Voice Activity Detection for Robust Speaker Identification System

El Bachir Tazi
UFR INTIC
Département d'Informatique
Faculté des sciences
Dhar Mahraz
USBM Fès Maroc

Abderrahim Benabbou
Département d'Informatique
Faculté des sciences et
Techniques Saiss
USBM Fès Maroc

Mostafa Harti
UFR INTIC
Département d'Informatique
Faculté des sciences
Dhar Mahraz
USBM Fès Maroc

ABSTRACT

The performances of Speaker Identification Systems (SIS) are strongly influenced by the quality of the speech signal. Most of these systems are based on Gaussian Mixture Models (GMM) that is trained using a training speech database. The mismatch between the training conditions and the testing conditions has a deep impact on the accuracy of these systems and represents a barrier for their operation in real conditions generally affected by noises disturbances. The Voice Activity Detection (VAD) is a very useful technique for improving the performance of these systems working in these scenarios. In this paper we have used within the feature extraction process, a robust VAD module, that yield high speech/non-speech discrimination accuracy and improve the performance of the SIS in noisy environments. A set of experiments which we have conducted on our proper database containing 37 Arabic speaker in order to evaluate the performances of our SIS based on gammatone frequency cepstral coefficients (GFCC) front-end combined to VAD algorithm show 7.84% average improvement of Identification Rate (IR) performance of our SIS based on GFCC robust method compared to a baseline MFCC method. 2.13% average improvement accuracy as a benefit of VAD technique is observed when the Rignal per Roise Ratio (SNR) changes from 40 dB to 0dB.

Keywords

Gaussian mixture models (GMM), Mel frequency cepstral coefficients (MFCC), Gammatone frequency cepstral coefficients (GFCC), Speaker identification system (SIS), Voice activity detection (VAD).

1. INTRODUCTION

A speaker recognition system (SRS), performing either speaker identification system (SIS) or speaker verification system (SVS), typically comprises three stages: feature extractor, pattern classifier using speaker modeling, and decision logic [1,2].

Typically, the extracted speaker features are short-term cepstral coefficients such as Mel-frequency cepstral coefficients (MFCC), Gammatone-frequency cepstral coefficients (GFCC) and perceptual linear predictive coefficients (PLPC), or long-term features such as prosody [3]. For speaker modeling, GMM are widely used to model the feature distributions and it considered actually as the state of the art, in text independent speaker identification task [4]. Such systems usually do not perform well under noisy conditions [5] because the extracted features are distorted by noise, causing mismatched likelihood calculation. A voice analysis is done after taking an input through microphone from a user. The design of the system involves manipulation of the input audio signal. At different levels, different operations are performed on the input signal such as voice activity detection, pre-emphasis, framing, windowing, spectral, cepstral analysis and identification/matching of the spoken utterance. The speaker identification task consists of two distinguished phases:

- The first phase is training sessions. This step consist to built a corpus of reference database that will serves as reference for comparing and identifying the speaker in the next step
- The second phase is a testing phase that consists of searching the identity of the speaker in test.

The following figure 1 describes the general architecture of our SIS on which we have implemented the proposed VAD algorithm.

The rest of the paper is organized as follows. Section 2 describes the VAD, GFCC and GMM techniques used to construct the SIS. The experimental conditions and evaluation results of our system are presented in Section 3. Section 4 concludes the paper.

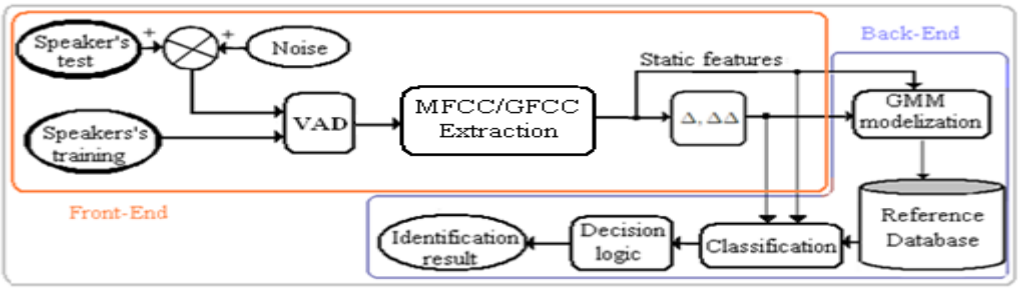

Fig. 1 General architecture of our speaker identification system (SRS)

