VOICE ACTIVITY DETECTION USING A SLIDING-WINDOW, MAXIMUM MARGIN CLUSTERING APPROACH

Phillip De Leon and Salvador Sanchez

New Mexico State University
Klipsch School of Electrical and Computer Engineering
Las Cruces, New Mexico, U.S.A.
pdeleon@nmsu.edu, salx00@gmail.com

ABSTRACT

Recently, an unsupervised, data clustering algorithm based on maximum margin, i.e. support vector machine (SVM) was reported. The maximum margin clustering (MMC) algorithm was later applied to the problem of voice activity detection, however, the application did not allow for real-time detection which is important in speech processing applications. In this paper, we propose a voice activity detector (VAD) based on a sliding window, MMC algorithm which allows for real-time detection. Our system requires a separate initialization stage which imposes an initial detection delay, however, once initialized the system can operate in real-time. Using TIMIT speech under several NOISEX-92 noise backgrounds at various SNRs, we show that our average speech and non-speech hit rates are better than state-of-the-art VADs.

Index Terms— Speech analysis, classification algorithms

1. INTRODUCTION

A voice activity detector (VAD) classifies an audio segment as to whether it contains speech or not. VADs are ubiquitous in many applications such as speech codecs [in order to reduce bandwidth by coding non-speech with fewer bits (if any)], hands-free telephony (in order to reduce acoustic echo by only activating the microphone during speech), and speech recognition (in order to improve accuracy by ignoring all non-speech segments) [1]. The challenge in VAD design is accurate classification in the presence of strong noise backgrounds.

A block diagram of the basic VAD is shown in Figure 1 where $\mathbf{s}_n$ is the signal segment, $\mathbf{x}_n$ is the feature vector extracted from $\mathbf{s}_n$, and y_n is the associated binary decision or label—either $+1$ (speech) or -1 (non-speech). The feature extraction stage computes discriminating features such as frequency-band energies or segment statistics [2, 3]. The classification stage may be based on heuristics, statistics, or pattern recognition based approaches [4].

Finally, the decision smoothing stage uses prior classified segments to produce a final decision regarding the current segment, thus improving robustness against misclassifications, i.e. a speech segment is classified as not speech or vice-versa. Misclassifications often occur at the beginning or ending of the word due to low speech levels being dominated by noise. Decision smoothing is also known as "hang-over" and is typically present in all VADs [5].

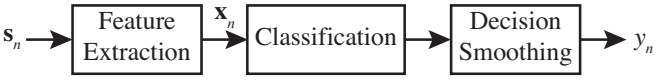

Fig. 1. Block diagram of a voice activity detector. The decision smoothing stage is also called a "hang-over" stage.

One popular VAD, proposed by Sohn, et. al., is based on spectral features which are assumed to be normally-distributed and a likelihood ratio test [3]. In addition, an initial non-speech segment is assumed in order to estimate distributional parameters and then subsequent, classified segments are used to update the distribution parameters. Sohn's VAD also employs a novel Hidden Markov Model (HMM) based hang-over scheme [3].

The ITU G.729B speech coding standard specifies a VAD which is often used in VAD performance evaluations [6]. The ITU G.729B VAD uses four features: full and low-band frame energies, line spectral frequencies (LSFs), and zero crossing rate (ZCR) [6]. Running averages of the feature vectors are calculated and the characteristic energies of the background noise. Difference measures are compared between features extracted from the current frame to the running averages. Classification is then based on a majority vote given by the averages using different features.

The ETSI AMR speech coding standard specifies two VADs: option 1 (in AMR-1) and option 2 (in AMR-2) [5]. The AMR-1 VAD separates the audio signal into different frequency bands and detects pitch and tone features present in the subbands. As in ITU G.729B, running averages of these features are computed and the decision is based on differences between current features and the averages. The AMR-2 VAD uses subband energy and power spectral density (PSD)

features. The subband energy in a current frame is compared to long-term energy estimates and a decision is made based on the SNR difference measure. Running estimates of the background noise are computed based on the deviation of the PSD in order to provide an adaptive measure of the SNR [5].

