Funding Program THALES, and NINDS R01-40596 (NEC)

References

1. Wolpaw J R, Birbaumer N et al. (2002) Brain-computer interfaces for communication and control.Clin Neurophysiol 113: 767–791
2. Schalk G, Kubanek J et al. (2007) Decoding two- dimensional movement trajectories using electrocorticographic signals in human. JNeural Eng 4: 264–275.
3. Benz H, Zhang H et al. (2012)Connectivity analysis as a novel approach to motor decoding for prosthesis control. IEEE Trans on Neural Systems and Rehabilitation Engineering 20:143-152.
4. Benz H L, CollardM et al. (2012)Directed Causality of the Human Electrocorticogram During Dexterous Movement. Engineering in Medicine and Biology Society (EMBC), 34th Annual International Conference of the IEEE, pp. 1872-1875.
5. StavrinouML, MoraruLet al. (2007)Evaluation of Cortical Connectivity During Real and Imagined Rhythmic Finger Tapping, Brain topography19:137-145.
6. FiferMS, AcharyaS et al. (2012) Toward Electrocorticographic Control of a Dexterous Upper Limb Prosthesis: Building Brain-Machine Interfaces. IEEE Pulse 3:38-42.
7. PeiX, Hill J et al. (2012)Silent communication: Toward using brain signals.IEEE Pulse3:43-46.
8. GuentherF H, Brumberg JS et al. (2009)A wireless brain–machine interface for real-time speech synthesis. PloS Biology 4: e8218.
9. KorzeniewskaA, FranaszczukP J et al. (2011) Dynamics of large-scale cortical interactions at high gamma frequencies during word production: Event related causality (ERC) analysis of human electrocorticography (ECoG). NeuroImage 56: 2218–2237.
10. CroneN E, BoatmanD et al. (2001)Induced electrocorticographic gamma activity during auditory perception. Clinical Neurophysiol112: 565-582.
11. DaSalla C S, KambaraH et al. (2009) Single- trial classification of vowel speech imagery using common spatial patterns.Neural Networks 22:1334-1339.
12. Pei X, Barbour D L et al (2011)Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans.J Neural Eng 8:046028
13. SchalkG, LeuthardtE C (2012)Brain-Computer Interfaces Using Electrocorticographic (ECoG) Signals. IEEE Reviews in Biomedical Engineering 4:140 – 154.
14. PasleyBN, DavidSV, et al. (2012)Reconstructing Speech from Human Auditory Cortex. PloS Biology 10:e1001251.
15. Linderman, M D, Santhanam G et al. (2008) Signal Processing Challenges for Neural Prostheses. IEEE Signal Processing Magazine 25:18-28.
16. ChangJ, Kim N S et al. (2006) Voice Activity Detection Based on Multiple Statistical Models. IEEE Trans on Signal Processing54: 1965-1976.
17. Duncan J S, Papademetris X et al. (2004)Geometric strategies for neuroanatomic analysis from MRI.Neuroimage23, Suppl. 1: S34-45.
18. Boersma P, Weeninck D(2001) Praat, a system for doing phonetics by computer.Glot International 5:341-345.
19. Monson H (1996)Statistical Digital Signal Processing and Modeling. John Wiley & Sons.
20. KononenkoI (1994)Estimating Attributes: Analysis and Extensions of RELIEF. In Proc. of the European Conference on Machine Learning, pp. 171-182.
21. Bashashati A, Fatourechi M et al. (2007) A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. J Neural Eng 4:32-57.

Author:	Vasileios G. Kanas
Institute:	University of Patras
Street:	Korinthou 1
City:	Patras
Country:	Greece
Email:	vaskanas@upatras.gr