Optimal Selection of Electrocorticographic Sensors for Voice Activity Detection

Vasileios G. Kanas[1], Iosif Mporas[1,2] Heather L. Benz[3], Kyriakos N. Sgarbas[1], Nathan E. Crone[4], and Anastasios Bezerianos[5,6]

[1] Department of Electrical and Computer Engineering, University of Patras, Patras, Greece
[2] Computer and Informatics Engineering Department, TEI of Western Greece, Patras, Greece
[3] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205 USA.
[4] Department of Neurology, Johns Hopkins University, Baltimore, MD 21205 USA.
[5] Singapore Institute for Neurotechnology, National University of Singapore, Singapore.
[6] Department of Medical Physics, School of Medicine, University of Patras, Patras, Greece.

Email: vaskanas@upatras.gr, imporas@upatras.gr, benz@jhu.edu, sgarbas@upatras.gr, ncrone@jhmi.edu, tassos.bezerianos@nus.edu.sg

***Abstract*— An effective speech brain machine interface requires selecting the best cortical recording sites and signal features for decoding speech production, but also minimal clinical risk for the patient. Motivated by this need to reduce patient risk, the purpose of this study is to detect voice activity (speech onset and offset) automatically from spatial-spectral features of electrocorticographic signals using the optimal number of sensors (minimal invasiveness). ECoG signals were recorded while a subject performed two different syllable repetition tasks. We found that the optimal frequency resolution for detecting voice activity is 8 Hz using 31 sensors out of 55, achieving 98.2% accuracy by employing support vector machines (SVM) as a classifier, and that acceptable accuracy of 96.7% was achieved using 15 sensors, which would permit a less invasive surgery for the placement of electrodes. The proposed voice activity detector may be utilized as a part of an ECoG-based automated natural speech BMI system.**

Keywords*— Brain machine interface (BMI), voice activity detection (VAD), electrocorticography (ECoG), minimal clinical risk

I. INTRODUCTION

Brain machine interface (BMI) systems use recorded neural activity for motor control [1]-[4] or communication [5]-[8]. Silent communication BMIs could be used as rehabilitation tools for locked-in patients [9] by enabling the intuitive articulation of synthesized speech. While most research into such prosthetic systems is based on non-invasive electroencephalographic (EEG) signals [10]-[15], electrocorticography (ECoG)-based BMI systems, in which electrodes are implanted on the surface of the brain, benefit from high spatial specificity and signal-to-noise ratio. Recent studies using ECoG activity have achieved discrimination between vowels and consonants during overt and covert speech [16], [17] recognition of a small set of spoken words [18], and even classification of whole sentences [19].

A clinically-viable permanently-implantable speech prosthetic system should be real-time and minimize power consumption [20] to be effective for realistic everyday use by patients. The detection of individual's speech activity (i.e., the time interval in which an individual speaks) automatically from neural activity is essential for meeting power constraints by limiting unnecessary signal processing [21]. Additionally, channel selection is an important problem in BMI. By using data from initial calibration sessions to train our models, we can remove the redundant or less informative ECoG channels. There are several reasons for reducing the number of channels used. First, the channel implantation involves clinical risk for the patient. A minimally invasive BMI system is therefore desirable. Moreover, when training a BMI system, the number of available trials is usually limited compared to the number of channels, leading to a risk of overfitting the classifier.

Current experimental protocols do not detect speech epochs autonomously, meaning that human intervention is needed to differentiate between speech and non-speech intervals. Building upon our previous study [21], here we extract spectral characteristics from the entire informative frequency bandwidth and select the optimal number of ECoG sensors to drive a less

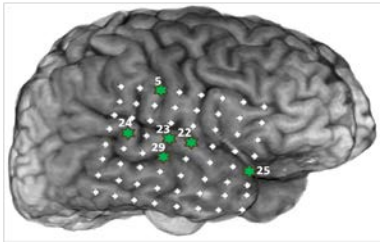

Fig.1 A reconstruction of electrode locations in the subject. The electrode grid covered areas of temporal, parietal, and frontal lobe in the right hemisphere that are analogues of language-processing areas in the left hemisphere. The five best channels are indicated with green stars.