Ying et. al. proposed a VAD based on an unsupervised learning framework [2]. This VAD uses features based on the energy distribution in Mel-scale frequency bands and a sequential Gaussian Mixture Model (SGMM) [2]. The SGMM is trained using an unsupervised learning process, whereby the initial frames were clustered into two Gaussian components, with the distribution with the lowest mean modeling non-speech frames and the distribution with the higher mean modeling speech frames. The distributions were used to determine a decision threshold for the classifier [2]. The VAD classifies the frame as speech or non-speech at each subband and the subband decisions are used to determine the final decision through a voting procedure. Using performance measurements of speech hit rate (proportion of correct speech segment classifications) and non-speech hit rate (proportion of correct non-speech segment classifications), Kola et. al. [7], provides average results using the NOISEX-92 noise corpus using various state-of-the-art VADs. These results are shown in Table 1 where we see that Ying's VAD, when averaged over the various noise signals and SNRs, is better than other VADs.

Table 1. Voice activity detector performance measured by average of speech and non-speech hit rates for NOISEX-92 noise signals [7].

SNR dB	Ying's	AMR2	Sohn's	ITU	AMR1
-12	0.48	0.27	0.41	0.31	**0.52**
-3	**0.64**	0.60	0.55	0.40	0.53
0	0.68	**0.69**	0.60	0.51	0.51
3	0.71	**0.74**	0.65	0.56	0.50
6	0.75	**0.76**	0.70	0.60	0.50
12	**0.81**	0.78	0.76	0.66	0.50
18	**0.85**	0.79	0.80	0.71	0.52
Average	**0.70**	0.66	0.64	0.53	0.51

Wu et. al. proposed a VAD based on maximum margin clustering (MMC) of speech and non-speech feature vectors [8]. Through the use of MMC, this VAD obtains a support vector machine (SVM) model for speech and classifies signal segments accordingly. However, in Wu's approach, feature vectors from the pre-recorded signal are required to construct the SVM and hence does not allow for real-time detection which is important in speech processing applications.

In this paper, we propose a MMC-based VAD which utilizes a sliding window in order to provide real-time operation. The sliding window approach not only solves the SVM initialization problem but also the problem of cluster updating in a dynamic noise environment which Wu's VAD does not address. Although the proposed sliding-window, MMC VAD suffers from an initial delay (1.25s in our implementation), the initial delay does not affect incoming signal segments after 1.25s and hence can operate in real-time. As we will demonstrate, our proposed VAD has higher average accuracy than Ying's VAD which in turn has been shown to outperform the reviewed VADs including Sohn's, ITU G.729B, ETSI AMR option-1, ETSI AMR option-2 [7].

This paper is organized as follows. In Section 2, we review maximum margin clustering as proposed in [9] and describe a sliding window approach to MMC. In Section 3, we provide details for the proposed VAD based on the sliding window, MMC including hang-over stages and feature buffer management necessary for the proposed VAD. In Section 4, we describe the simulations and results. Finally, in Section 5, we conclude the paper.

2. SLIDING WINDOW, MAXIMUM MARGIN CLUSTERING

MMC seeks to assign class labels $y \in \{+1, -1\}$ to data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ such that the separation or margin between the two data clusters is maximized [9]. The MMC algorithm with label-generation, begins by randomly assigning $+1$ labels to half of the data points and -1 labels to the other half [10]. A SVM is constructed using the labeled data and the "most-violated" data point, i.e. the $+1$ labeled point which lies furtherest from the margin on the -1 side of the hyperplane or vice versa is relabeled to the other class [9]. The process of SVM construction and identification and relabeling of the most-violated data point is then repeated until convergence of the cluster membership. Alternately, a fixed number of iterations can be used or some other convergence criteria.

In the adaptation of the MMC algorithm for real-time voice activity detection, we use a sliding window of feature vectors as the data set. The sliding window consists of the current and past $M - 1$ feature vectors, allowing for the classification of the current signal segment. For each incoming feature vector, the MMC algorithm is performed using a balance constraint of zero and the maximum number of iterations of 100. The number feature vectors, M used within the sliding window affects accuracy and the initial delay.

