VOICE ACTIVITY DETECTION USING NEUROGRAMS

Wissam A. Jassim and Naomi Harte

Sigmedia, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland

ABSTRACT

Existing acoustic-signal-based algorithms for Voice Activity Detection (VAD) do not perform well in the presence of noise. In this study, we propose a method to improve VAD accuracy by employing another type of signal representation which is derived from the response of the human Auditory-Nerve (AN) system. The neural responses referred to as a *neurogram* are simulated using a computational model of the AN system for a range of Characteristic Frequencies (CFs). Features are extracted from neurograms using the Discrete Cosine Transform (DCT), and are then trained using a Multilayer Perceptron (MLP) classifier to predict the VAD intervals. The proposed method was evaluated using the QUT-NOISE-TIMIT corpus, and the NIST scoring algorithm for VAD was employed as an accuracy measure. The proposed neural-response-based method exhibited an overall better VAD accuracy over most of the existing methods.

Index Terms— Speech activity detection, neurogram, and auditory-nerve system

1. INTRODUCTION

VAD is the process of detecting the presence (speech) or absence (non-speech) events in speech signals. It is an important pre-processing step in many speech processing applications, such as speech recognition, speaker recognition, and speech enhancement. The accuracy of speech/non-speech detection is severely degraded when the speech signal is distorted by noise. Therefore, a reliable VAD algorithm is required as its robustness against noise can substantially improve the performance of subsequent speech processing applications.

A typical VAD technique consists of two parts. In the first part, features are extracted from speech, and fed to a classification module to detect the speech/non-speech in the second part. Improving the performance of these two elements has received remarkable research interests over the years. The VAD module of the ITU-T G.729 coding system [1] is one of the most well-know algorithms to detect voice activity in the signal. It uses different parameters such as the full band energy, the low band energy, the zero-crossing rate, and a spectral measure to distinguish between active and inactive periods. Another common approach is the VAD algorithm designed by Sohn *et al.* [2] in which a first-order Markov process modeling of speech occurrences is employed to derive the decision rule. This algorithm showed good performance in various environmental conditions where signals are distorted by different types of noise such as Vehicle, White, and Babble noise. Tan and Lindberg [3] have proposed a VAD algorithm in which a moving average is applied to the frames selected by a low-complexity Variable Frame Rate (VFR) analysis. The current frame is assigned as speech if the moving average is greater than a specific threshold value. This method outperformed other recent VAD algorithms in different conditions indicating its effectiveness in speech recognition. Recently, several studies have proposed new algorithms to improve the performance of VAD by combining different types of features [4, 5]. In the study by Segbroeck *et al.* [6], four different types of 2D representations were combined together for VAD: the spectral-shape-based features (Gammatone Frequency Cepstral Coefficients, GFCC), the spectro-temporal modulation patterns of speech (Gabor features), Harmonicity-based features, and the Long-Term Signal Variability (LTSV) measure. The total number of features in the combined set is 184 [7]. The decision rule was derived using a MLP classifier trained on the combined feature set. This method was shown to be very competitive with current state-of-the art systems on the DARPA RATS corpora, even with low feature dimensionality. In [8], a new VAD algorithm based on the property of complex subbands of speech was proposed. This method achieved superior performance over existing algorithms on the QUT-NOISE-TIMIT corpus.

Improving the performance of VAD under noisy conditions remains a challenge however. Unlike the acoustic-signal-based methods, this study proposes an approach to detect speech activity using AN-response-based features. This idea was motivated by the fact that the neural responses (a series of brief electrical action potentials transmitted on individual fibers of the auditory neurons) exhibit strong robustness to noise. This observation is supported by the phase locking property, i.e. the behaviour that nerve neurons tend to fire potentials at times corresponding to a peak in the sound stimuli. In this study, the neural responses correspond-