2. DESCRIPTION OF THE USED METHODS

2.1 Voice Activity Detection

The VAD technique is frequently used in a number of applications including speech coding, speech enhancement, speech identification and speaker identification. This technique consists of extracting only the parts containing the useful speech signal by removing the parts corresponding to a silence and background noise. This will reduce the duration of recordings to their useful parts only. Hence there improved speed and performance of the SIS systems. Several implementations are reported in the literature to design a VAD module [6,7,8]. In this study we have choose the solution using the Zero Crossing Rate (ZCR) combined to the energy of the speech signal. Indeed a low rate of zero crossing and high energy are a good indicator of the presence of a speech signal, while a high rate of zero crossing rate and a low energy characterize a silence zone containing only background noise [9]. Given the fact that the noise is characterized by its random nature, and then usually it has a zero-crossing rate higher than the parts corresponding to a speech signal. In this implementation we have used the equation (1) to compute the zero crossing rate.

$$ZCR = 0.5 * \sum_{n=0}^{N-1} |sign(s_n) - sign(s_{n-1})| \qquad (1)$$

Where sign(s_n) is the sign of the instantaneous sample value of signal s(n) acquired at time n and N is the total length of the processing speech signal. In practice to discriminate between the presence and absence of the speech signal we have fixed two thresholds one for the energy and one other for the ZCR. Bellow here are the main steps of the proposed algorithm:

Step 0: initialize all parameters like thresholds of energy and ZCR (thr_zcr, thr_energy), length of frame (lengthf) etc.

Step 1: for i=1 to length of noisy speech signal to process

Step 2: framing the speech signal using the initialized lengthf

Step 3: for j=1 to length of frames do

Step 4: calculate the energy and the ZCR of the j^{th} frame

Step 5: if ZCR>thr_zcr and energy <t hr_energy

Step 6: suppression the j^{th} frame from original speech signal next j

Step 7: Improved VAD speech signal ← speech resulting in step 6 next i

The figure 2 shows an example of the resulting signal after VAD processing applied to an utterance of speech signal corrupted by 20dB SNR white Gaussian noise. We can show that the length of the resulting signal is short than the original one. This indicates that some parts of the original signal were suppressed by the VAD algorithm. These parts correspond to silence/non speech segments of the original signal and background noise. This action reducing the signal, will subsequently contribute to accelerate the speaker identification process and increase the accuracy of the SIS.

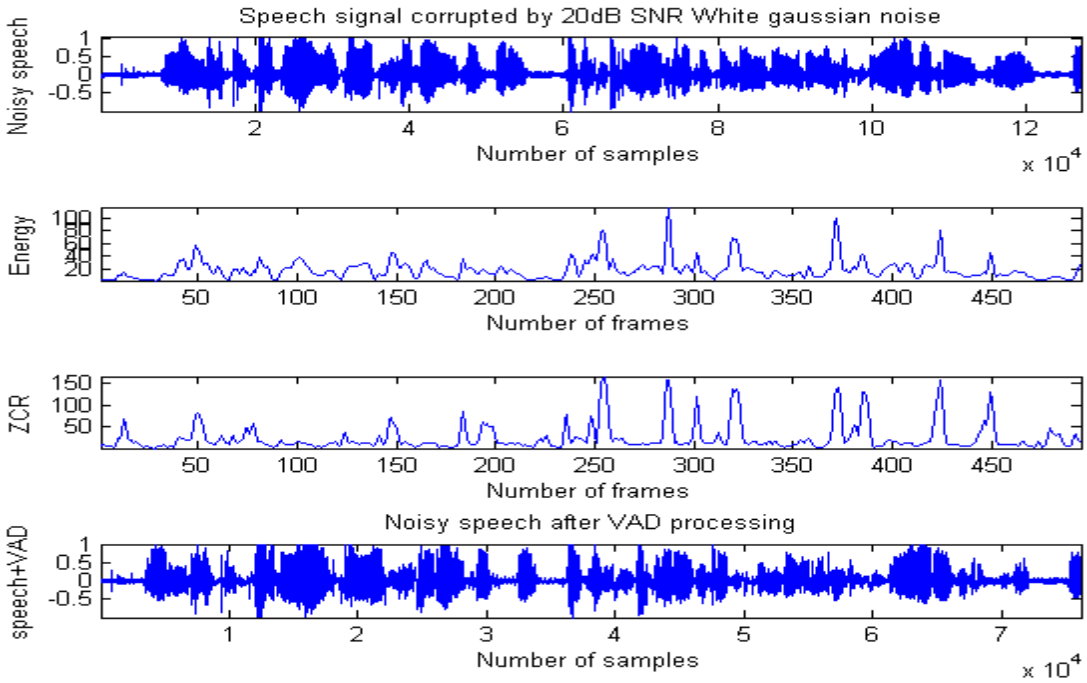

Fig. 2 Example of the input and output signals of VAD module.

