# A Method of Automatic Speaker Recognition Using Cepstral Features and Vectorial Quantization

José Ramón Calvo de Lara

Advanced Technologies Application Center, CENATAV, Cuba
jcalvo@cenatav.co.cu

**Abstract.** *Automatic Speaker Recognition* techniques are increasing the use of the speaker's voice to control access to personalized telephonic services. This paper describes the use of vector quantization as a feature matching method, in an automatic speaker recognition system, evaluated with speech samples from a SALA Spanish Venezuelan database for fixed telephone network. Results obtained reflect a good performance of the method in a text independent job in the context of sequences of digits.

## 1 Introduction

*Automatic Speaker Recognition* techniques make it possible to use the speaker's voice to verify their identity and control access to services such as voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers [1].

These techniques can be classified into identification and verification. *Speaker identification* is the process of determining which registered speaker provides a given utterance. *Speaker verification* is the process of accepting or rejecting the identity claim of a speaker.

Speaker Recognition methods can be divided into *text-independent* and *text dependent*. In a *text-independent* method, speaker models capture characteristics of speaker's speech *irrespective of what one is saying*. In a *text-dependent* method the recognition of the speaker's identity is based on his/her *speaking specific phrases*, like passwords, card numbers, PIN codes, etc.

Speaker Recognition systems contain two main processes: *feature extraction* and *feature matching*. *Feature extraction* extracts a small amount of data from the voice signal that can be used later to represent each speaker. *Feature matching* involves the procedure to identify the unknown speaker by comparing extracted features from his/her voice input with the ones from a set of known speakers.

An Automatic Speaker Recognizer has to serve two pattern recognition phases. The first one is the *training phase* while the second one is the *testing phase*. In the *training phase*, each registered speaker provides samples of their speech so that the system can train a reference model for that speaker. In case of speaker verification

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 146 – 153, 2005.

systems, in addition, a speaker-specific threshold is also computed from the training samples. During the *testing phase*, the input speech is matched with stored reference model(s) and recognition decision is made.

This paper refers the author's experience in the design and test of a text independent speaker recognition method, with a vector quantization algorithm of feature matching, evaluated with speech samples obtained from SALA database for fixed telephone network.

## 2 Feature Extraction from Speech Samples

The feature extraction from the speech samples consists of a filtering process with pre-emphasis and an extraction process of spectral features using a short term analysis [2]. The 8bit μ-law samples of corpus recorded at a sampling rate of 8 kHz were converted to linear 16 bit PCM samples.

### 2.1 Filtering Process with Pre-emphasis

Pre-emphasis refers to filtering that emphasizes the higher frequencies of speech; its purpose is to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region. The pre-emphasis makes the upper harmonics of the fundamental frequency more distinct, and the distribution of energy across the frequency range more balanced.

### 2.2 Extraction of Spectral Features

The extraction process of spectral features using a short term analysis consists in:

- A frame blocking, where the continuous speech signal is blocked into frames of 256 samples, with adjacent frames separated by 100 samples.
- A frame windowing, a Hamming window is applied to each individual frame in order to minimize the signal discontinuities, and consequently the spectral distortion, at the beginning and end of each frame.
- A Discrete Fourier Transform process using a FFT algorithm, which converts each frame of 256 samples from the time domain into the frequency domain, the result obtained is the *signal's periodogram.*

A wide range of possibilities exist for representing the speech signal in Automatic Speech and Speaker Recognition with spectral features as Linear Prediction Coefficients (LPC), Linear Prediction Cepstrals Coefficients (LPCC) and Mel-Frequency Cepstrals Coefficients (MFCC) and others [3].

MFCC are perhaps the best known and most popular spectral features for representing the speech signal, widely used in many speech and speaker recognizers [4], these are used in this speaker recognizer. Dynamic spectral features known as *delta* and *delta-delta* features are calculated too, and appended to MFCC.