3. VOICE ACTIVITY DETECTOR BASED ON SLIDING WINDOW, MMC

3.1. Overview

The design of the proposed VAD is illustrated in Figure 2 and is composed of six stages: feature extraction, initialization, classification, hang-over 1, hang-over 2, and feature vector buffer (FVB) management. The initialization and hang-over 1 stages are used once at system startup and then for subsequent

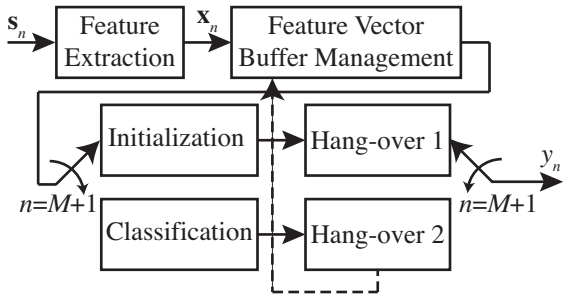

Fig. 2. Diagram of proposed sliding window, MMC VAD. For the first M feature vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_M$ the initialization and hang-over 1 stages are in effect. Once these initial feature vectors are processed, classification and hang-over 2 stages are in effect for $n > M$.

operation, the current feature vector classification and hang-over 2 stages are used in place.

3.2. Feature Extraction Stage

As in Ying's VAD, the proposed VAD uses logarithmic, Mel-weighted power spectrum (LMWPS) coefficients as feature elements [2]. The power spectrum is computed from 0-8000 Hz using a 20 ms Hamming-windowed segment with 50% overlap and a 12-channel, Mel-scale filterbank. The 12-coefficients are separated into three groups and each group is summed leading to a 3-D feature vector. We have investigated varying the number of groups but had the best results using three [11].

3.3. Initialization Stage

In order to initialize the system, we first extract M feature vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_M$ and apply the MMC algorithm to obtain associated class labels $y_1, \ldots, y_M$. Since we assume the first signal segment is non-speech, if $\mathbf{x}_1$ is labeled $+1$ we change all $+1$ labels to -1 and vice-versa; otherwise we make no label change. We have investigated performance with different values for M and have found that $M = 125$ had the best results [11]. The initialization of the clusters imposes a 1.25s startup delay before the first speech/non-speech decisions are made but thereafter, decisions are made in real-time.

3.4. Hang-over Stages

The hang-over 1 stage for the proposed VAD uses a modified version of the ETSI AMR-2 VAD's hang-over scheme [2]. In this stage, the feature vector labels $y_1, \ldots, y_M$ resulting from initialization may be changed based on hang-over parameters. The modification is to allow the possibility of all M class labels to be changed. Initial hang-over parameters are the same as in [2] with the exception of setting the hang-over counter parameter to 13 [11].

The hang-over 2 stage is also based on the ETSI AMR-2 VAD's hang-over scheme and carries over the current counters from the hang-over 1 stage beginning with $\mathbf{x}_{M+1}$. However, unlike hang-over 1, only the current label resulting from the classification stage (see below) may be changed.

3.5. Feature Vector Buffer Management Stage

In order to achieve accurate classification with MMC, the data should be roughly balanced among the two classes. We have implemented a buffer management algorithm in order to maintain this balance. The FVB management stage performs the tasks of permanently storing $\mathbf{x}_1$ (reference non-speech feature vector), balancing the M feature vectors used by MMC, and ensuring at least $M/2$ most recent feature vectors are in the buffer. The non-speech counter used in the hang-over 2 stage is initialized to one (because of $\mathbf{x}_1$) and is incremented or decremented based on decision changes produced by the hang-over 2 stage, thus allowing a the $M/2$ speech features to be present within the FVB. The FVB management algorithm is presented in Algorithm 1.

