A Robust Voice Activity Detection for Real-Time Automatic Speech Recognition *

Omid Ghahabi, Wei Zhou, Volker Fischer

EML European Media Laboratory GmbH, Berliner Straße 45, 69120 Heidelberg, Germany
omid.ghahabi@eml.org

Abstract: Voice Activity Detection (VAD), locating speech segments within an audio recording, is a main part of most speech technology applications. Non-speech segments, e.g., silence, noise, and music, usually do not carry any interesting information in speech recognition applications and they even degrade the performance of the recognition system in terms of both the accuracy and computational cost. Various VAD techniques have been developed, but not all of them are appropriate for a real-time application where the robustness, accuracy, and the processing time are the main keys. In this paper, we propose a fast and robust VAD for a real-time Automatic Speech Recognition (ASR) task. The main goal is to efficiently filter out the non-speech segments before processing the speech segments of the audio signal by the decoder. The proposed technique is a hybrid supervised/unsupervised model based on zero-order Baum-Welch statistics obtained from a Universal Background Model (UBM). We will show that not only the processing time for the whole speech recognition task is decreased by 39%, but also the Word Error Rate (WER) is reduced by about 1.9% relative.

Introduction

Voice Activity Detection (VAD) is a fundamental signal processing step in almost every speech processing application like speech coding, speech enhancement, speaker, and language recognition. The non-speech frames (e.g., silence, noise, and music) are usually not as interesting as the speech frames in these applications and are typically discarded before further processing.

Various VAD algorithms have been proposed from simple energy based ones in the time or frequency domain (e.g., [1, 2]) to more complicated deep learning based statistical models (e.g., [3, 4, 5, 6]). Energy based algorithms work usually well on clean conditions but the performance degrades rapidly in presence of environmental noise, background speech, or other acoustic events like music. Statistical models, on the other hand, can learn the statistical properties of speech and non-speech frames based on the available training data. However, these techniques are usually more complicated and may not work well on unseen environments which have not been in the training data. Statistical models can be supervised, unsupervised, or a combination of them. Supervised techniques are typically more accurate but more sensitive to unseen environments. Although there are some adaptive techniques which adapt the decision threshold or the statistical models to the new environment, they usually need the whole testing utterance for adaptation which is not applicable in real-time applications. Some other

*Part of this work has been carried out under the EU-funded H2020-MSCA-RISE-2014 project LISTEN, GA number 644283; www.listen-project.eu.

proposed techniques rely only on the evaluating audio signal to model and separate speech and non-speech segments [7, 8]. Nevertheless, these techniques usually work in two stages which need again the whole utterance for an offline process.

Depending on the application, an external VAD is usually required (e.g., in speaker and language recognition) while in some other applications like Automatic Speech Recognition (ASR) the VAD can be embedded in the acoustic model. The acoustic models in ASR are nowadays based on deep learning techniques which make it very costly to process non-speech parts of an audio recording. The impact will grow with the amount of non-speech parts, where certain accuracy degradation can also be expected. For example, when the ASR system is working in an always listening mode, continuous acoustic model evaluation all the time will be too time and energy consuming and may not work well in noisy or background speech conditions. Thus, in this paper we propose a fast, robust, and accurate enough external VAD in front of the speech recognition system to mainly save time and energy and to increase the ASR accuracy where it is possible. The proposed VAD takes advantage of a large amount of unlabeled data to train a Universal Background Model (UBM) and a few amount of labeled data to model the speech and non-speech classes with two very low dimensional vectors based on zero order Baum-Welch statistics obtained from the UBM. In the testing phase, the Baum-Welch statistics of an unknown audio segment is compared with these two speech and non-speech VAD vectors and the decision is made based on a robust threshold.

The rest of the paper is organized as follows. Section 2 describes the proposed VAD algorithm in details. Section 3 analyzes the performance of the proposed VAD in terms of accuracy, computational cost, and robustness. Section 4 summarizes the paper and discusses future work.