#### 2.2.1 MFCC Features

Psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. MFCC features are based

on the known variation of the human ear's critical bandwidths with frequency; filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech.

Thus for each tone with an actual frequency, *f*, measured in Hz, a subjective pitch is measured on a scale called the *'mel'* scale. The *mel-frequency* scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 *mels*.

In order to simulate the frequency warping process, we use a filter bank, one filter for each desired *mel-frequency* component. That filter bank has a triangular band-pass frequency response, and the spacing as well as the bandwidth is determined by a constant m*el-frequency* interval. A *mel-spaced* filter bank with 12 filters is given in figure 1.

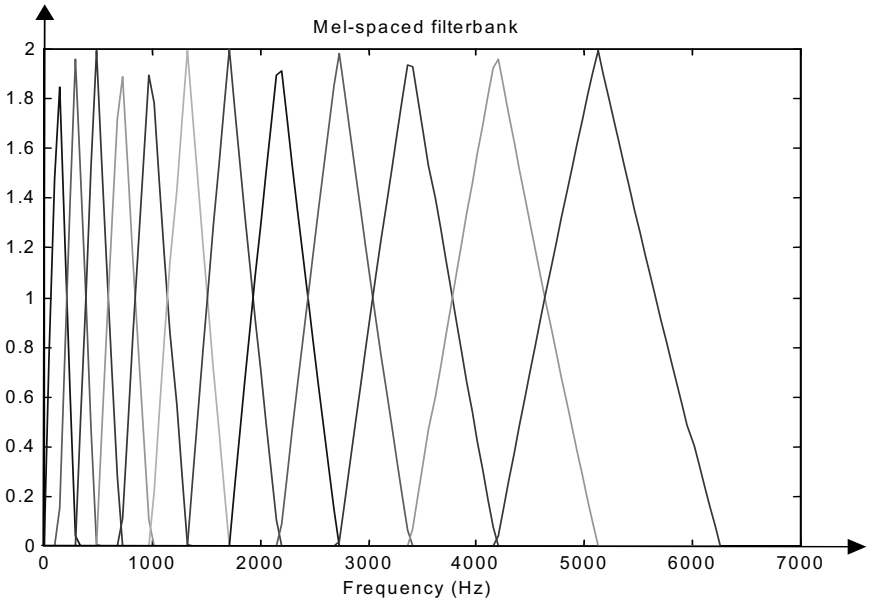


**Fig. 1.** Mel-spaced filter bank with 12 filters [1]

The modified or mel power spectrum consists of the output power of these filters applied to the periodogram. The number of mel-spaced filters and mel power spectrum coefficients is typically chosen as 20.

At last, we convert the log mel spectrum back to time, to obtain the mel-frequency Cepstrum Coefficients (MFCC). Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT).

The first component k= 0 is excluded from the DCT since it represents the mean value of the input signal which carried little speaker specific information. Twelve cepstral coefficients of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis.

By applying the procedure described above for each speech frame, an acoustic vector of 12 mel-frequency cepstrum coefficients is computed. These are result of a cosine transform of the logarithm of the short-term power spectrum expressed on a mel-frequency scale. Therefore each input utterance is transformed into a temporal sequence of acoustic vectors. A block diagram of the MFCC extraction process is given in figure 2.