Work funded by the ADAPT Centre for Digital Content Technology, which is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

ing to a speech signal are simulated using a computational model of the auditory periphery by Zilany *et al.* [9, 10]. The model takes an input speech stimulus, and generates the time-varying spike counts for AN fibers tuned to a CF as a function of time. CF is defined as the most sensitive frequency for an AN fiber. The generated spike counts as a function of time for a range of CFs values can be represented as a 2D array (time-frequency) referred to as a neurogram. Neurograms are more informative than other 2D representations such as spectrogram, as they reflect most of the non-linear behaviours in the auditory periphery. Features from neurograms have been employed in several applications such as assessment of speech intelligibility [11, 12], speech quality [13], and identifying emotions in speech [14].

2. PROPOSED METHOD

The proposed VAD method consists of two stages: training and testing. In the training stage, the neural responses are simulated for the input speech signals using the AN model, and feature extraction is then applied. An MLP classifier is trained with features from true VAD events. In the testing stage, the trained model predicts the VAD events for an input feature set. Figure 1 shows an overview of the proposed VAD.

In this study, the proposed method was tested using speech signals taken from the QUT-NOISE-TIMIT corpus [15]. This database is specifically designed to evaluate VAD algorithms across a wide variety of common background noise scenarios (cafe, home, street, car, and reverberant noise) at different Signal to Noise Ratio (SNR) levels (15, 10, 5, 0, -5, and -10 dB). 720 speech files randomly taken from the *development* set were used in the training stage, whereas 720 speech files taken from the *enrolment* and *verification* sets were used in the testing stage. The total number of speech files is 1440 sampled at 16 kHz. Note that 480 files contain less than 25% speech, 480 files had between 25% and 75% speech, and the remaining 480 files had more than 75% speech. The VAD performance was evaluated based on the ground truth event-label files created alongside the QUT-NOISE-TIMIT corpus. The scenario of data partitioning is similar to the one employed for the evaluation of noisy speaker recognition in [16].

2.1. Neurogram

The AN model requires each input speech signal to be upsampled to 100 kHz [9, 10]. The Sound Pressure Level (SPL) of the upsampled speech was adjusted to 65 dB (preferred listening level), and the resultant signal was then fed to the AN model. The responses corresponding to 64 values of CF spaced logarithmically from 180 Hz to 8 kHz were simulated. For each CF, the spike timing was averaged with a bin size of 10 μs. The binned stream was then smoothed using a 32-samples Hamming window with 50% overlap. The resultant timing information accounted for spike synchronization to frequencies up to 6.25 kHz. Note that, the smoothed neural responses represent the Temporal Fine Structure (TFS) version of neurogram [11].

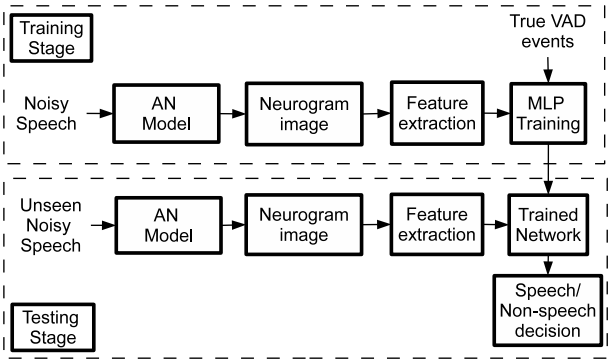

Fig. 1. Block diagram of the proposed VAD algorithm.

Three types of AN fibers are described in the literature, based on their Spontaneous Rates (SR): High Spontaneous Rates (HSR) (18-250 spikes/s), Medium Spontaneous Rates (MSR) (0.5-18 spikes/s), and Low Spontaneous Rates (LSR) (< 0.5 spikes/s) [17]. The AN model employed in this study is capable of simulating neural responses corresponding to the three types of fibers. Figure 2 shows the three versions of neurogram representations for a short segment of speech. As shown in the figure, the HSR fibers are more sensitive to signal changes than the MSR and LSR fibers. The LSR fibers have lower sensitivities for higher values of CF (> 5 kHz). However, they tend be more affected by signal changes at louder presentation levels [18].