2.2 Gammatone frequency cepstral coefficients

The extraction of the best parametric representation of acoustic signals is an important task to produce a better identification performance. The efficiency of this phase is important for the next phase since it affects its behavior. The overall process of the GFCC algorithm is shown in the block diagram at the following figure 3.

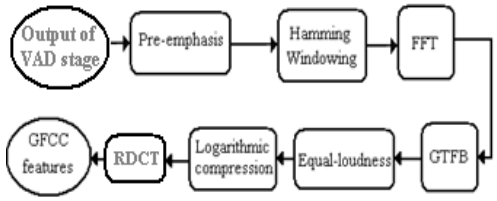

Fig. 3 Block Diagram of the GFCC process

The GFCC algorithm is another FFT-based feature extraction technique in SIS. The technique is based on the GammaTone Filter Bank (GTFB), which attempts to model the human auditory system as a series of overlapping bandpass filters [9,10]. Like the conventional MFCC previously studied in [11], feature vectors in novel and robust GFCC technique are calculated from the spectra of a series of windowed speech frames of 32ms and overlapping by 16ms. First, the spectrum of a speech frame is obtained by applying the Fast Fourier Transformation (FFT), 512 point. Then the speech spectrum is passed through 20 filter bank gammatone GTFB. Equal-loudness is applied to each of the filter output, according to the centre frequency of the filter. After that, logarithm is taken to each of the filter outputs. Finally, in order to obtain the cepstral coefficients GFCC we must transit from spectral domain to cepstral domain. For this the Reverse Discrete Cosine Transform (RDCT) is applied to the filter outputs.

2.3 Gaussian mixture modelization

The classifier of our system is based on the GMM which is considered actually as the state of the art in text independent speaker identification task [12,13]. Let i be the number corresponding to one speaker in the database, xi represents a signal belonging to the speaker i and Xxi represents the model of speaker i resulting of the signal xi. We also note £(xi/ Xxj) the likelihood of xi knowing the model Xxj.

For a yt vector of d dimension, the multi-dimensional Gaussian distribution denoted N(μ,Σ) has a probability density function $\mathcal{F}\mu,\Sigma(yt)$ given by (2).

$$\mathcal{F}\mu,\Sigma(\mathrm{yt}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Sigma)}} e^{\left(-\frac{1}{2}(\mathrm{yt}-\mu)^{T}\,\Sigma^{-1}(yt-\mu)\right)} \qquad (2)$$

Where μ and Σ are respectively the average vector of d dimension and the covariance matrix of dxd dimension of the distribution. The function £(yt/μ,Σ)=$\mathcal{F}\mu,\Sigma(yt)$ is called the likelihood function of the distribution.

The Xxi models used are the GMM (Gaussian Mixture Models). Each GMM X is a weighted sum of multivariate Gaussians (3) defined by the vector of parameters Θx=(c1, ... ck, μ1, ... μk, Σ1, ..., Σk).

Where k is the number of Gaussian components and ck the weight of the mixture associated with the kth component given that: Ck>=0 and $\sum_{i=1}^{K} \mathrm{ci} = 1$.

The likelihood for a test vector yt is produced by the mixture of Gaussian GMM X is expressed by (3)

$$£(yt/X) = £(yt/\Theta\mathrm{x}) = \sum_{i=1}^{K} \mathrm{ci}\; £(\mathrm{yt}/\mu\mathrm{i}\,, \Sigma\mathrm{i}) \qquad (3)$$

For a speech signal y containing n samples y=(y1, y2, y3, ..yn), the likelihood of this signal knowing the GMM X model is given by (4)

$$£(y/X) = \prod_{i=1}^{N} £(y_i/X) \qquad (4)$$

Where *yi* is the ith sample of *y* signal.

The Learning phase aims to estimate the parameters of Gaussian distributions that make up the models corresponding to all acoustics vectors in the database. These parameters are obtained by the K-means algorithm, and then the optimization of the values of these parameters is provided by the Expectation Maximization algorithm (EM) described in [14].