Algorithm 1 Feature vector buffer management stage

```
buffer(1) = x_1
if nonSpeechCounter < M/2 then
    buffer(i) = buffer(i + 1), i = 1, 2, ..., M − 1
    buffer(M+1) = x_n (current feature vector)
else
    buffer(i) = buffer(i+1), i = M/2, M/2+1, ..., M−1
    buffer(M+1) = x_n (current feature vector)
end if
```

3.6. Classification Stage

In order to classify the current feature vector, $\mathbf{x}_n$ we apply the MMC algorithm to the data in the FVB to obtain $M + 1$ associated class labels. Since we assume the first buffer element, $\mathbf{x}_1$ is non-speech, if $\mathbf{x}_1$ is labeled $+1$ we change all $+1$ labels to -1 and vice-versa; otherwise we make no label change.

4. SIMULATIONS AND RESULTS

4.1. Corpora and Performance Measures

In order to evaluate the proposed VAD, we used the TIMIT corpus (284 speakers chosen at random from 630) and the NOISEX-92 corpus for various noise signals (speech babble, Volvo vehicle, and white) [12, 13]. Speech was mixed with noise at various SNRs for the simulation. We used speech, non-speech, and average hit rates to evaluate performance [4, 14]. As discussed in Section 1, researchers have recently conducted a study of various state-of-the-art VADs where it

Table 2. Speech, non-speech, and average hit rates for proposed VAD and Ying's VAD under clean speech. The proposed VAD is calibrated to matched Ying's VAD results under clean speech prior to evaluating under noisy speech.

VAD	Speech Hit Rate	Non-speech Hit Rate	Average Hit Rate
Proposed	0.94	0.76	0.85
Ying's	0.94	0.75	0.85

is shown that Ying's VAD on average outperforms others [7]. Therefore, we compare our VAD only to Ying's VAD.

We implemented Ying's VAD according to [2] using recommended parameters and compared performance to that reported in [7]. We adjusted parameters on the proposed VAD to calibrate the speech/non-speech hit rates using clean TIMIT speech before evaluating under noisy speech. The results are shown in Table 2 where we see that the performance of the proposed VAD and Ying's VAD using clean speech signals have virtually the same hit rates. The results reveal that both VADs favor a higher speech hit rate resulting in a lower non-speech hit rate.

4.2. Results

For white noise, speech babble, and Volvo noise backgrounds tables 3-5, give the speech, non-speech, and average hit rates for the proposed VAD based on the sliding window, MMC algorithm and Ying's VAD. For the white noise background, we find that the proposed VAD has higher speech hit rates across all SNRs as well as average hit rates slightly lower at low SNRs but higher at higher SNRs as compared to Ying's. For the speech babble background, we find results which are similar to the white noise results. Finally, for the Volvo noise, we find that the proposed VAD has higher speech hit rates across SNRs than Ying's as well as higher average hit rate performance. We find similar results for the other NOISEX-92 noise signals (F-16, factory, and pink) [11].

Table 3. Speech, non-speech, and average hit rates for proposed VAD based on the sliding window MMC algorithm (cols. 2, 4, 6) and Ying's VAD (cols. 3, 5, 7) with white noise.

SNR (dB)	Speech Hit Rate		Non-speech Hit Rate		Average Hit Rate	
-10	0.99	0.82	0.06	0.23	0.52	**0.53**
-5	0.97	0.82	0.10	0.29	0.54	**0.55**
0	0.97	0.78	0.29	0.41	**0.63**	0.59
5	0.87	0.72	0.59	0.59	**0.73**	0.65
10	0.84	0.67	0.75	0.76	**0.80**	0.72

Table 4. Speech, non-speech, and average hit rates for proposed VAD based on the sliding window MMC algorithm (cols. 2, 4, 6) and Ying's VAD (cols. 3, 5, 7) with speech babble background.