To extract features from neurogram, the responses for each CF (one-dimensional stream) were divided into frames using a Hamming window with a time span of 20 ms and a 10 ms frame shift. An expansion of context information was then utilized by computing the DCT coefficients for a 400 ms moving time window centred around the frame of interest, and the first 5 DCT coefficients are selected. Note that this technique of feature extraction has previously been employed in [6] for the one-dimensional pitch frequency and LTSV streams. As a result, the selected coefficients across all CFs form the final 320-dimensional (64*5) feature vector for each frame. In this study, features extracted from the three types of neurogram were tested in the proposed VAD algorithm.

2.2. Training and performance evaluation

The feature vectors was normalized by mapping the mean and standard deviation across observations to 0 and 1, respectively. The mapping parameters were saved to normalize the test set. A standard MLP neural network was trained on the normalized feature vectors. The network consists of four layers: an input layer with a size equal to the feature dimension, two hidden layers with 64 nodes for each, and an out-

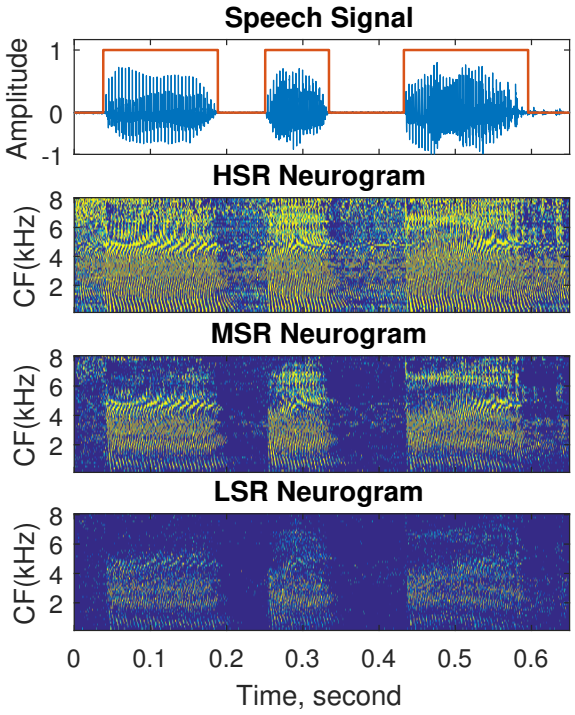

Fig. 2. Neurogram representations with 64 CFs of a short segment of speech presented at 65 dB SPL.

put layer with two nodes corresponding to speech/non-speech event. The trained network was saved to be used in the testing stage.

The NIST Open Speech-Activity-Detection (OpenSAD15) scoring software [19] was employed for performance evaluation. It computes the Detection Cost Function (DCF) error based on the time that is misclassified in a VAD algorithm as compared to true speech/non-speech events. Note that $DCF = 0.75 \times P_{Miss} + 0.25 \times P_{FA}$, where P_{Miss} and P_{FA} are the miss rate and false-alarm rate, respectively. The goal is to minimize DCF values for better VAD performance. The metric adds a collar in seconds at the beginning and end of each speech region, within which the false alarm errors are not scored. In this study, the experiments were run for collar lengths of 0.25 seconds, 0.5 seconds, 1 second, 2 seconds, and no-collars. However, only the DCF values with 0.5 seconds collar were reported here as recommended by the OpenSAD15 technical report [19].

3. EXPERIMENTAL RESULTS

The performance of the proposed method was compared to the results from four existing methods. The software by ITU-U was used to run the G.729 VAD algorithm [1]. The statistical-model-based method by Sohn *et al.* [2] was run using the Voicebox toolbox [20]. The rVAD code [21] was used to run the low complexity method by Tan and Lindberg [3]. For the feature-combination-based method by Segbroeck *et al.* [6], the Matlab code provided in [7] with its default parameter setting was employed to extract the combined feature set for the training set. An MLP network with the same structure as the one employed for the proposed method was then trained on the extracted features.

