Real-Time Voice Activity Detection for ECoG-Based Speech Brain Machine Interfaces

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

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

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

***Abstract*— In this article, we investigated the performance of a real-time voice activity detection module exploiting different time-frequency methods for extracting signal features in a subject with implanted electrocorticographic (ECoG) electrodes. We used ECoG signals recorded while the subject performed a syllable repetition task. The voice activity detection module used, as input, ECoG data streams, on which it performed feature extraction and classification. With this approach we were able to detect voice activity (speech onset and offset) from ECoG signals with high accuracy. The results demonstrate that different time-frequency representations carried complementary information about voice activity, with the S-transform achieving 92% accuracy using the 86 best features and support vector machines as the classifier. The proposed real-time voice activity detector may be used as a part of an automated natural speech BMI system for rehabilitating individuals with communication deficits.**

Keywords*— *Brain–machine interfaces (BMIs), electrocorticography (ECoG), time-frequency analysis, voice activity detection

I. Introduction

Brain machine interfaces (BMIs) were expected to be the next breakthrough in the field of rehabilitation for severely handicapped individuals. Several studies have made advances toward the development of effective motor [1]-[4] and speech prostheses [5]-[8] based on biological signals. These speech prostheses aim to completely replace the vocal mechanism of a locked-in individual [9] and enable the articulation of words directly or indirectly from neural activity. Recent studies have demonstrated the feasibility of using information in brain signals to discriminate between vowels and consonants during overt and covert speech [10], [11], to recognize a small set of spoken words [12], and even to classify whole sentences [13].

As with automatic speech recognition systems, such neuroprosthetic devices need to be real-time and minimize power consumption [14] to be effective for realistic everyday use by patients. The experimental protocols proposed in current literature require human intervention or a trial-based system to differentiate between speech modes (speech versus silence), resulting in non-autonomous speech prosthetic systems. Meeting power constraints by limiting unnecessary signal processing was a crucial concern for clinically-viable permanently-implantable speech prosthetic systems. Therefore, the detection of individual's speech activity (i.e., the time interval in which an individual speaks) was essential for their operation [15].

Motivated by the need for automatic speech activity detection, we proposed a module for automatic real-time voice activity detection (VAD) from ECoG signals during syllable articulation. In addition, considering every time-frequency (TFR) approach should be viewed as a measurement device [16], meaning that the outcome was related not only to the inherent ECoG signal characteristics but also to the specific properties of the implemented transforms, we assume that the features extracted using different TFR techniques carry complementary information related to the non-stationary properties of the ECoG signals. We therefore examine different parametric and nonparametric TFR techniques in a machine learning scheme that may be used to recognize speech and non-speech intervals online and reliably using ECoG signals.

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.

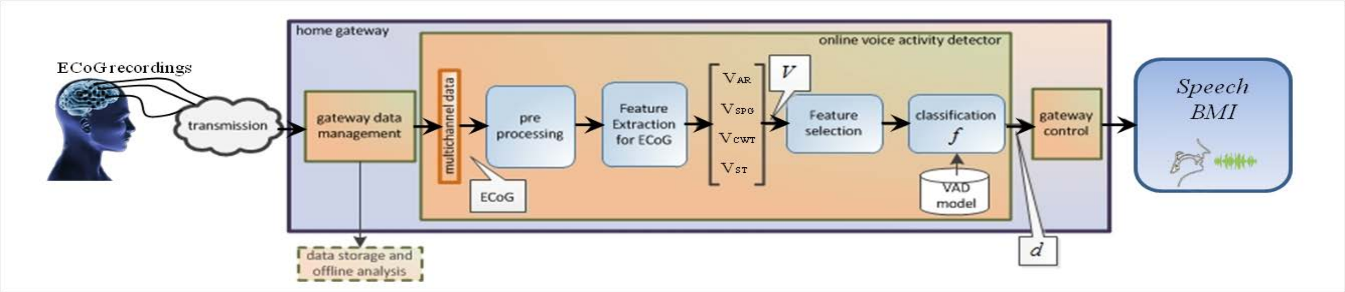

Fig.1 Schematic of the real-time VAD module. Human ECoG recordings from human subjects were input to the VAD module, which outputs gateway control of voice activity for a speech BMI. The module pre-processes multichannel ECoG data to remove noise, extracts features relevant to VAD, selects highly predictive features, and classifies speech and non-speech intervals.

II. Real-Time Voice Activity Detection Module

The presented module for voice activity detection aimed to automatically process brain signals, enabling ECoG-based speech BMIs to be used in unconstrained environments in everyday life. In general, the purpose of such an online tool is to monitor and identify speech activity using a user's brain activity. A block diagram of the VAD architecture is illustrated in Fig. 1.The recorded multichannel ECoG signals were transmitted to a local gateway for online processing. During online VAD, brain signals were initially preprocessed. Filtering was followed by frame blocking of the incoming data streams, neural data and speech data, to epochs of constant length with constant time-shift. After preprocessing, time-frequency ECoG features were extracted. The extracted feature vector was next used as the input to the feature selection block. At this stage the discriminative ability of each TFR feature of each of the ECoG channels was ranked from the most discriminative to the least discriminative. Feature evaluation was performed to select the subset of features that most contribute to the accurate detection of speech activity while rejecting the features that will reduce the overall performance, either because they increase noise or because they do not contribute enough new information [17].

