Effect of Feature Extraction Techniques on the Performance of Speaker Identification

M. Elkholy and N. Korany
Electrical Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt
Email: eng_marwa_t@yahoo.com and nokorany@hotmail.com

***Abstract* —In this paper, the effect of features extracted on the performance of speaker identification engine is investigated. Vector Quantization (VQ) is implemented and used as identification engine. Three type of speech features, Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Predictive (PLP), and Relative Spectral Technique-Perceptual Linear Predictive (RASTA-PLP) are extracted and used for the classification problem. One word per speaker is used within the train phase and the identification rate is calculated for each feature extraction technique. The calculation is repeated using various word of different spoken time, and the paper specifies the feature extraction technique that fits with the Vector Quantization (VQ) recognition engine.**

***Index Terms*—speaker recognition, speaker identification, vector quantization, relative spectral technique - perceptual linear predictive (RASTA-PLP), perceptual linear prediction (PLP), mel frequency cepstral coefficients**

I. Introduction

Speaker recognition [1] is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas, and remote access to computers. Speaker recognition can be classified into Identification and Verification. Speaker identification is the process of determining which registered speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Fig. 1 and Fig. 2 show the basic structures of speaker recognition systems for speaker identification and those in verification systems respectively.

The system is classified as text-independent speaker identification system since its task is to identify the person who speaks regardless of what is saying. However this task has been challenged by the highly variant of input speech signals. Speech signals in train and test phases differ greatly due to many facts such as people voice change with time, health conditions (e.g. the speaker has a cold), speaking rates, and so on. There are also other factors, beyond speaker variability, that present a challenge to speaker recognition technology, e.g. acoustical noise and variations in recording environments.

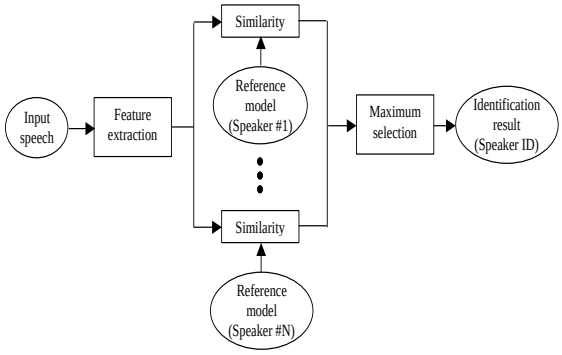

Figure 1. Basic structures of speaker identification.

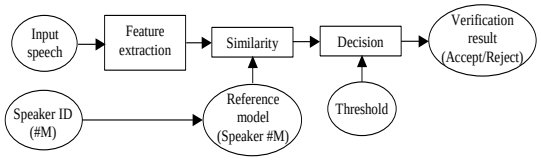

Figure 2. Basic structures of speaker verification system.

An interesting issue is how much relevant information related to speaker recognition is lost within this analysis. Thus, it is concerned with Feature Extraction from different parametric representations Relative Spectral Technique-Perceptual Linear Predictive (RASTA-PLP), Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual linear predictive (PLP).

II. Speaker Identification Engine

Vector quantization (VQ) is a lossy data compression method based on the principle of block coding. It is a fixed-to-fixed length algorithm. In the earlier days, the design of a vector quantize (VQ) is considered to be a challenging problem due to the need for multidimensional integration. In 1980, Linde Buzo Gray (LBG) proposed a VQ design algorithm based on a training sequence. In the following, the test template is denoted as

Manuscript received June 13, 2013; revised August 23, 2013

doi: 10.12720/ijsps.1.1.93-97

$$\chi = \{x_1, ..., x_T\} \quad (1)$$

and the reference template as

$$R = \{r_1, ..., r_k\} \quad (2)$$

Theory of vector quantization (VQ) [2] can be applied in template matching. The average quantization distortion of X, using R as the quantizer is defined as

$$D_Q(\chi, R) = \frac{1}{T}\sum_{t=1}^{T} \min_{1 \le k \le K} d(x_t, r_k) \quad (3)$$

where d (¢; ¢) is Euclidean distance as distance measure for vectors. Fig. 3 shows a conceptual diagram to illustrate these recognition processes, only two speakers and two dimensions of the acoustic space are shown. The circles refer to the acoustic vectors from the speaker 1 whereas the triangles are from the 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 a vector to the closest codeword of a codebook is called a VQ-distortion. In the recognition 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 distortion is identified as the speaker of the input utterance. One speaker can be discriminated from another based on the location of centroids [3].