3.1. Neurogram-based VAD algorithm

Each unseen noisy signal from the enrolment and verification sets was first upsampled to 100 kHz, and its SPL was adjusted to 65 dB. The resultant signal was then fed to the AN model to simulate the neural responses with 64 CFs and three types of fibers. VAD decisions are made in 10 ms increments using the trained network based on the 320-dimensional feature vector extracted from neurogram.

Table 1 shows DCF errors values of the VAD events detected by four existing methods and the proposed neurogram-based method (HSR, MSR, and LSR neurograms) as a function of SNR for the enrolment set. In general, the proposed method outperformed three traditional algorithms (G.729, Sohn *et al.*, and Tan and Lindberg) across the SNR levels. However, the method of Segbroeck *et al.* outperformed the HSR-based method for every SNR value in this data set. Also, it outperformed the LSR-based method for the three lowest SNR levels. It can be seen that the MSR neurogram set achieved better results than that of the HSR and LSR neurogram sets. It outperformed the method of Segbroeck *et al.* in four of the six SNR levels. For the verification set, the MSR-based method achieved overall results comparable to that of the method by Segbroeck *et al.* as shown in Table 2. However, the method of Tan and Lindberg was better than any of the neurogram feature sets at -10 dB SNR, and it outperformed the HSR-based method at -5 dB SNR.

For all the systems, the VAD is less accurate on the enrolment set. This suggests this dataset is more challenging. The noise recording locations of the development set are different to that of both the enrolment and verification set. Furthermore, while the enrolment and verification sets have the same environment, the recording sessions are different. These factors may contribute to the less consistent pattern observed of noise conditions where the MSR features gave the best performance. It was difficult to run comparative simulations for the complex-subbands-based method [8] which was originally evaluated on the same database, as it uses different detection thresholds. However, the DCF errors were computed for the P_{Miss} and P_{FA} reported in that paper. The DCF values for the low (15 or 10 dB), medium (5 or 0 dB), high (-5 or -10 dB) levels of noise are 9.40, 14.58, and 31.64, respectively. Thus, it is expected that our proposed approach would outperform the complex-subbands-based method for the QUT-NOISE-TIMIT corpus.

In general, the MSR neurogram was more robust to noise than the HSR and LSR neurograms for VAD. To explore this, the 2D correlation coefficient (distance measure) was com-

Table 1. DCF (%) errors for the enrolment set. The best result is highlighted for each SNR value

SNR, dB / Method	15	10	5	0	-5	-10
G.729 [1]	19.30	21.20	22.21	23.90	28.33	30.34
Sohn *et al.* [2]	11.60	13.69	19.94	25.59	30.31	36.27
Tan & Lindberg [3]	8.18	9.53	11.20	14.13	19.18	26.33
Segbroeck *et al.* [6]	7.15	7.39	10.85	10.60	17.75	24.10
HSR	8.80	8.90	10.95	14.57	20.66	25.45
MSR	6.12	6.93	9.29	10.83	19.02	22.63
LSR	7.15	7.18	9.97	12.57	19.77	24.74

Table 2. DCF (%) errors for the verification set

SNR, dB / Method	15	10	5	0	-5	-10
G.729 [1]	14.69	20.17	21.52	24.65	29.94	32.05
Sohn *et al.* [2]	10.57	13.04	19.48	26.23	32.38	38.98
Tan & Lindberg [3]	8.80	8.91	10.98	15.79	17.33	22.28
Segbroeck *et al.* [6]	4.99	4.59	9.98	11.04	18.90	22.27
HSR	7.23	6.38	10.48	11.82	18.75	26.21
MSR	4.82	4.96	10.68	9.62	15.98	24.86
LSR	5.52	5.38	10.95	10.72	15.35	26.66

