An adaptive speaker identification system for noisy speech

Denis Pirttiaho Cardoso
Universidade de São Paulo - USP
São Paulo - SP - Brazil
denis_dpc@yahoo.com.br

Miguel Arjona Ramírez
Universidade de São Paulo - USP
São Paulo - SP - Brazil
miguel@lps.usp.br

***Abstract*— Speaker identification is concerned with the selection of one speaker within a set of enrolled members and in this work the experiments were performed using a text-independent cohort Gaussian mixture model (GMM) speaker identification system. In order to perform the tests, TIMIT speech database is used and its corresponding version corrupted by a noisy telephone channel, i.e., NTIMIT. The vocal tract is represented by Mel-frequency cepstral coefficients (MFCC) with filter banks (FB) or, alternatively, by linear prediction cepstral coefficients (LPCC). Additionally, the cepstral mean subtraction (CMS) technique is applied to minimize the intrinsic channel distortion when the NTIMIT database is used. The utterance component for which the MFCC are calculated is obtained using a voice activity detector (VAD). However, the VADs are generally sensitive to the signal-to-noise ratio (SNR) of the utterance, being necessary to adapt them to the system operating conditions. It is provided by the proposed integration into the VAD of an SNR estimator which is based on Minima Controlled Recursive Average (MCRA), so that is necessary in order to handle both clean and noisy speech. It is observed that in high SNR utterances, such as those from the TIMIT database, the more appropriate extraction method for the MFCC was the baseline one consisting of FB, while for noisy speech the technique of CMS coupled with the extraction of MFCC from LPCC provided best results.**

***Index Terms*— Speaker identification, Gaussian mixture model, mel cepstral coeffcient, Minima Controlled Recursive Averaging.**

I. Introduction

A GMM speaker identification system [3] has two main two components. One is the training phase where speaker models are estimated and the other one is the identification phase where the most compatible model for an utterance input to the system is selected. The speech signal before being input to the identification system goes through to a preprocessing [4] stage. The MFCC algorithm is selected according to the SNR of the speech signal directly as this improves the performance of the identification system, so that clean and noisy speech may be used from NTIMIT or TIMIT databases to highlight the most appropriate technique of MFCC extraction. The clean component of the speech signal, from which the MFCC are obtained, is extracted with the use of a VAD. The VAD are generally sensitive to the SNR level of the utterance, being necessary to adapt them to conditions of operation of the system. This is handled by a noise estimator based on the method of MCRA [2] which is integrated into the VAD, thus allowing the use of clean and noisy speech. This paper is organized as follows. In section II signal preprocessing is described. Section III presents the simulation framework. Section IV presents the results as well as the experimental conditions. Section V concludes the paper.

This work was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) under Grants no. 2007/08288-2 and 2004/13206-7.

II. Preprocessing

An utterance $x(n)$ is modeled as a time-varying excitation $e(n)$ filtered by a short-time-varying filter $h(n)$ that can be considered stable over a period of typically around 10-30 ms [5]. This short-time stationary behavior can be exploited dividing the speech signal into frames and serves to characterize the vocal tract configuration given by $h(n)$ in Equation (1), which allows each speaker to be exclusively identified.

$$x(n) = e(n) * h(n) \tag{1}$$

In the preprocessing stage the speech signal undergoes pre-emphasis, segmentation, windowing and voice activity detection, as shown in Figure 1. The characterization of the filters $h(n)$ is made from each of the frames $y(m, l)$ extracted from the speech signal, where m corresponds to the sample index and l to the frame number.

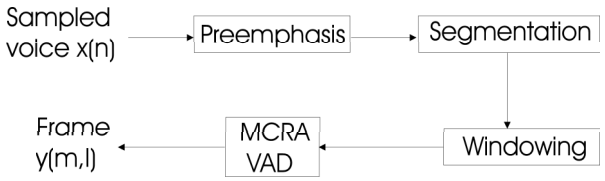

Fig. 1. Utterance preprocessing

A. Voice Activity Detector

The signal $x(n) = d(n) + \hat{x}(n)$ is composed of uncorrelated additive noise $d(n)$ and the speech signal $\hat{x}(n)$. The noise $d(n)$ must be discarded by the VAD since it impairs the performance of the identification system. The VAD is calibrated according to the quality of the signal, in order to maintain its effectiveness in situations where there are changes in the SNR of the speech signal. To overcome this limitation, allowing the use of TIMIT and NTIMIT databases, the MCRA method was adopted. It estimates the speech presence observing the ratio

between the local energy of the noisy speech and its minimum within a specified time window and this method is formulated below.