SNR (dB)	Speech Hit Rate		Non-speech Hit Rate		Average Hit Rate	
-10	0.78	0.70	0.20	0.31	0.49	**0.51**
-5	0.79	0.71	0.23	0.36	0.50	**0.53**
0	0.80	0.71	0.31	0.42	0.56	**0.57**
5	0.82	0.73	0.45	0.51	**0.63**	0.62
10	0.84	0.75	0.61	0.63	**0.73**	0.69

Table 5. Speech, non-speech, and average hit rates for proposed VAD based on the sliding window MMC algorithm (cols. 2, 4, 6) and Ying's VAD (cols. 3, 5, 7) with Volvo noise.

SNR (dB)	Speech Hit Rate		Non-speech Hit Rate		Average Hit Rate	
-10	0.84	0.70	0.70	0.70	**0.77**	0.70
-5	0.85	0.72	0.79	0.81	**0.82**	0.76
0	0.88	0.76	0.80	0.86	**0.84**	0.81
5	0.90	0.80	0.79	0.87	**0.85**	0.84
10	0.91	0.84	0.78	0.86	**0.85**	**0.85**

4.3. Remarks

Due to the iterative nature of the MMC algorithm, the computational complexity is greater than that of other VADs. There are several ways that the complexity may be reduced with only a small reduction in performance. These include reducing the maximum number of iterations, reducing the window size, and advancing the sliding window by more than one sample. It is difficult to compare computational complexity of the proposed VAD with other VADs since these analyses are not found in the literature. However, informal benchmarking on a standard PC revealed slightly more processing time than Ying's VAD.

5. CONCLUSIONS

We have proposed a voice activity detector (VAD) based on a sliding window of feature vectors and a maximum margin clustering (MMC) algorithm. The use of the sliding window allows the proposed VAD to be used in real-time speech processing applications unlike a previously-proposed VAD which also used MMC. The proposed VAD was compared to Ying's sequential Gaussian Mixture Model (SGMM) VAD using NOISEX-92 signals at various SNRs and was shown to have higher average speech and non-speech hit rates.

6. REFERENCES

[1] J. Ramirez, J. Benitez, L. Garcia, and A. Rubio, "Statistical Voice Detection Using a Multiple Observation Likelihood Ratio Test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689–691, October 2005.

[2] D. Ying, Y. Yan, J. Dang, and F. Soong, "Voice Activity Detection Based on an Unsupervised Learning Framework," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 19, no. 8, pp. 2624–2633, November 2011.

[3] J. Sohn, N. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, January 1999.

[4] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 19, no. 3, pp. 600–613, March 2011.

[5] ETS1 EN 301 708, *Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels*, 1999.

[6] International Telecommunication Union, *Coding of Speech at 8 kbits/s Using Conjugate Structure Algebraic Code-Excited Linear-Prediction (CS-ACELP). Annex B: A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70*, 1996.

[7] J. Kola, C. Espy-Wilson, and T. Pruthi, "Voice Activity Detection," *MERIT BIEN*, pp. 1–6, 2011.

[8] J. Wu and X. L. Zhang, "Maximum Margin Clustering Based Statistical VAD with Multiple Observation Compound Feature," *IEEE Signal Process. Lett.*, vol. 18, no. 5, pp. 283–286, May 2011.

[9] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans, "Maximum Margin Clustering," *in Advances in Neural Information Processing Systems*, vol. 17, 2005.

[10] Y. F. Li, I. W. Tsang, J. T. Kwok, and Z. H. Zhou, "Tighter and Convex Maximum Margin Clustering," *AISTATS JMLR*, vol. 5, 2009.

[11] S. Sanchez, "Voice Activity Detection Using a Sliding Window, Maximum Margin Clustering Algorithm," M.S. thesis, New Mexico State University, Nov. 2012.

[12] J. S. Garofolo, *Getting Started With the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database Nat. Inst. Standards Technol. (NIST)*, Gaithersburg, MD, December 1988.

[13] A. Varga and H. Steeneken, "Assessment for Automatic Speech Recognition: NOISEX-92: a Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, July 1993.

[14] E. Nemer, R. Goubran, and S. Mahmoud, "Robust Voice Activity Detection Using Higher-Order Statistics in the LPC Residual Domain," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 217–231, March 2001.