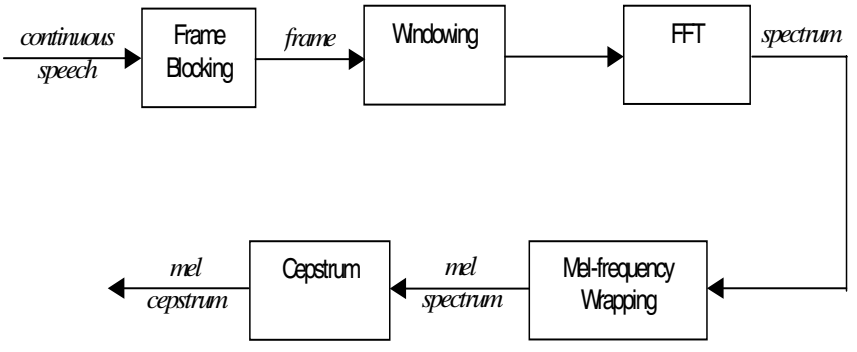


**Fig. 2.** Mel-Frequency Cepstrum Coefficients extraction process [1]

### 2.2.2 Extracting Delta and Delta-Delta Features

A widely method to encode some of the dynamic information over time of spectral features is known as *delta features* "Δ" [3, 5]. The time derivatives of each cepstral coefficient are obtained by differentiation and zero padding at begin and end of the utterance, then, the estimate of the derivative is appended to the acoustic vector, yielding a higher-dimensional feature vector. The time derivatives of the delta features are estimated also, using the same method, yielding *delta-delta features* "ΔΔ". These are again appended to the dimensional feature space. In our case we obtained a 36-dimmension acoustic vector: 12 MFCC + 12Δ + 12 ΔΔ.

## 3 Feature Matching

The problem of automatic speaker recognition is a pattern recognition problem. The goal of pattern recognition is to classify objects into one of a number of classes. In our case, the objects or patterns are sequences of acoustic vectors that are extracted from an input speech using the techniques described in the previous section. The classes refer to individual speakers. Since the classification procedure in our case is applied on extracted features, it can be referred to as feature matching.

Furthermore, if there are a set of patterns which classes are known, then it is a problem of supervised pattern recognition. During the training phase, we label the sequence of acoustic vectors of each input speech with the ID of the known speakers; these patterns comprise the training set and are used to derive a classification algorithm. The remaining sequences of acoustic vectors are then used to test the classification algorithm; these patterns are referred to as the test set. If the correct classes of

the individual pattern in the test set are also known, then one can evaluate the performance of the algorithm.

### 3.1 Vector Quantization Method of Feature Matching

The state-of-the-art in feature matching techniques used in speaker recognition includes Dynamic Time Warping (DTW), Hidden Markov Modelling (HMM), and Vector Quantization (VQ). In this system, the VQ approach is used, due to ease of implementation and high accuracy.

VQ is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codeword is called a *codebook.* Figure 3 shows a diagram to illustrate this process.

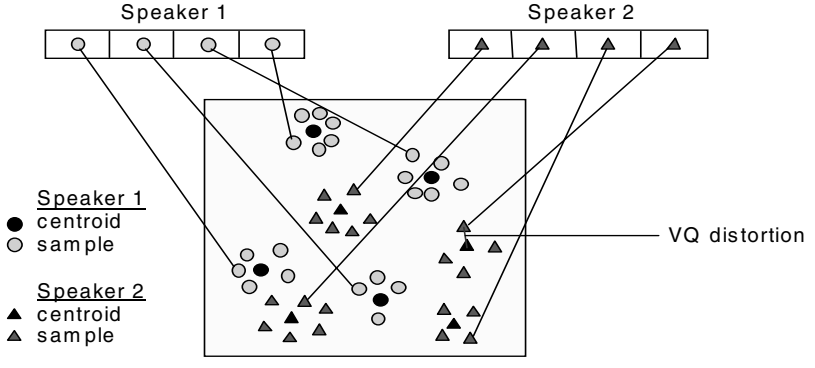


**Fig. 3.** Conceptual diagram illustrating vector quantization codebook formation [6]

In the figure, only two speakers and two dimensions of the acoustic vectors space are shown. The circles refer to the acoustic vectors from speaker 1 while the triangles are from speaker 2. In the *training phase*, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors.

The result codewords (centroids) are shown by black circles and black triangles for speaker 1 and 2, respectively. The distance from any acoustic vector to the closest codeword of a codebook is called a VQ-distortion. In the *testing phase*, an input utterance of an unknown voice is *"vector-quantized"* using each trained codebook and the total VQ distortion is computed. The speaker corresponding to the VQ codebook with smallest total VQ-distortion is identified.