Proposed Voice Activity Detection

Figure 1 shows the block diagram of the proposed VAD algorithm in both training and test phases. Training phase takes advantage of a large amount of unlabeled data to train a Universal Background Model (UBM) in an unsupervised manner and a small amount of labeled data to train the proposed VAD vectors based on Baum-Welch statistics given the UBM in a supervised manner. In the test phase, the Baum-Welch statistics of an unknown audio segment is compared with the VAD vectors. The main parts of the algorithm are described in more details as follows.

UBM and Baum-Welch Statistics

UBM, in this work, is a Gaussian Mixture Model (GMM) which is a weighted sum of M Gaussian densities as given by,

$$p(x|\lambda) = \sum_{i=1}^{M} w_i g(x|\mu_i, \Sigma_i) \tag{1}$$

where x is a D-dimensional feature vector, i is the index of the i th Gaussian mixture, $g(x|\mu_i, \Sigma_i)$ are Gaussian mixtures defined as,

$$g(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2} (x-\mu_i)' \Sigma_i^{-1} (x-\mu_i) \right\}, \tag{2}$$

and w_i, μ_i and Σ_i are the weight, the mean vector, and the covariance matrix of the i th Gaussian density, respectively. The UBM parameters are estimated using the Expectation-Maximization (EM) algorithm as in [9]. UBM represents the whole acoustic space and is trained with a large amount of unlabeled data.

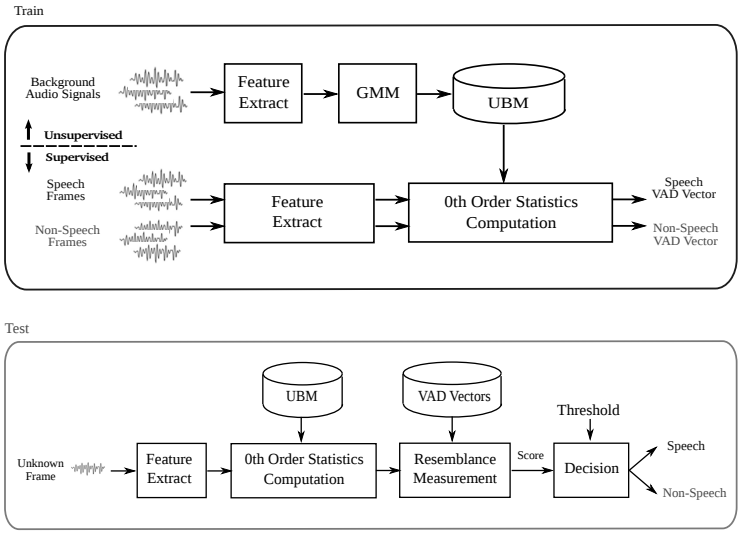

Figure 1 – Block diagram of the proposed VAD algorithm in both train and test phases.

Baum-Welch statistics, are calculated given a set of feature vectors $u = \{x_1, x_2, ..., x_T\}$ and the UBM as follows,

$$\mathscr{N}_i(u) = \sum_{t=1}^{T} Pr(i|x_t, \lambda_{ubm}) \tag{3}$$

$$\mathscr{F}_i(u) = \sum_{t=1}^{T} Pr(i|x_t, \lambda_{ubm}) x_t \tag{4}$$

where $\mathscr{N}_i(u)$ and $\mathscr{F}_i(u)$ are the zeroth and the first order statistics, respectively, and $Pr(i|x_t, \lambda_{ubm})$ is the a posteriori probability for the Gaussian mixture i calculated as follows,

$$Pr(i|x_t, \lambda_{ubm}) = \frac{w_i g(x_t|\mu_i^{ubm}, \Sigma_i^{ubm})}{\sum_{k=1}^{M} w_i g(x_t|\mu_k^{ubm}, \Sigma_k^{ubm})} \tag{5}$$

VAD Vectors and Resemblance Measurement

Given a small amount of labeled speech and non-speech feature vectors and the UBM, the zeroth order Baum-Welch statistics are computed for each class and saved as the VAD vectors. Thus, each class will be represented by a vector of dimension M as follows,

$$\omega_{sp} = (\mathscr{N}_1(u_{sp}), \mathscr{N}_2(u_{sp}), .., \mathscr{N}_M(u_{sp})) \tag{6}$$

$$\omega_{nsp} = (\mathscr{N}_1(u_{nsp}), \mathscr{N}_2(u_{nsp}), .., \mathscr{N}_M(u_{nsp})) \tag{7}$$

where ω_{sp} and ω_{nsp} are speech and non-speech VAD vectors, respectively, M is the number of Gaussian densities in the UBM, and u_{sp} and u_{nsp} are, respectively, all the labeled speech and non-speech background feature vectors.

