Combining Finite State Machines and LDA for Voice Activity Detection

Elias Rentzeperis, Christos Boukis, Aristodemos Pnevmatikakis, and Lazaros C. Polymenakos

Athens Information Technology, 19.5 Km Markopoulo Ave., Peania/Athens 19002, Greece {eren,cbou,apne,lcp}@ait.edu.gr

Abstract. A robust speech activity detection system is presented in this paper. The proposed approach combines the well-known linear discriminant analysis with a finite state machine in order to successfully identify speech patterns within a recorded audio signal. The derived method is compared with existing ones to demonstrate its superiority, especially when performing on noisy audio signals, obtained with far field microphones.

1 Introduction

Voice activity detection (VAD) is a fundamental component of several modern speech processing systems like automatic speech recognition (ASR), voice commanding and teleconferencing. Providing such systems with accurate information about the existence of speech within an audio signal can result in reduction of the computational and energy requirements and improved performance of the overlying system.

Most VAD systems monitor a quantity and they compare it to a threshold in order to decide whether the observed signal is speech or not [1]. This quantity is usually the energy of the observed signal, which has presented remarkable performance with close talking (CT) microphones. The threshold can be chosen either with heuristic methods or adaptively [2], so as to be able to cope with non-stationary environments. Another approach is to use classification techniques, like the well-documented linear discriminant analysis [3], in order to distinguish speech from non-speech patterns. These techniques have noticeable results for both CT and far field (FF) microphones. The same holds for VAD systems that rely on the use of Hidden Markov Models (HMM).

The use of finite state machines (FSMs) in VAD was proposed as well [4]. These models pose some lower bounds on the duration of silence and speech intervals. Hence more accurate separation is performed since segments of very small duration characterised as speech within a silent interval are neglected and vice versa. In this paper we propose the use of a five state automaton, as was presented in [4, 5], which uses the LDA method applied to Mel Frequency Cepstral Coefficients (MFCC) as primary criterion for transition between states contrary to the approaches presented in [4, 5] which use the energy instead.

Our approach was found to have improved performance. The energy was completely neglected, since it might vary depending on the relative position of the microphone and the speaker.

This paper is organised as follows: Section 2 provides the basic background and summarises the previous VAD methods that employ FSMs. In Section 3 the proposed system is presented. The results of the performance of the introduced approach are provided in Section 4 and are compared to those of other methods. Finally Section 5 concludes the paper.

2 Background

2.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) are the dominant features used in speech applications. They are obtained by taking the inverse Fourier transform of the log spectrum after it is wrapped according to a nonlinear scale that is matching by properties of human hearing, the Mel scale. It was shown in our experiments that the addition of the first and second derivatives of the MFCC as well as of the energy of each preprocessed frame enhances the performance of the algorithm.

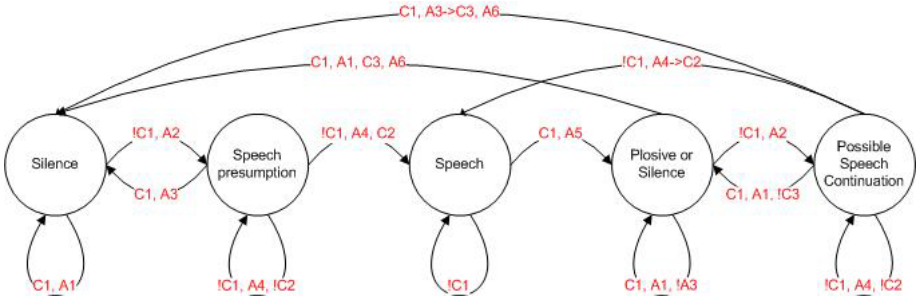

Fig. 1. Finite State Machine

2.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a method that efficiently separates data into classes [3]. In the case of VAD there are two classes to be discriminated, speech and non speech. The optimal discriminating line is derived by maximising the following criterion function

$$(\quad) = \frac{{}^{t} \quad B}{{}^{t} \quad W} \tag{1}$$

where $_B$ is the *between-class scatter matrix* and $_W$ is the *within-class scatter matrix*. $_B$ is a measure of the separation of the means of the clusters, while $_W$ is a measure of the spread of the clusters. The maximization problem reduces to a general eigenvalue one, given by

$$\overset{-1}{W} \quad B \quad = \tag{2}$$

The eigenvector that corresponds to the greatest eigenvalue from the solutions is chosen as the projecting vector of the test vectors.