#### 3.1.1 LBG Algorithm

In the training phase, a speaker-specific VQ codebook is generated for each known speaker by clustering his/her training acoustic vectors using a well-know algorithm namely LBG [7], this recursive algorithm cluster a set $X = \{x_1,....., x_T\}$ of acoustic vectors into a codebook $C = \{c_1,.....c_M\}$ of M codewords (M power of 2) . The algorithm is formally implemented by the following recursive procedure [1]:

1. Design a 1-vector codebook, this is the centroid of the set of training vectors
2. Double the size of the codebook by splitting each current codebook $\mathbf{y}_n$ according to the rule:

$$\mathbf{y}_n^+ = \mathbf{y}_n(1+\varepsilon) \tag{1}$$

$$\mathbf{y}_n^- = \mathbf{y}_n(1-\varepsilon) \tag{2}$$

   Where $n$ varies from 1 to the current size of the codebook, and $\varepsilon$ is a splitting parameter ($\varepsilon$ =0.01).
3. Nearest-Neighbor Search: for each training acoustic vector, find the codeword in the current codebook that is closest in terms of VQ-distortion, and assign that vector to the corresponding cluster associated with the closest codeword.
4. Centroid Update: update the codeword in each cluster using the centroid of the training acoustic vectors assigned to that cluster.
5. Repeat steps 3 and 4 until the VQ distortion falls below a preset threshold.
6. Repeat steps 2, 3 and 4 until a codebook size of $M$ is designed.

The generated codebook $C$ contains the codewords that better represents the training set of acoustics vectors $X$ in terms of VQ-distortion.

### 3.1.2 Measure of VQ-Distortion

Consider an acoustic vector $x_i$ generated by any speaker, and a codebook $C$, the VQ-distortion $d_q$ of the acoustic vector $x_i$ with respect to $C$ is given by:

$$d_q(x_i, C) = \min d(x_i, c_j) \tag{3}$$

Where $d(.,.)$ is a distance measure defined for the acoustic vectors. The codeword cj for which $d(x_i, c_j)$ is minimum, is the nearest neighbor of $x_i$ in the codebook $C$.

Euclidean distance is a distance measure used due the straightforward implementation and intuitive notion (Euclidean distance between two cepstral features, measures the squared distance between the corresponding short term log spectra) [3].

In the testing phase, all the sequences of acoustic vectors from an unknown speaker is "vector-quantized" computing the average quantization distortion $D_Q$ with each trained codebook $C$, the known speaker corresponding to the codebook C with smallest $D_Q$ is assigned to unknown speaker. The average quantization distortion $D_Q$ is defined as the average of the individual distortions:

$$D_Q(X, C) = \frac{1}{T}\sum_{i=1}^{T} d_q(x_i, C) \tag{4}$$

## 4 Experimental Results

The proposed speaker recognizer was evaluated with sequences of digits obtained from 347 speakers of SALA database. A sequence of about 15 sec of duration was

used for training and other sequence of similar duration was used for testing. Until now SALA Database had been used only in speech recognition studies[8].

### 4.1 SALA Database

The SALA Spanish Venezuelan Database for fixed telephone network was recorded within the scope of the SpeechDat Across Latin America project. [9] The design of the corpus and the collection was performed at the Universidad de los Andes, Mérida Venezuela, transcription and formatting was performed at the Universidad Politécnica de Cataluña, Spain.

This database comprises telephone recording from 1000 speakers recorded directly over the PSTN using two analogue lines, signals were sampled at 8 kHz and μ-law encoded without automatic gain control. Every speaker pronounces 44 different utterances.

The database has the following speaker demographic structure:

- Five dialectal regions: Central, Zuliana, Llanos, Sud-Oriental and Andes
- Five age groups: under 16, 16 to 30, 31 to 45, 46 to 60 and over 60

13 speakers called more than once using the same prompt sheet.