In Equation (2), $Y(k,l)$ is the short-time Fourier transform of $y(m,l)$, the frame output by the preprocessing phase, and $b(i)$ is a Hanning window of length 3. It is observed that $S_f(k,l)$ is a smoothed version of the energy spectrum of $Y(k,l)$ around frequency k.

$$S_f(k,l) = \sum_{i=-1}^{1} b(i)|Y(k-i,l)|^2 \tag{2}$$

In Equation (3), the spectrum $S(k,l)$ at each frame is a complementary linear combination between $S(k,l-1)$ and $S_f(k,l)$. The maximum value of the parameter α_s is 0.9 and it is chosen to be greater than 0.6 to reduce the influence of $S_f(k,l)$ in the composition of $S(k,l)$. This choice allows $S(k,l)$ to adapt gradually to $S_f(k,l)$ without displaying sharp peaks.

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s) S_f(k,l) \tag{3}$$

In Equation (4), the minimum value of $S(k,l)$ is found over the past D frames obtained in the preprocessing stage. This minimum value $S_{min}(k,l)$ stores the smoothed noise energy for a given frequency around the frame l.

$$S_{min}(k,l) = \arg \min_{1 \le n \le D} S(k,l-n) \tag{4}$$

In Equation (5), the integrated log spectrum normalized with respect to S_{min} is computed in the frequency band from F_{min} through F_{max} where the speech energy is concentrated. This operation ensures that the value of $S_m(l)$ rises in presence of clean speech in a similar way irrespective of the utterance quality.

$$S_m(l) = \frac{1}{N2 - N1 + 1} \sum_{k=N1}^{N2} log \frac{S(k,l)}{S_{min}(k,l)} \tag{5}$$

$$N1 = \frac{F_{min}}{F_s} M \tag{6}$$

$$N2 = \frac{F_{max}}{F_s} M \tag{7}$$

In Figure 2 the evolution of $S_m(l)$ is displayed for corresponding phrases in TIMIT and NTIMIT databases. The speech presence is detected in frames where $S_m(l)$ exceeds the threshold δ represented by the horizontal line. It is observed that the value of δ must be experimentally calibrated and its value is independent of the frame length due to the frequency averaging in Equation (5). The calibration process consist of rebuilding the utterance from the frames that already passed through the VAD and whose value of $S_m(l)$ exceeds δ. The noise frames are replaced by null frames in the rebuilding process. When a good acoustic quality is reached for the reconstructed signal, the corresponding minimum value of the average log spectrum is assigned to δ.

After an initial period of stabilization, a strong correlation has been observed between frames marked for speech presence for signals extracted from either database. This is a confirmation of the effectiveness of the method.

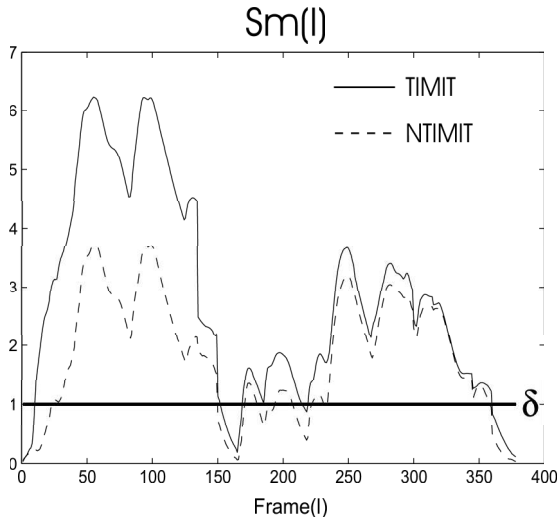

Fig. 2. VAD result with MCRA for TIMIT and NTIMIT databases

III. System framework

The system employed in the simulations consists of a module for preprocessing, extraction of the MFCC, identification and training. The MFCC are extracted from FB or LPCC [4] in accordance with the choice made by the user of the system as shown in Figure 3.

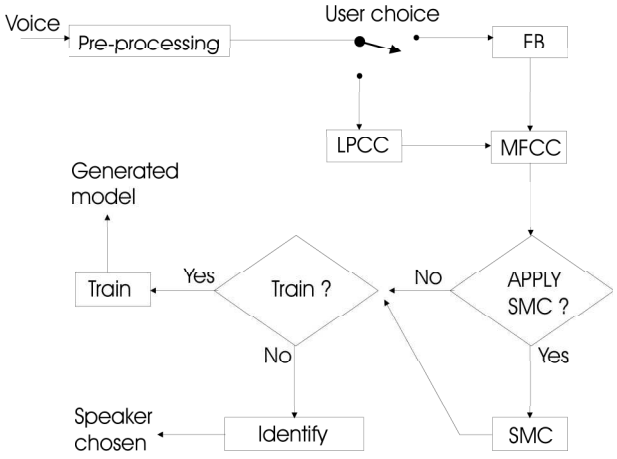

Fig. 3. System Framework diagram

A. MFCC derivation from FB

In the derivation of MFCC the filter bank proposed by *Slaney* [6] was used which provides a better speaker discrimination [8]. As can be seen in Figure 4, this filter bank is composed of 40 filters whose center frequencies for the first 13 filters are linearly spaced while the other 27 ones are

logarithmically spaced. The frequency scales for these filters are represented by Equations (8) and (9) respectively.