This research has been co-financed by the European Union (European Social Fund, ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF); Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

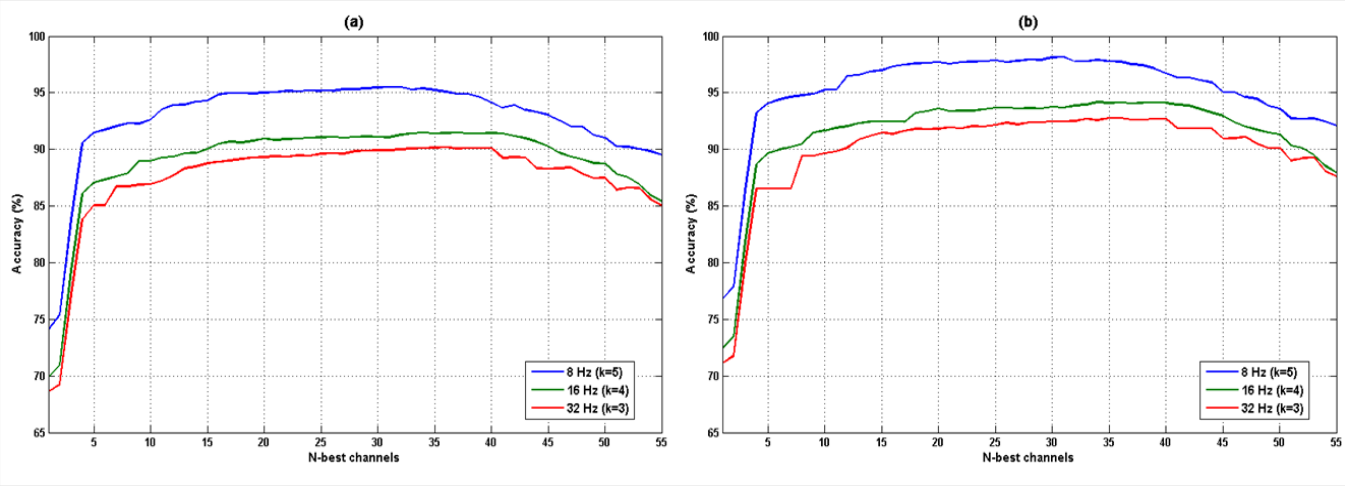

Fig.2 Classification accuracy, in percentage, for three frequency resolutions (k=1,2,3) along with the N-best (as evaluated by the RelieF algorithm) ECoG channels for (a) without post-processing and (b) with post-processing using SVM as classifier. The y axis is the classification accuracy, and x axis is the number of the selected channels.

invasive and automatic voice activity detector (VAD), which could be used as a preliminary processing stage for speech prosthetic systems. We evaluate the most informative cortical areas by exploring voice activity jointly in the spatial and spectral domains to reveal how the neural representation of voice activity is organized within different channels and frequency bands.

II. Materials and Methods

A. *Experiment and data acquisition*

One male patient was enrolled in this study following implantation of an ECoG grid for the treatment of epilepsy. The patient consented to participate in the experiment, which was approved by the Johns Hopkins Medicine Institutional Review Board. The ECoG grid was composed of 64 electrodes with 1 cm center-to-center spacing (Ad-Tech, Racine, Wisconsin, USA). ECoG electrodes were visualized on the brain surface using co-registration of pre-implantation volumetric MRI with post-implantation volumetric CT in Bioimage [22].

The array on the brain surface is shown in Fig. 1. Two syllable tasks, cued with visual and auditory stimuli, were performed by the patient during ECoG recording. 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"). Each of the 12 syllables was presented 10 times, for a total of 120 trials in each task. In the auditory version of the task, each trial was 4,000 ms long, while in the visual version each trial was 3,072 ms long. In both syllable tasks, between trials a fixation cross was displayed on the screen for 1,024 ms.

A NeuroPort System amplifier (Blackrock Microsystems, Salt Lake City, Utah) was used to record ECoG data at a sampling rate of 10 kHz. Data was subsequently low pass filtered with a cutoff frequency of 500 Hz. The patient's spoken responses were recorded with a Zoom H2 recorder (Samson Technologies, Hauppauge, New York), also at 10 kHz, and were time-aligned with ECoG recordings through an analog input on the NeuroPort amplifier. Each dataset was visually inspected and all channels that did not contain clean ECoG signals were excluded, which left $M = 55$ channels for our analyses. Recorded data from each ECoG sensor were re-referenced by subtracting the common average (CAR) [23] of signals on all electrode. The ECoG signals of each channel were also normalized by subtracting the average value and dividing by the standard deviation within each channel. The open source Praat software [24] was 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 were excluded from the evaluation.