Table 3. Averaged values of correlation coefficient in % as a function of SNR

SNR, dB / Method	15	10	5	0	-5	-10
HSR	54.77	46.05	37.41	29.49	22.40	15.31
MSR	56.42	49.64	42.11	34.62	26.97	19.28
LSR	41.75	37.60	32.41	26.58	20.65	14.90

puted between the clean and corresponding noisy neurogram images for 20 speech signals randomly taken from the development set. The same parameters of CF and SPL are used for neurogram computation. Table 3 shows the averaged correlation coefficient values as a function of SNR. The results show that the distance between the clean and noisy neurogram images with MSR fibers is less than that of the other neurogram types. Thus they are more robust to noise. However, a more comprehensive analysis is required to test this behaviour for different SPL values as the neural responses may give different behaviour at different loudness levels. In this paper, the results are reported for a preferred listening level of 65 dB.

3.2. Combining Systems

The 320-dimensional neurogram feature vector was concatenated together with the 184-dimensional baseline feature set by Segbroeck *et al.* [6], and the result is a feature vector of 504 elements. The same MLP training and testing processes were repeated for the new combined form. Figures 3 and 4 show the errors rates for the enrolment and verification sets, respectively. It is clear that the performance of the existing VAD algorithm is substantially improved by adding the neural-response-based features to the baseline set. Despite the better performance of the MSR neurogram in the previous experiments, they are not always the optimal additional feature set in this combined system. It could be that the two

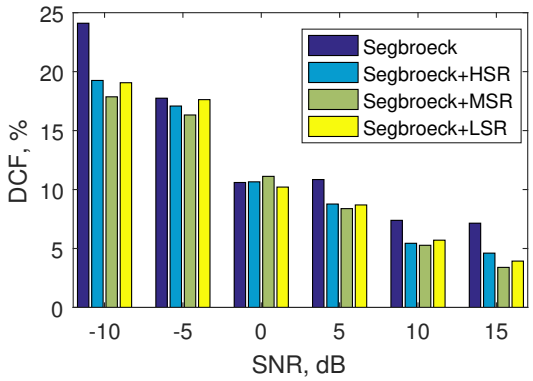

Fig. 3. DCF (%) error combining features for enrolment set

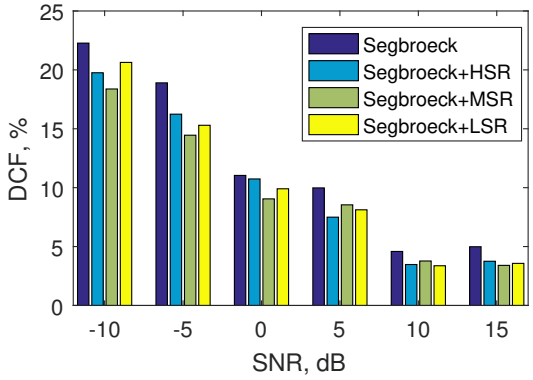

Fig. 4. DCF (%) error combining features for verification set

feature sets are correlated, and thus the overall VAD accuracy is not increased. Combining features from the three types of neurogram with the baseline features did not achieve better performance (results are not shown) or justify the high dimensionality of the combined set (1144 elements). However, it might be beneficial to employ an efficient feature selection to reduce the dimensionality of the combined features before training them with a classifier that is less sensitive to correlation of variables.

4. CONCLUSION

In this study, a neural-response-based method was proposed to detect the activity of speech. Three types of AN fibers with different SR were tested. The performance of the VAD system was evaluated under noisy conditions at different SNR levels. The proposed method achieved an overall better results over most of the existing methods. The robustness of the employed features can be attributed to the phase-locking property of the neurons in the peripheral auditory system. The experimental results also showed that the proposed features can be combined with other baseline features to improve the overall robustness of speech detection. Future work will be directed towards employing deep learning approaches to automatically learn features for speech event detection.