$$F_{linear} = 133.33 + 66.66i \qquad 1 \leq i \leq 13 \qquad (8)$$

$$F_{log} = 1000(1.0711703)^{(i-13)} \qquad 14 \leq i \leq 27 \qquad (9)$$

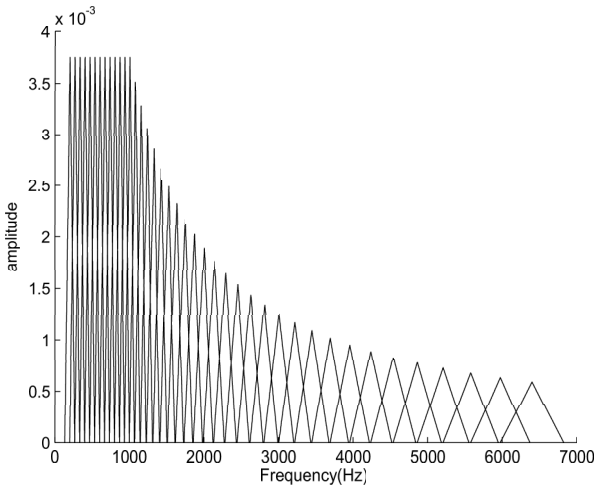

Fig. 4. Spectrum of Slaney filter bank

B. MFCC derivation from LPCC

The technique of linear prediction consists of estimating the current value of a signal $s(n)$ from its previous *P* samples. The prediction coefficients $a(i)$ in Equation (10) are computed [7] where $R(m)$ is the autocorrelation function of the signal $s(n)$.

$$\begin{bmatrix} R(0) & \dots & R(P-1) \\ R(1) & \dots & R(P-2) \\ \dots & \dots & \dots \\ R(P-1) & \dots & R(0) \end{bmatrix} \begin{bmatrix} a(1) \\ a(2) \\ \dots \\ a(P) \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \dots \\ R(P) \end{bmatrix} \qquad (10)$$

The autocorrelation function $R(m)$ is given by Equation (11) where M represents the length of frame obtained in the preprocessing phase.

$$R(m) = \sum_{n=0}^{M-m-1} s(n)s(n+m) \qquad m = 0, 1, \dots, P \qquad (11)$$

As can be seen, the linear system represented by Equation (10) presents Toeplitz simmetry and, therefore, can be solved using the Durbin algorithm [10] described by Equations (12) to (14).

$$E^{(0)} = R(0) \qquad (12)$$

$$k_i = \frac{R(i) - \sum_{j=1}^{i-1} \alpha_j^{i-1} R(|i-j|)}{E^{i-1}} \qquad 1 \leq i \leq P \qquad (13)$$

$$\begin{aligned} &For \quad j = 1, \dots, i-1 \qquad (14) \\ &\alpha_i^i = k_i \\ &\alpha_j^i = \alpha_j^{i-1} - k_i \alpha_{i-j}^{i-1} \\ &E^{(i)} = (1 - k_j^2) E^{(i-1)} \end{aligned}$$

Equations 13 and 14 are iterated for increasing prediction order $1 \leq i \leq P$. Finally, the linear prediction coefficients are obtained at prediction order P as Equation (15).

$$For \quad l = 1, \dots, P \qquad a(l) = \alpha_l^P \qquad (15)$$

The LPCC given by $\widetilde{a}^{(i)}(m)$ are computed by the algorithm in Equations (16) where the constants C and P correspond respectively to the number of cepstral coefficients and prediction coefficients defined in Table I.

$$For \ i = -P, \dots, -2, -1, \quad \widetilde{a}^{(i)}(m) \quad is \ equals \ to: \qquad (16)$$

$$\begin{cases} a(-i) + \alpha . \widetilde{a}^{(i-1)}(0) & m = 0 \\ (1 - a^2)\widetilde{a}^{(i-1)}(0) + \alpha . \widetilde{a}^{(i-1)}(1) & m = 1 \\ \widetilde{a}^{(i-1)}(m-1) + \alpha(\widetilde{a}^{(i-1)}(m) - \widetilde{a}^{(i)}(m-1)) & m = 2, \dots, C \end{cases}$$

After that, the normalized coefficients $\widetilde{a}(k)$ are determined through Equation (17).