In the testing phase, the zeroth order statistics vector of an unknown short duration segment is first computed (ω) and the resemblance ratio score will be based on the cosine of the angle

between ω and each VAD vector as follows,

$$\begin{aligned} S_{sp}(\omega) &= cos(\omega, \omega_{sp}) - cos(\omega, \omega_{nsp}) \\ &= \frac{\omega^T}{\|\omega\|} \left(\frac{\omega_{sp}}{\|\omega_{sp}\|} - \frac{\omega_{nsp}}{\|\omega_{nsp}\|} \right) \end{aligned} \tag{8}$$

where $\|.\|$ denotes the Euclidean length and T is the transpose operation.

Experiments

In this section, we first explain the databases used for training and testing the proposed VAD model and the ASR system, the frontends, configurations, and setups. Then we will evaluate the performance of the proposed VAD algorithm in some experiments.

Database and Setup for VAD

Feature vectors are 12 dimensional Mel-Frequency Cepstral Coefficients (MFCCs) appended with delta coefficients. The UBM consists 64 Gaussian mixtures trained on the unlabeled background data. The background data is composed of about 750h broadcast news (BCN) recordings, downsampled to 8Khz, in three languages EN, ES, and DE, and of about 750h telephony signals, in DE, from inhouse data. About 28h of BCN data is labeled (25h speech and 3h non-speech) out of which 70% is used for training of VAD vectors and 30% for testing (Test Set 1). Almost all of the telephony data is labeled but we only use the same amount as labeled BCN data for training of VAD vectors and testing (Test Set 2). In other words, both Test Set 1 and 2 include about 8.5h data. Labeling is performed by using a forced alignment process.

Database and Setup for ASR

The acoustic model is a Bidirectional Long Short Term Memory (BLSTM) presented in [10] which is trained on approximately 6000h of telephony speech, including inhouse data, but also publicly available corpora such as Fisher and Switchboard. Feature vectors are 12 dimensional MFCCs. Temporal dynamics are captured by the concatenation of 9 consecutive frames and an LDA transformation is used to reduce the feature vector dimension to 45. The language model is 4gram Kneser-Ney trained on 1.5M words with approximately 2M pronunciations. More details regarding the ASR system can be found in [10]. The test set is collected from an inhouse dataset which consists of about 1500 utterances ($\approx$ 2.5h), including a total number of approximately 11,000 words.

Results

In this section, the performance of the proposed VAD is evaluated first as a classification task, then the effect on an ASR system will be presented. The evaluation for a two-class classification task (speech/non-speech) is performed by Equal Error Rate (EER) and the minimum of the Decision Cost Function (minDCF). EER is referred to the equal False Alarm Rate (FAR) and False Reject Rate (FRR) and DCF is a weighted sum of FAR and FRR in terms of the decision threshold th,

$$DCF(th) = \alpha_1 FRR(th) + \alpha_2 FAR(th) \tag{9}$$

where the weights α_1 and α_2 are defined based on the application. We have chosen $\alpha_1 = 0.75$ and $\alpha_2 = 0.25$ as proposed in [11] meaning that missing a speech segment will be more costly.

Table 1 – The effect of the data used for training the UBM and VAD vectors and the mismatch between train and test.

UBM	**VAD Vectors**	**Test Set 1**		**Test Set 2**	
		EER%	minDCF	EER%	minDCF
BCN	BCN	15.85	0.57287	25.02	0.84271
	BCN + Telephony	15.85	0.56907	23.55	0.79773
BCN + Telephony	BCN	15.81	0.57207	24.94	0.81343
	BCN + Telephony	**15.74**	**0.56323**	**23.37**	**0.76333**

Table 1 shows the performance of the proposed VAD in terms of EER and minDCF for Test Sets 1 (only BCN) and 2 (only Telephony) for segment length of 20 frames (200ms). It also shows the effect of the databases used for training of UBM and VAD vectors. As it can be seen in this table, adding telephony signals to the training process of UBM and VAD vectors improves the performance for both test sets although the improvement is not so significant. Another observation is that the performance on the BCN test data is better than on the telephony data. One reason could be due to the length of the window used for the feature normalization (mean normalization) which has been 200 frames (2s) for the BCN data and 100 frames for the telephony data, or it could be due to the background speech in the telephony data, or even some errors in the labeled data obtained by the forced alignment, which needs more investigation.