5. REFERENCES

[1] A. Benyassine, E. Shlomot, H. Y. Su, D. Massaloux, C. Lamblin, and J. P. Petit, "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *Comm. Mag.*, vol. 35, no. 9, pp. 64–73, Sept. 1997.

[2] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[3] Z. H. Tan and B. Lindberg, "Low-complexity variable frame rate analysis for speech recognition and voice activity detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 798–807, 2010.

[4] Samuel Thomas, Sri Harish Reddy Mallidi, Thomas Janu, Hynek Hermansky, Nima Mesgarani, Xinhui Zhou, Shihab A. Shamma, Tim Ng, Bing Zhang, Long Nguyen, and Spyridon Matsoukas, "Acoustic and data-driven features for robust speech activity detection," in *INTERSPEECH*, 2012.

[5] Masakiyo Fujimoto, Kentaro Ishizuka, and Tomohiro Nakatani, "A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme," in *2008 ICASSP*, March 2008, pp. 4441–4444.

[6] Maarten Van Segbroeck, Andreas Tsiartas, and Shrikanth Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice.," in *INTERSPEECH*, 2013, pp. 704–708.

[7] Maarten Van Segbroeck, "Voice activity detection system," `https://github.com/mvansegbroeck/vad`, 2013.

[8] S. Wisdom, G. Okopal, L. Atlas, and J. Pitton, "Voice activity detection using subband noncircularity," in *2015 ICASSP*, April 2015, pp. 4505–4509.

[9] Muhammad S. A. Zilany, Ian C. Bruce, and Laurel H. Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery," *The Journal of the Acoustical Society of America*, vol. 135, no. 1, pp. 283–286, 2014.

[10] Muhammad S. A. Zilany, Ian C. Bruce, Paul C. Nelson, and Laurel H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, 2009.

[11] Andrew Hines and Naomi Harte, "Speech intelligibility prediction using a neurogram similarity index measure," *Speech Communication*, vol. 54, no. 2, pp. 306 – 320, 2012.

[12] Michael R. Wirtzfeld, Rasha A. Ibrahim, and Ian C. Bruce, "Predictions of speech chimaera intelligibility using auditory nerve mean-rate and spike-timing neural cues," *Journal of the Association for Research in Otolaryngology*, vol. 18, no. 5, pp. 687–710, Oct 2017.

[13] Wissam A. Jassim and Muhammad S.A. Zilany, "Speech quality assessment using 2D neurogram orthogonal moments," *Speech Communication*, vol. 80, no. Supplement C, pp. 34 – 48, 2016.

[14] Wissam A. Jassim, R. Paramesran, and Naomi Harte, "Speech emotion classification using combined neurogram and INTERSPEECH 2010 paralinguistic challenge features," *IET Signal Processing*, vol. 11, no. 5, pp. 587–595, 2017.

[15] David B. Dean, Sridha Sridharan, Robert J. Vogt, and Michael W. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *INTERSPEECH 2010*, September 2010.

[16] David B. Dean, Ahilan Kanagasundaram, Houman Ghaemmaghami, Md Hafizur Rahman, and Sridha Sridharan, "The QUT-NOISE-SRE protocol for the evaluation of noisy speaker recognition," in *INTERSPEECH 2015*, Dresden, Germany, September 2015, pp. 3456–3460.

[17] M. C. Liberman, "Auditory nerve response from cats raised in a low noise chamber," *Journal of the Acoustical Society of America*, vol. 63, no. 2, pp. 442–455, 1978.

[18] Muhammad S. A. Zilany, *Modeling the Neural Representation of Speech in Normal Hearing and Hearing Impaired Listeners*, Ph.D. thesis, Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, 2007.

[19] National Institute of Standards and Technology (NIST), "NIST open speech-activity-detection evaluation," https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation, May 2016.

[20] Mike Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 2005.

[21] Zheng-Hua Tan, "rVAD: Noise-robust voice activity detection source code," `http://kom.aau.dk/~zt/online/rVAD/index.htm`, 2014.