2.3 Finite State Model

In [4] the use of a five state automaton was proposed for VAD. Its five states were *silence, speech presumption, speech, plosive or silence and possible speech continuation*. The transitions between states were controlled by comparing the derived short and long term energy estimates with an energy threshold. From Fig. 1 and Tab. 1, where the introduced FSM and the associated conditions and actions are presented, it is observed that a segment is characterised as speech if its duration is longer than 64 AND its energy is above the employed threshold. Similarly, a silent interval smaller than 240 is classified as plosive, and thus speech.

Table 1. Conditions and Actions of the energy controlled five state automaton for VAD

Conditions	
C1	Energy<Energy_Threshold
C2	Speech_Duration (SD)>=64ms
C3	Silence_Duration (SiD)>=240ms
Actions	
A1	$SiD = SiD + l$
A2	$SD = l$
A3	$SiD = SiD + SD$
A4	$SD = SD + l$
A5	$SiD = l$
A6	$SiD = SD = 0$

In order to improve the performance of this system the introduction of an extra criterion was proposed in [5]. This system characterised as speech segments that satisfied not only the energy but the LDA criterion as well. It does not clarify though what happens when the results of the energy and the LDA criteria do not match. The LDA was trained by using two learning databases where the speech and non-speech intervals have been manually segmented. The LDA threshold was derived from these databases as well.

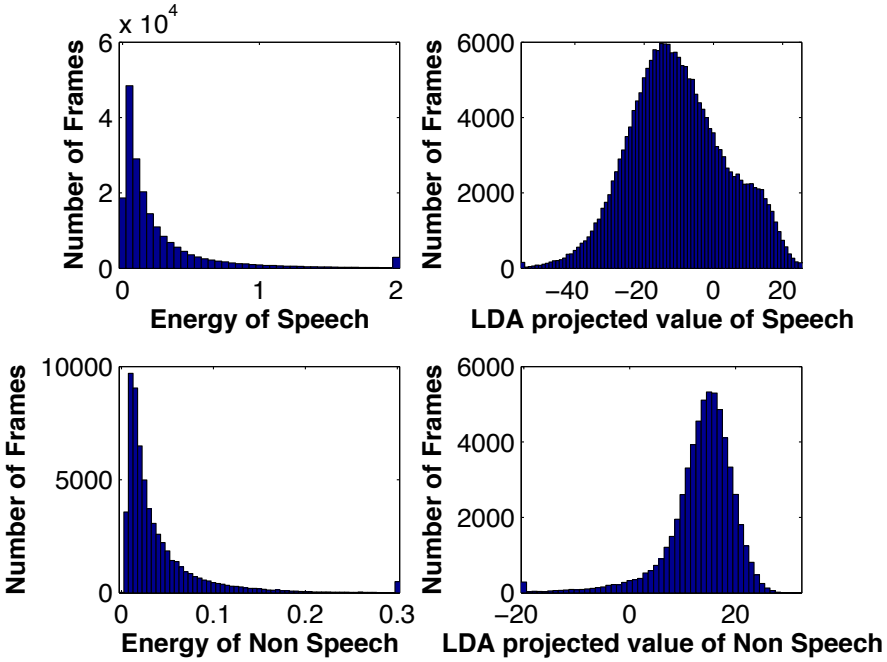

Fig. 2. Histograms of the energy and the LDA projected values of the speech/non-speech segments of the training data.

3 Proposed System

Embarking upon the observation that LDA provides more accurate discrimination between speech and non-speech classes than simply comparing the energy estimate with a threshold, and adopting the FSM of [4] a robust VAD system was developed. The choice to use LDA projection instead of energy is justified from Fig. 2 where is illustrated that the speech and silent segments have similar energy values but different LDA projections of their MFCC.

The proposed architecture used the five state automaton of Fig. 1, but the primary criterion that controlled the transition between states was derived by comparing the linear combination of the MFCC provided by the LDA, with a threshold. The LDA classifier was trained with manually segmented speech/non-speech signals. The threshold was obtained from the provided training data as well. Moreover, median filtering was applied to the results obtained from FSM in order to remove spiky decision regions and get improved error rates.

The audio signal was processed in frames. For each frame the corresponding MFCC were computed and subsequently their linear combination, which was derived by LDA, was compared to the *Threshold_LDA* to decide whether this is speech or not. Notice that the duration bounds and the time counters ( )

Table 2. Conditions and Actions of the proposed LDA controlled five state automaton for VAD

Conditions	
C1	Linear MFCC Combination<Threshold_LDA
C2	Speech_Duration (SD)>=5 frames
C3	Silence_Duration (SiD)>=16 frames
Actions	
A1	$SiD = SiD + 1$
A2	$SD = 1$
A3	$SiD = SiD + SD$
A4	$SD = SD + 1$
A5	$SiD = 1$
A6	$SiD = SD = 0$

are expressed in frames instead of msec. The proposed approach is summarised in Tab. 2.