B. *ECoG feature extraction*

The neural correlates of voice activity captured by the subdural ECoG sensors might appear in different frequency bands disparately distributed over the cortical area. We characterize the spectral characteristics separately for each ECoG sensor, aiming to examine the frequency bands in parallel with the cortical areas that provide the highest performance for the VAD task. Here, we are interested in spectral characteristics extracted from the entire frequency bandwidth and not only from predefined frequency regions of interest in the literature (i.e. delta, theta, alpha, beta, gamma, high gamma). Each ECoG channel is segmented with a sliding Hamming window with length of 256 samples and a step size of 128 samples. For each of the overlapping windows, the power spectra are calculated with a fast Fourier transform (FFT). Power estimates 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 $K = 2^k$ frequency bands to obtain the final spectral features per ECoG channel, resulting in different sets of feature vectors $V \in \mathbb{R}^{M \times K}, k = 0,1,...,8$ .Then a total of 55-14080 features (depending on the number of averaged frequency bands) are

TABLE I. THE 10 BEST RANKED ECOG CHANNELS AS EVALUATED BY THE RELIEF ALGORITHM FOR THREE FREQUENCY RESOLUTIONS

	Frequency Resolution					
	8 Hz (k=5)		***16 Hz (k=4)***		***32 Hz (k=3)***	
Ranking	**ECoG channel**	**Ranking Score**	**ECoG channel**	**Ranking Score**	**ECoG channel**	**Ranking Score**
1	24	0.00658	24	0.00698	24	0.00725
2	29	0.00424	29	0.00433	29	0.00432
3	23	0.00404	23	0.00423	23	0.00422
4	22	0.00352	22	0.00375	22	0.00386
5	5	0.00325	5	0.00355	5	0.00382
6	25	0.00303	25	0.00318	25	0.00317
7	27	0.00287	27	0.00295	20	0.00309
8	32	0.00276	32	0.00292	31	0.00302
9	46	0.00268	20	0.00283	46	0.00297
10	17	0.00259	46	0.00281	32	0.00291

used in our analysis. In the remainder of the paper, we use the term "frequency resolution" to denote the spectral components, with 1 Hz distance between them, contained in each of the $K = 2^k$ frequency bands. Here, we present the results of the best performing frequency resolutions, 8 Hz, 16 Hz, and 32 Hz.

C. Optimal channel selection and classification

The discriminative ability of each feature for the VAD task is evaluated using the RelieF algorithm [25] applied separately to each feature vector set (i.e., for $k = 0,1,...,8$). The RelieF algorithm evaluates the worth of a feature and generates a ranking score by repeatedly sampling an instance of the feature and finding the value of the given feature for discriminating the nearest instance of the class in which it was found from the alternative class (here speech or silence). Then, for each feature vector (i.e., for $k = 0,1,...,8$), we average the ranking scores across each channel to determine the discriminative ability of each ECoG sensor. Finally, the features derived from the optimal subset of ECoG sensors are used as inputs to the classification model.

For classification we tested three classifiers used in the literature [26] to examine the robustness of our method: support vector machines (SVMs), K-nearest neighbors (KNN), and Naive Bayes. The evaluation of results was estimated using 10-fold cross validation. For the SVM kernel we used the radial basis function (RBF), with parameters C=10.0 and γ=0.01, which were found to be optimal values after a grid search at C= {1.0, 5.0, 10.0, 20.0, 30.0} and γ= {0.001, 0.01, 0.1, 0.5, 1.0, 2.0, 5.0}.

As a final stage, a post-processing step over the sequence of frame-based decisions is applied, eliminating sporadic erroneous labeling of the current ECoG frame. At this stage, we smooth each decision with respect to its closest neighbors. In particular, if the $l \geq 0$ preceding and the $l \geq 0$ successive ECoG frames are classified as a given label (speech or non-speech), then the current frame is also relabeled as this label. After testing, we selected $l = 2$.

III. EXPERIMENTAL RESULTS

In this study, we are interested in the development of a less invasive voice activity detection system. The ten best ECoG channels, as evaluated by the RelieF algorithm, for different frequency resolutions are shown in Table I. Channels 24, 29, 23, 22, 5 and 25 carry the most information common to the three frequency resolutions. These channels are located in the right hemisphere analogues of left hemisphere cortex relevant to speech, including superior temporal gyrus, superior temporal sulcus, speech sensorimotor areas, and other temporal lobe areas.