3. EXPERIMENTS AND RESULTS

3.1 Experimental conditions

In this study, we have interested to evaluate the benefit of the VAD method. For this we have used it in order to improve the GFCC front-end extraction in a text-independent monaural speaker identification context. First we have built our proper corpus database which corresponding to a population of 37 Arabic-speakers (21 male and 16 female). Each speaker had participated by 2 different recordings: one for learning the database for about 20s and one other for the test step for about 10s. All the productions sound from the speakers, were directly digitized to .wav format with a sampling frequency of 16 kHz and 16-bit monophonic quantification using the well-known software Wavesurfer®8 [15]. A white Gaussian noise, with 0 mean and unit variance, of variable level was added to the recorded signals to examine the robustness of described techniques in noisy environments that are inevitable in most real applications. The features extractors that will be considered in this set of experiments are GFCC, ΔGFCC, ΔΔGFCC without and with the VAD stage. The entire identification system is implemented under MATLAB®7 programming environment. The following table 1 describes the experiment conditions in detail.

Table 1 Experiment Conditions of the Speaker Identification Systems

Task system	Text-independent automatic speaker identification
language	Arabic
Front-ends	ΔΔMFCC, ΔΔGFCC without and with VAD stage
Back-end	Gaussian mixture models (GMM) with 2 mixture
Number of coefficients in a feature vector	36 (12 static + 12 delta + 12 delta-delta)
Window size	32 ms
Step size	16 ms
Sampling rate	16kHz
Training set	37 speakers (one utterance per speaker for about 20s)
Test set	37 speakers (one utterance per speaker for about 10s)
Noise Type	White Gaussian Noise (WGN) with 0 mean and unit variance
SNR range	0, dB, 5dB, 10dB, 15dB, 20dB, 25dB , 30dB, 35dB, 40dB
Platform	HP Elite book core i5 2.4Ghz
Programming Language	MATLAB®7
Acquisition tool	Wavesurfer®8

3.2 Experimental results

The evaluation of the identification performances of our systems was done by applying the empirical equation (5) [16].

$$C = \frac{H}{N} \times 100\% \qquad (5)$$

Where C is the percentage of correctly identified speakers called identification rate, H is the number of correctly identified speakers and N is the total number of speakers that have participated to tests identification.

The following table 2 and figure 4 show the identification rate of GFCC, ΔGFCC, ΔΔGFCC without and with VAD front-ends in various SNR conditions. These results indicate clearly that the algorithm GFCC combined to VAD stage produces interesting results. The experiments show also that:

- The VAD module improve the identification rate for both the standard MFCC and the robust GFCC front-ends.
- The dynamic variants ΔΔMFCC and ΔΔGFCC give better accuracy than the static variant but they occur a long time to estimate the parameters of the GMM models.
- Large values of Gaussian mixture number NG give better result for MFCC front-end but they occur a long time to estimate the parameters of the GMM models.
- The minimum value of NG=1 is sufficient for GFCC method to give the good result.
- The GFCC method gives better IR and robustness than conventional and classical MFCC method.

Table 2 Percentage of correctly identified speakers (C) in various SNR environments

Standart MFCC (Ncoef=36, NG=2)			Robust GFCC (Ncoef=36, NG=2)	
SNR(dB)	**ΔΔMFCC**	**ΔΔMFCC+VAD**	**ΔΔGFCC**	**ΔΔGFCC+ VAD**
0	08.10%	10.81%	10.81%	10.81%
5	16.21%	16.21%	21.62%	21.62%
10	18.91%	21.62%	24.32%	27.02%
15	24.32%	29.72%	45.94%	56.75%
20	56.75%	59.45%	70.27%	75.97%
25	78.37%	83.78%	94.59%	94.59%
30	89.18%	91.89%	100%	100%
35	97.29%	97.29%	100%	100%
40	100%	100%	100%	100%
IR Average	**54.34%**	**57.35%**	**63.06%**	**65.19%**