During the training phase of the voice activity detector a data set of feature vectors with known class labels was used to train a binary classification model (speech vs. non-speech). At the test phase the existing VAD model was used to choose for each epoch, using the feature vector, the corresponding speech class.

III. Materials And Methods

A. *Experiment and Data Acquisition*

One male patient diagnosed with intractable epilepsy participated in this study. 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 are shown in Fig. 2. Localization of the ECoG electrodes after surgery was performed using Bioimage by co-registration of pre-implantation volumetric MRI with post-implantation volumetric CT [18].

Two syllable tasks (e.g. visual and auditory stimulus) were performed by the patient during ECoG recording. Syllable stimuli were presented and 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.

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. Channels that did not contain clean ECoG signals were excluded from our analysis.

B. *Preprocessing and ECoG feature extraction*

Recorded data from each ECoG electrode were re-referenced by subtracting the common average (CAR) of electrodes in the same array, 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 $x[n]_{ch}^{CAR}$ were 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 were also normalized by subtracting the average value and dividing by the standard deviation. The open source Praat software [19] 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.

Due to the non-stationary nature of the ECoG signals, spectral features do not provide any time domain information. In this

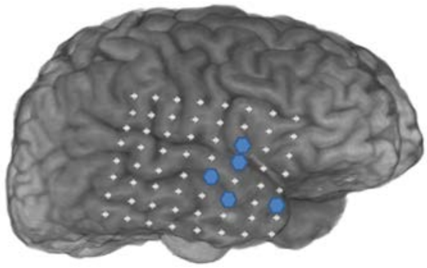

Fig.2 Electrode locations in the subject. The markers show the five best ECoG electrodes for real-time VAD module. These electrodes were located over the right hemisphere homologue of Broca's area, superior temporal gyrus, and other temporal cortices.

study, mixed time–frequency representations were used to map a one-dimensional signal into a 2D plane of time and frequency in order to analyze the channel-specific time-varying spectral content of the brain signals. More specifically, we applied $Q=4$ different time-frequency techniques to calculate the power spectrum in the 2D plane. In our analysis, the TFR-based ECoG features $\{V_{SPG},V_{AR},V_{CWT},V_{ST}\}\subseteq V\in\mathbb{R}^{N\times Q\times L}$ were averaged across time (within each sliding window) and frequency, separately for each dimension. In the frequency domain, we focused on $L=6$ ECoG frequency bands, i.e., delta (1-4 Hz), theta (5-7 Hz), alpha (8-12 Hz), beta (18-26 Hz), gamma (30-70 Hz) and high-gamma (80-200 Hz). Here, we used the range 80-200 Hz to describe the high-gamma activity, similar to Canolty et al. [22]. In the time domain, we frame blocked each ECoG channel, using a Hamming window with length of 256 samples and a step size of 128 samples. The feature extraction process resulted in the $V\in\mathbb{R}^{N\times Q\times L}$ feature vector.

Here, we used the spectrogram (SPG), autoregressive model (AR), continuous wavelet transform (CWT) and S-transform (ST) as TFR representations. After testing, for the CWT we used a Morlet mother wavelet, and for AR modeling we used a model order of 20.

C. Feature selection and classification

The evaluation of the ECoG parameters was jointly based on spatial (i.e., the selected electrodes) and time-frequency characteristics and performed using the RelieF algorithm [20]. 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 and the alternative class (here speech or silence). We evaluated the performance of our VAD system by examining each TFR technique separately; meaning that one of the feature vectors $\{V_{SPG},V_{AR},V_{CWT},V_{ST}\}\subseteq V\in\mathbb{R}^{N\times Q\times L}$ was forwarded to the feature selection stage.

For classification we tested three classifiers used in literature [21] to examine the robustness of our method: support vector machines (SVM), K-nearest neighbors (KNN), and Naive Bayes. The evaluation of results was estimated using 70% of the data for training and the remaining data to test our classification model. For the SVM kernel we used the radial basis function (RBF), with parameters C=10.0 and γ=0.5, which were found as 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}.Additionally, for the KNN classifier K=30 was found as the optimal parameter value after searching at K= {1, 10, 15, 20, 25, 30, 40, 50}.