### 4.2 Evaluation Results

The following table shows the 30 distribution groups of the 347 speakers:

**Table 1.** Distribution of groups of speakers for the evaluation

| **Age** | **16-30** | | **31-45** | | **46-60** | |
|---|---|---|---|---|---|---|
| **Regions** | **F** | **M** | **F** | **M** | **F** | **M** |
| **Central** | 12 | 12 | 12 | 12 | 8 | 12 |
| **Zuliana** | 12 | 11 | 12 | 12 | 12 | 11 |
| **Llanos** | 12 | 12 | 12 | 12 | 12 | 12 |
| **Sud-Oriental** | 12 | 12 | 12 | 12 | 5 | 12 |
| **Andes** | 12 | 12 | 12 | 12 | 12 | 12 |

The speaker recognizer was evaluated within every one of the 30 groups, obtaining the following results:

**Table 2.** Results of the evaluation

| | **speakers** | **identified** | **%** |
|---|---|---|---|
| **F** | 169 | 168 | **99.4** |
| **M** | 178 | 175 | **98.3** |
| | 347 | 343 | **98.8** |

An additional evaluation, using the 13 speakers that called more than once, taking a sequence of digits of the first call for training and other sequence of digits of the second call for testing, shows a 92.3 % of identification rate.

## 5 Conclusion and Future Work

This paper describes the result of the application of a vector quantization speaker recognition method, used in a text independent job in the context of sequences of continuos digits and evaluated with a database for fixed telephone network. This kind of job and environment isn't usual for vector quantization methods [3,4].

Many as 98.8% of speakers in a group of 347 speakers of SALA Database were identified correctly. Such a result may be regarded as a promising way to a high-performance speaker identification system. However, it has to be taken into account that the speech data used in the experiments were recorded during one session. More exhaustive test must be performed in order to probe the method when there is a time interval between the recording of training and testing sentences.

## References

1. Minh N. Do, "An Automatic Speaker Recognition System", Audio Visual Communications Laboratory, Swiss Federal Institute of Technology, Lausanne, Switzerland, 2001. http://lcavwww.epfl.ch/~minhdo/asr_project/asr_project.doc
2. Douglas A. Reynolds, "An Overview of Automatic Speaker Recognition Technology", MIT Lincoln Laboratory, Lexington, MA, USA, This paper appears in ICASSP 2002, pp 4072-4075.
3. Tomi Kinnunen, "Spectral Features for Automatic Text-Independent Speaker Recognition", University of Joensuu, Department of Computer Science, Joensuu, Finland, December 21, 2003. ftp://cs.joensuu.fi/pub/PhLic/2004_PhLic_Kinnunen_Tomi.pdf
4. Joseph P. Campbell, Jr, "Speaker Recognition: A tutorial", DoD. Proceedings of the IEEE, Vol 85, No. 9 September 1997, pp. 1437-1462.
5. Mason, J., and Zhang, X. "Velocity and acceleration features in speaker recognition", Department of Electrical & Electronic Engineering, Univ. Coll., Swansea. This paper appears in ICASSP 1991, pp. 3673–3676.
6. F.K. Song, A.E. Rosenberg and B.H. Juang, "A vector quantisation approach to speaker recognition", AT&T Technical Journal, Vol. 66-2, pp. 14-26, March 1987.
7. Y. Linde, A. Buzo & R. Gray, "An algorithm for vector quantizer design", *IEEE Transactions on Communications*, Vol. 28, pp.84-95, 1980.
8. L. Maldonado, E. Mora; Universidad de los Andes, Mérida, Venezuela: Personal communications with the author, 2004.
9. A. Moreno, R. Comeyne, K. Haslam, H. van den Heuvel, H. Höge, S. Horbach, G. Micca : "SALA: SpeechDat Across Latin America: .Results Of The First Phase",, LREC2000: 2nd International Conference on Language Resources & Evaluation, Athens, Greece 2000.