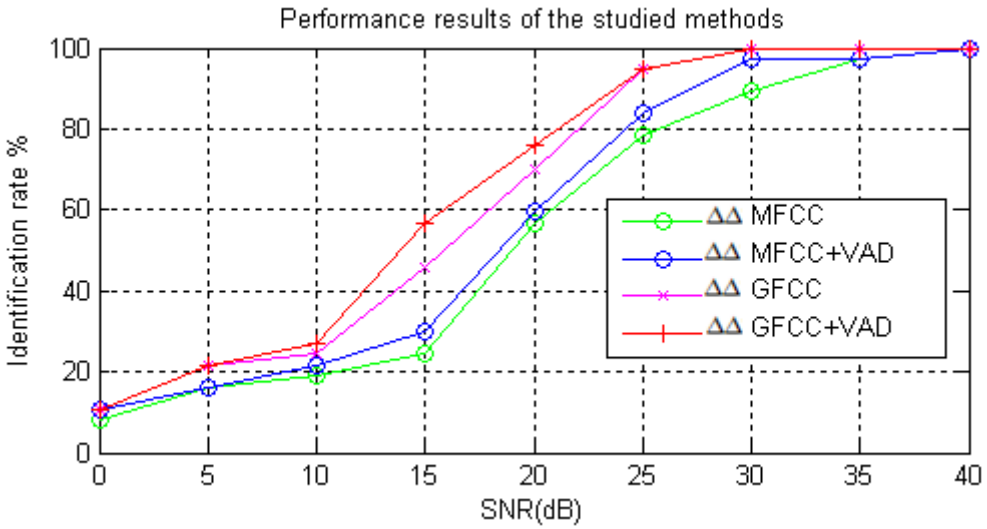

Fig. 4 Performances of the robust GFCC and standard MFCC front-ends without and with VAD technique versus SNR

4. CONCLUSION

In this paper a Voice activity detection technique combined to GFCC and MFCC font-ends for speaker identification system was studied. A mel/gammatone filtering approach is performed on white Gaussian noise reduced signal. For reduction of additive distortion the detection of speech/non speech frames based on VAD is applied. At final stage, a noise robust feature vectors, which consists of 12 mel/gammatone cepstral coefficients and their derivatives, was created. For evaluation of improvement performance of speaker identification with proposed front-ends, we have used our proper database containing 37 Arabic speakers corrupted by additive white Gaussian noise. The average improvement of 10.84% relative to the baseline MFCC front-end is achieved with 2.13% improvement is due only to the VAD module

when SNR changes from 40 dB to 0 dB. The experimental results obtained with our database show that the GFCC front-end combined to VAD method gives considerable speed and identification rate improvement when compared to the MFCC baseline system. The algorithms presented in this paper are found to be suitable for real-time applications, with acceptable quality and quantity of speech.

Since automatic speech identification and speaker identification typically share the same front-end, it is interesting to study the presented techniques also in speech identification task.

5. REFERENCES

[1] J.P. Campbell, "Speaker identification: A tutorial," Proc. IEEE, vol. 85, pp. 1437-1462, 1997.

[2] S. Furui, Digital speech processing, synthesis, and identification. New York: Marcel Dekker, 2001.

[3] D.A. Reynolds, et al., "The SuperSID project: exploiting high-level information for high-accuracy speaker identification," in Proc. ICASSP, pp. 784-787, 2003.

[4] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," Speech Comm., vol. 17, pp. 91108, 1995.

[5] Y. Shao and D.L. Wang, "Robust speaker identification using binary time-frequency masks," in Proc. ICASSP, vol. I, pp. 645-648, 2006.

[6] Sohn, J., Sung, W., 1998. A voice activity detector employing soft decision based noise spectrum adaptation. In: Internat.Conf. on Acoust. Speech Signal Process., Vol. 1, pp. 365–368

[7] J. A. Haigh and J. S. Mason, "Robust voice activity detection using cepstral features," in IEEE TEN-CON, 1993, pp. 321–324

[8] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd, "The voice activity detector for the pan European digital cellular mobile telephone service," in Proc. Int. Conf. Acoustics, Speech, Signal Processing, May 1989, pp. 369–372.

[9] W. Abdulla, "Auditory based feature vectors for speech recognition systems" Advances in Communications and Software Technologies, N. E. Mastorakis & V. V. Kluev, Editor. WSEAS Press. pp 231-236, 2002.

[10] M. Kleinschmidt, J. Tchorz and B. Kollmeier, Combining speech enhancement and auditory feature extraction for robust speech recognition, Speech Communication, Vol. 34, Issues 1-2, pp. 75-91, 2001.

[11] B. Tazi, A. Benabbou, M. Harti, "Improved Feature Extraction for Text independent Automatic Speaker Identification System" in CMT'2012, EST USMBA Fez 22,23 and 24 Mars 2012

[12] Douglas A. Reynolds et Richard C. Rose; " Robust text-independent speaker identification using gaussian mixture speaker models". IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 3, N° 1 pp: 72-83, january 1995.

[13] Reynolds, Douglas A. Thomas F. Quatieri, and Robert B. Dunn. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing. vol. 10, pp. 19-41, 2000.

[14] Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, B, 39, 1–38. December 1976.

[15] http://www.speech.kth.se/wavesurfer/

[16] S. Furui, An Overview of speaker recognition technology In Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, pages 1-9, Martigny, Switzerland, April 1994.