4 Experiments

To evaluate its performance the introduced VAD system was compared to

- the approach of [4] that uses the same five state automaton, but the state transitions are controlled by the comparison of the energy estimates with an energy threshold
- the stand-alone LDA applied to MFFCs for the discrimination of the speech from the non-speech class
- the Energy Based Adaptive algorithm presented in [1] which relies on an estimation of the instantaneous SNR for the distinction of speech and non speech segments

The VAD systems were evaluated on a database collected by the University of Karlsruhe (ISL-UKA). The database is comprised of seven seminars. Each seminar contains four segments of audio data that are approximately five minutes long. The audio segments are sampled at a rate of 16.0 kHz. All the data were obtained from FF microphones resulting in comparable energy values of speech and non-speech segments (Fig. 2). Segments three and four were used for the training of the algorithm while one and two for testing. Manual human transcriptions were provided for the separation of the training segments and evaluation of the testing recordings.

The following metrics were used for the evaluation of the algorithms:

- Mismatch Rate (MR): the ratio of the incorrect decisions over the total time of the tested segment.

Table 3. Comparison of the proposed VAD with exiting approaches

Method	LDA Threshold	Energy Threshold	MR	SDER	NDER	ADER	Wpeps
LDA	4.9	-	10.09%	10.40%	8.62%	9.51%	0.09
Adaptive Energy Thresholding	-	-	18.10%	18.40%	15.60%	17.00%	0.08
FSM+LDA	4.9	-	**9.94%**	**10.19%**	**8.65%**	**9.42%**	**0.08**
FSM+Energy	-	0.043	17.28%	17.69%	14.63%	16.16%	0.08

– Speech Detection Error Rate (SDER): the ratio of incorrect decisions at speech segments over the total time of speech segments.
– Non Speech Detection Error Rate (NDER): the ratio of incorrect decisions at non speech segments over the total time of non speech segments.
– Average Detection Error Rate (ADER): the average of SDER and NDER.
– Working Point Epsilon (WPeps): an indicator of the balance between SDER and NDER. It is the absolute value of the difference between SDER and NDER over their sum.

Considering that SDER and NDER should be relatively balanced in order to draw any conclusions for the value of the algorithms, we required WPeps to be between 0 and 0.1 for the results to be valid. Under this constraint the parameter that we seek to optimize is the ADER.

Each frame consisted of 1024 samples. Furthermore the amount of overlapping between neighbouring frames was 75%. The LDA method was trained with manually segmented speech and nonspeech data. The SD threshold was 5 frames and the SiD one 16 frames, which correspond to 128 msec and 304 msec respectively, since the sampling rate was 16.0 kHz. The window size in the median filtering step was 29 frames long.

The performance of the compared VAD systems is presented in Tab. 3. From this table it is observed that the proposed method presents improved performance compared to the other approaches.

5 Conclusions

A robust voice activity detection system has been proposed in this paper, which combines a finite state machine along with the linear discriminant analysis in order to perform accurate segmentation of audio signals to speech/non-speech segments. This approach was found to outperform the stand-alone LDA and the existing approaches that combine FSMs with the energy criterion for VAD. Its performance was evaluated with noisy far field microphone recordings.

Acknowledgments: This work is sponsored by the European Union under the integrated project CHIL, contract number 506909.

References

1. D. A. Reynolds, R. C. Rose and M. J. T. Smith, PC-Based TMS320C30 Implementation of the Gaussian Mixture Model Text-Independent Speaker Recognition System, in International Conference on Signal Processing Applications and Technology, Hyatt Regency, Cambridge, Massachusetts, pp. 967–973, November 1992
2. S.Gökhun Tanyer and Hamza Özer, Voice Activity Detection in Nonstationary Noise, IEEE Trans. Sp. Au. Proc., vol. 8, no. 4, pp 479–482, Jul. 2000
3. R.O. Duda P.E. Hart and D.G. Stork, Pattern Classification, John Willey & Sons, 2001
4. L. Mauuary and J. Monné, Speech/non-speech Detection for Voice Response Systems, in *Eurospeech'93*, Berlin, Germany, 1993, pp 1097–1100
5. A. Martin, D. Charlet and L. Mauuary, Robust Speech/Non-Speech Detection Using LDA Applied to MFCC, ICASSP, 2001