Figure 2 shows the Detection Error Trade-off (DET) curves for segment lengths from 10 to 30 frames. DET curves are obtained on the test sets 1 and 2 (pooled). As it was expected, the performance is improved by increasing the length of the processing segments. However, in order to have a reasonable accuracy and speed and have short enough segments containing only speech or non-speech signals, we use the segment length of 20 frames in the rest of the experiments. As it can be seen in this figure, the decision threshold corresponding to the EER is quite stable and the same for all the segment lengths. Having a nonsensitive threshold is important in real applications.

Figures 3 and 4 show the VAD segmentation and the corresponding confidence scores for two example utterances form BCN and telephony datasets, respectively. The quality is similar

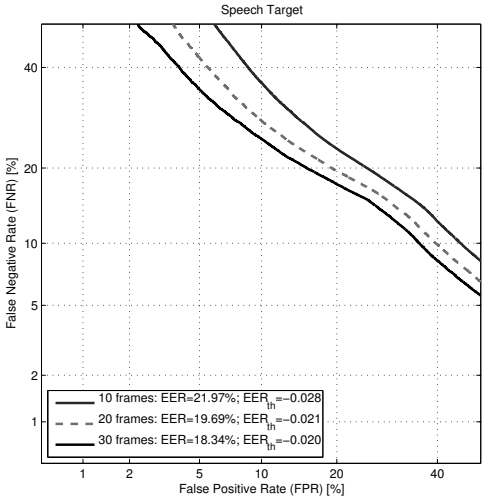

Figure 2 – DET curves for different segment lengths obtained on test sets 1 and 2 (pooled).

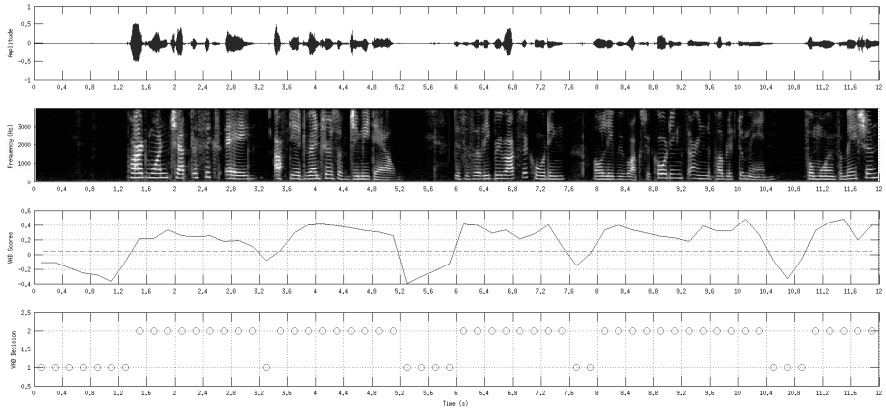

Figure 3 – An example utterance form BCN data in time domain, time and frequency domain, and the corresponding VAD scores and decisions.

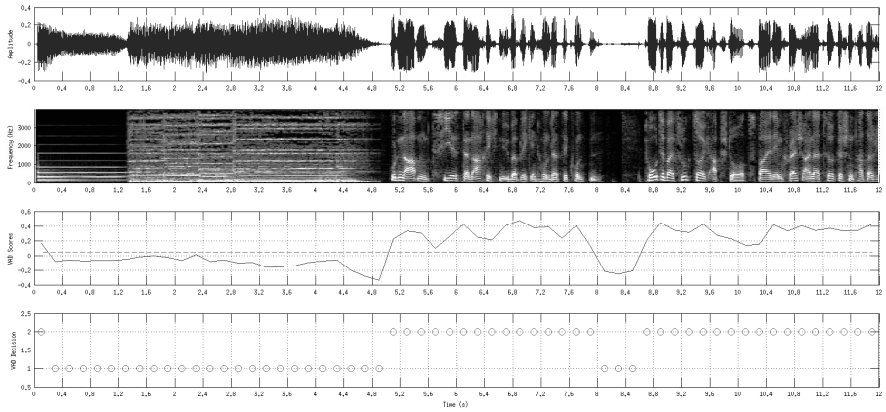

Figure 4 – An example utterance form telephony data in time domain, time and frequency domain, and the corresponding VAD scores and decisions.