Additionally, we investigate how spectral features $V \in \mathbb{R}^{M \times K}, k = 0,1,...,8$, derived from additional best ECoG sensors, influence the accuracy of the VAD system. SVMs are found to outperform the other classification algorithms and achieved classification accuracy of 95.50% and 98.2% (before and after post-processing step, respectively), while the second best classifier, KNN, achieves 87.90% and 89%. Fig. 2 shows the VAD accuracy, in percentage correct classifications, for the N-best ECoG channels, with the post-processing step improving the VAD performance. The best VAD accuracy (98.2%) is achieved using 8 Hz frequency resolution and 31 ECoG channels out of the total 55, while using only the 15-best channels results in a an approximately 1% drop of the VAD accuracy. For this reason, we find that using 15 channels provides an optimal equilibrium between system accuracy and patients' clinical risk.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we investigate the effect of ECoG sensor selection on the classification accuracy of the voice activity detection task, aiming to provide information about the

optimal number of channels for a less invasive ECoG-based BMI system. Spatial-spectral features are extracted from the entire frequency bandwidth, to characterize the ECoG signals in different frequency resolutions. Three classification algorithms were evaluated, among which the support vector machine algorithm was found to achieve the highest accuracy. Our results show that 31 channels and 8 Hz frequency resolution are optimal for detecting human articulation with high accuracy (98.2%). However, 4 channels are sufficient to provide a classification performance higher than 90%. Channels 24, 29, 23, 22, 5 and 25, which were located in cortex typically relevant to speech production and perception, hold the most information about speech production across the three different frequency resolutions. The distributed locations of the best ECoG channels suggest that language processing involves large-scale cortical networks. Such networks are engaged in phonological analysis, speech articulation and other processes [27]. Although, this study shows preliminary results and further research is needed to extend our approach to more subjects, the high accuracy and low number of channels needed after channel selection reveal that BMI is a viable tool for rehabilitation with minimal clinical risk. Although The results support developing a less invasive ECoG-based speech BMI in a hierarchical fashion, with the voice detector segmenting data during model training and online operation, and a decoder processing only ECoG data related to speech epochs using the most informative channels.

REFERENCES

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

[2] 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, Sem. 2007, pp. 264-275.

[3] 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 Prostheswas: Building Brain-Machine Interfaces," IEEE Pulse, vol. 3, Jan. 2012, pp. 38-42.

[4] J. Kubanek, K.J. Miller, J.G. Ojemann, J.R. Wolpaw, and G. Schalk "Decoding flexion of individual fingers using electrocorticographic signals in humans," Journal of neural engineering, vol. 6, Dec. 2009, 066001.

[5] 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 syntheswas," PloS Biology, vol. 4, Dec. 2009, e8218.

[6] 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," Journal of Neural Engineering, vol. 8, Aug. 2011, 046028.

[7] 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," Journal of Neural Engineering, vol. 7, Oct. 2010, 056007.

[8] 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, Jan. 2012, e1001251.

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

[10] 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, Mar.1999, pp.297–298.

[11] 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, Jun. 2003, pp. 120–123.

[12] 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, Jun. 2000, pp. 174–179.

[13] 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, Jan. 2008, pp. 15–21.

[14] 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, Jun. 2006, pp. 229–233.

[15] 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, Jun. 2004, pp. 979–984.

[16] C.S. DaSalla, H. Kambara et al., "Single- trial classification of vowel speech imagery using common spatial patterns", Neural Networks, vol. 22, 2009, pp. 1334-1339.

[17] X. Pei, D.L. Barbour et al, "Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans", J Neural Eng, vol. 8, 2011, 046028

[18] S. Kellwas, K. Miller, K. Thomson, R. Brown, P. House and B. Greger. "Decoding spoken words using local field potentials recorded from the cortical surface," Journal of Neural Engineering, vol. 7, Oct. 2010, 056007.

[19] D. Zhang, E. Gong, W. Wu, J. Lin, W. Zhou, and B. Hong, "Spoken sentences decoding based on intracranial high gamma response using dynamic time warping," Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE, 2012.

[20] M.D. Linderman, G. Santhanam, C.T. Kemere, V. Gilja, S. O'Drwascoll, 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, 2008, pp.18-28.

[21] V.G. Kanas, I. Mporas, H.L. Benz, K. Sgarbas, A. Bezerianos, N.E. Crone, "Joint Spatial-Spectral Feature Space Clustering for Speech Activity Detection from ECoG signals.", vol. 61, no. 4, April 2014, pp.1241-1250.

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

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

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

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

[26] Bashashati A, Fatourechi M et al., "A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals". J Neural Eng , vol. 4, 2007, pp. 32-57.

[27] 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, Jun. 2011, pp. 2218–2237.