$$\widetilde{a}(k) = \frac{\widetilde{a}^{(0)}(k)}{\widetilde{a}^{(0)}(0)} \qquad 1 \leq k \leq C \qquad (17)$$

The MFCC represented by $\widetilde{c}(m)$ is given in Equation (18).

$$\widetilde{c}(m) = \widetilde{a}(m) + \sum_{k=1}^{m-1} \frac{k}{m} \widetilde{c}(k)\widetilde{a}(m-k) \qquad m = 1, \dots, C \qquad (18)$$

IV. Results and experimental conditions

The TIMIT and NTIMIT databases are formed by 8 dialect region directories identified as *DR1*, *DR2*, ..., *DR8* and these directories store speech signals sampled at a rate of 16 kHz.

The simulations were performed to investigate the effect of noise on the performance of the identification system in accordance with the method for MFCC derivation. In the simulations involving noisy speech the CMS technique [9] was applied to reduce the effect of telephone channel interference. The constants adopted in the framework tests given by Figure 3, as well the Equations where they were initially used are reported in Table I.

TABLE I

CONSTANTS OF SIMULATION

constant	value	equation
α_s	0.8	3
D	120	4
F_{min}	500	6
F_{max}	3400	7
F_s	16000	6
P	14	10
M	256	11
C	23	16

In Table II the performance is shown for the identification system using DR1 directory. The DR1 directory contains sentences uttered by 22 different speakers, from which models with 24 mixtures were trained using 23 MFCCs. The average signal was 23 s long to be trained and 5.5 s long to undergo the identification tests.

TABLE II

PERFORMANCE OF THE IDENTIFICATION SYSTEM

database	MFCC from filter bank	MFCC from LPCC
TIMIT	100 %	95%
NTIMIT+CMS	68 %	86%

It is worth noting that similar results are achieved when using speech sentences from other directories, i.e., the system offers better performance for signals with high SNR from the TIMIT database when the MFCCs are derived from FB. For signals from the NTIMIT database with lower SNRs, it is observed that the best MFCC are obtained by linear prediction.

V. FINAL COMMENTS

In this work a framework for the construction of a GMM identification system has been proposed. With the use of the MCRA method integrated into the VAD it has become possible to employ signals with different SNR characteristics, like those from TIMIT and NTIMIT databases, with no need to make adjustments to the identifying and training system to tailor them to the speech signal quality.

The framework was proposed in a modular fashion, as can be seen in Figure 3, so that it is possible to expand it with the inclusion of new techniques for characterization of the vocal tract.

It was observed by the results shown in table II that the systems whose input signals have high SNR exhibit better performance when the MFCC are derived from FB. For poorer quality signals, such as those from the NTIMIT database, it was found that the derivation of MFCC from LPCC is the better one.

Therefore, the choice of extraction method for obtaining the MFCCs in an identification system based on GMM depends on the observation of the speech quality to guarantee its maximum performance.

REFERENCES

[1] K. Tokuda, K. Kobayashi, “Recursive calculation of mel-cepstrum from LP coefficients”, *Journal of Tokyo Institute of Technology*, pp. 1-7. April 1994.

[2] Cohen, I.; Berdugo, B., “Noise estimation by minima controlled recursive averaging for robust speech enhancement”, *Signal Processing Letters*, vol. 9, Issue 1, pp. 12-15. January 2002.

[3] Reynolds, D.A.; Rose, R.C., “Robust text-independent speaker identification using Gaussian mixture speaker models”, *Speech and Audio Processing, IEEE Transactions*, vol. 3, Issue 1, pp. 72-83. January 1995.

[4] Lawrence Rabiner, *Fundamentals of speech recognition*, pp. 113–114, 1993, Ed. Prentice Hall PTR, United States.

[5] Campbell, J.P., “Speaker recognition: A tutorial”, *Proc. of the IEEE*, vol. 85, no. 9, pp. 1437-1462. September 1997.

[6] Slaney, M., “Auditory Tool Box Version 2”, *Technical Report #1998-010*, Interval Research Corporation, 1998.

[7] Rabiner L., “Fundamentals of speech recognition”, pp. 411, 1993, Ed. Prentice Hall PTR, United States.

[8] Ganchev T., Fakotakis N., Kokkinakis G., “Comparative evaluation of various MFCC implementations on the speaker verification task”, *10th International Conference on Speech and Computer (SPECOM 2005)*, vol. 1, pp. 191-194, 2005

[9] C. Kermorvant, “A comparison of noise reduction techniques for robust speech recognition”, *IDIAP-RR 10*, 1999

[10] Rabiner L., “Fundamentals of speech recognition”, pp. 115, 1993, Ed. Prentice Hall PTR, United States.