for other examples. It is worth noting that no smoothing window is used unlike in other common VAD algorithms. The proposed VAD shows quite stable and reliable scores to detect accurately all kinds of non-speech segments including music and noise.

Table 2 shows the performance of the proposed VAD in the context of an ASR application. The configuration of the ASR system and the datasets were mentioned in section 3.2. The use of VAD shows, on this test set, a total relative improvement of 1.9% and 39% in WER and speed, respectively.

Conclusions

A hybrid supervised/unsupervised Voice Activity Detection (VAD) algorithm was proposed in this paper based on Baum-Welch zero order statistics obtained from a Universal Background Model (UBM). The proposed VAD was evaluated as a speech/non-speech classification task as well as in the context of an ASR application. The experimental results showed the stability

Table 2 – The effect of the proposed VAD on the performance of the ASR system.

	%SUB	%DEL	%INS	%WER	Speed (×RT)
ASR	24.1	3.9	4.0	32.0	0.89
ASR + VAD	23.6	3.9	3.9	31.4	0.54

and reliability of the VAD for both BCN and telephony data. Additionally, the proposed VAD not only improves the accuracy of an ASR system but also decreases the computational time to a great extent by filtering out the non-speech segments before decoding. The work presented in this paper is an ongoing research and, therefore, more improvements are expected in future work.

References

[1] WOO, K.-H., T.-Y. YANG, K.-J. PARK, and C. LEE: *Robust voice activity detection algorithm for estimating noise spectrum. Electronics Letters*, 36(2), pp. 180–181, 2000.

[2] ALAM, J., P. KENNY, P. OUELLET, T. STAFYLAKIS, and P. DUMOUCHEL: *Supervised/unsupervised voice activity detectors for text-dependent speaker recognition on the rsr2015 corpus*. In *Proc. Odyssey*. 2014.

[3] RYANT, N., M. LIBERMAN, and J. YUAN: *Speech activity detection on youtube using deep neural networks.* In *Proc. Interspeech*, pp. 728–731. 2013.

[4] HUGHES, T. and K. MIERLE: *Recurrent neural networks for voice activity detection*. In *Proc. ICASSP*, pp. 7378–7382. IEEE, 2013.

[5] EYBEN, F., F. WENINGER, S. SQUARTINI, and B. SCHULLER: *Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies*. In *Proc. ICASSP*, pp. 483–487. IEEE, 2013.

[6] ZHANG, X.-L. and J. WU: *Deep belief networks based voice activity detection. IEEE Transactions on Audio, Speech, and Language Processing*, 21(4), pp. 697–710, 2013.

[7] HUIJBREGTS, M., C. WOOTERS, and R. ORDELMAN: *Filtering the unknown: Speech activity detection in heterogeneous video collections*. pp. 1990–9772, 2007.

[8] KHOURY, E. and M. GARLAND: *I-vectors for speech activity detection. Proc. Odyssey*, pp. 334–339, 2016.

[9] REYNOLDS, D. and R. ROSE: *Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing*, 3(1), pp. 72–83, 1995. doi:10.1109/89.365379.

[10] FISCHER, V., O. GHAHABI, and S. KUNZMANN: *Recent improvements to neural network based acoustic modeling in the eml real-time transcription platform*. In *Proc. ESSV*. 2018.

[11] NIST: *Open speech analytic technologies pilot evaluation opensat pilot*. 2017. [Online]. Available: `https://www.nist.gov/sites/default/files/documents/2017/01/27/nist_2017_pilot_opensat_eval_plan_01-24-17_v1.1_1.pdf`.