TABLE I. HIGHEST ACHIEVED VAD ACCURACY (%) FOR DIFFERENT CLASSIFIERS USING THE N-BEST EXTRACTED TFR ECOG FEATURES

	SVM		***Naïve Bayes***		***KNN***	
	Accuracy (%)	***N-best features***	***Accuracy (%)***	***N-best features***	***Accuracy (%)***	***N-best features***
AR	78.32	147	70.71	24	71.95	36
CWT	68.21	111	52.44	8	61.89	117
SPG	75.77	120	64.69	20	69.31	233
ST	91.53	86	87.12	6	88.22	9

IV. EXPERIMENTAL RESULTS

The online VAD module described in Section II was evaluated following the experimental setup described in Section III. Online analysis requires a module that performs in real-time and with minimal computational effort. For this purpose we examined the performance of the VAD architecture, in terms of accuracy, for different numbers of features, chosen using RelieF-based feature ranking. The VAD accuracy, in percentage, for the N-best ECoG features for the tested classifiers is shown in Fig. 3. The best performance was achieved using the S-transform as a TFR representation. More specifically, we achieved 87%, 92%, and 88% accuracy using the 6-best, 86-best and 9-best features for Naive Bayes, SVM and KNN, respectively. Table I shows the highest accuracy achieved with different classifiers and TFR representations using the N-best ECoG features.

Finally, in order to investigate the channels' significance in relationship to their location on the brain, we averaged the feature ranking scores for each channel. The five electrodes that were ranked highest by the RelieF algorithm, shown in Fig. 2 with enlarged circles on the right hemisphere, were located in cortical areas typically involved in speech and language processing in the left hemisphere. Channel 19 was located over the right hemisphere area homologous to Broca's area, which was active in speech production. Channel 27 was located over the right superior temporal gyrus (STG), which contains auditory association cortex, typically involved in speech perception. Channels 36, 41, and 43 were located over temporal cortex, which was involved in speech and sound perception.

V. CONCLUSIONS

In this paper, we examined the performance of an online voice activity detection module as part of a framework for an automated natural speech BMI, which may enable people to

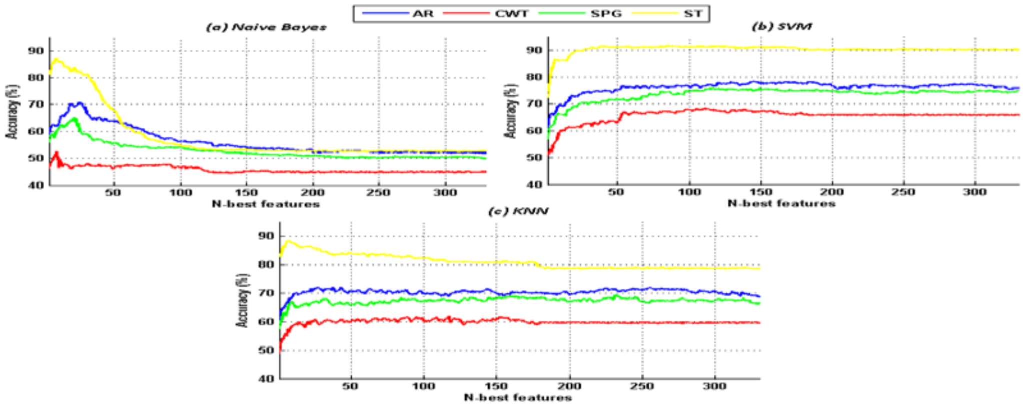

Fig. 3 VAD accuracy, in percentage, for different number of features (N-best), TFR techniques and classifiers. The ST was consistently the best-performing technique. For few features the Naïve Bayes and KNN classifiers performed best, but SVM achieved the highest accuracy overall.

communicate silently using brain activity. Data streams from ECoG recordings were used as input to the VAD module. We tested different time-frequency methods for feature extraction to study how TFR approaches influence the performance of our VAD architecture. Three classification algorithms were evaluated, among which the support vector machine algorithm was found to achieve the highest accuracy for this subject. Additionally, features extracted using the S-transform carried more information (achieving the highest performance) than features extracted using the other three methods tested. Finally, we found that channels 19, 27, 36, 41 and 43, which were located in cortex typically relevant to speech production and perception, were the most informative.

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. 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.

[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] DaSalla C S, KambaraH et al. (2009) Single- trial classification of vowel speech imagery using common spatial patterns.Neural Networks 22:1334-1339.

[11] 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

[12] 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.

[13] 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.

[14] 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, pp.18-28, 2008.

[15] 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, pp.1241-1250, April 2014.

[16] Auger, F.; Flandrin, P.; Yu-Ting Lin; McLaughlin, S.; Meignen, S.; Oberlin, T.; Hau-Tieng Wu, "Time-Frequency Reassignment and Synchrosqueezing: An Overview," Signal Processing Magazine, IEEE , vol.30, no.6, pp.32,41, Nov. 2013

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

[18] 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, pp. S34-45, 2004

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

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

[21] 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, pp. 32-57, 2007.

[22] Canolty R.T., Soltani M., Dalal S.S., et al., "Spatiotemporal dynamics of word processing in the human brain", Front Neurosci, vol. 1 pp. 1185–1196, 2007