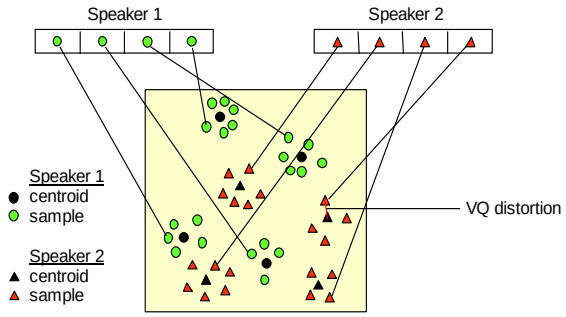

Figure 3. Conceptual diagram illustrating vector quantization codebook formation [3].

After the enrollment session, the acoustic vectors extracted from the input speech of each speaker provide a set of training vectors for that speaker. In Fig. 4, The LBG algorithm in a flow diagram is formally implemented by the following recursive procedure:

1. Design a 1-vector codebook; this is the centroid of the entire set of training vectors.
2. Double the size of the codebook by splitting each current codebook y_n as

$$y_n^+ = y_n(1+\varepsilon), y_n^- = y_n(1-\varepsilon) \quad (4)$$

where n varies from 1 to the current size of the codebook, and ε is a splitting parameter (choose ε =0.01).

3. Nearest-Neighbor Search: for each training vector, find the closest codeword in the current codebook, and assign it to the corresponding cell.
4. Centroid Update: update the codeword in each cell using the centroid of the training vectors assigned to that cell.
5. Repeat steps 3 and 4 until the average distance falls below a preset threshold.
6. Repeat steps 2, 3 and 4 until a codebook size of M is designed.

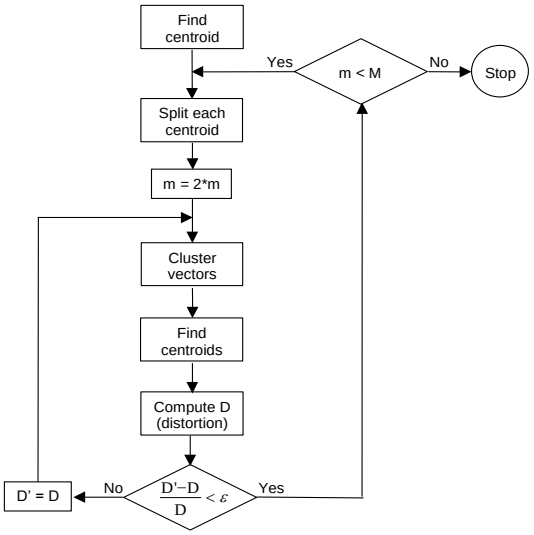

Figure 4. Flow diagram of the LBG algorithm.

III. Features Extraction

A. Criteria for Feature Selection

Feature extraction is necessary for several reasons. First, speech is a highly complex signal which carries several features mixed together [4]. In speaker recognition will be interested in the features that correlate with the physiological and behavioral characteristics of the speaker. Other information sources are considered as undesirable noise whose effect must be minimized. The second reason is a mathematical one, and relates to the phenomenon known as curse of dimensionality, which implies that the number of needed training vectors increases exponentially with the dimensionality.

The ideal feature should [5]:

- Have large Euclidean distance between-speaker and small distance within -speaker variability.
- Be difficult to impersonate/mimic.
- Not be affected by the speaker's health or long-term variations in voice.
- Occur frequently and naturally in speech.
- Be robust against noises and distortions.

B. Mel Frequency Cepstral Coefficients

The MFCC is a representation of the speech signal defined as the real cepstral of a windowed short-time signal derived from the FFT of that signal which, is first subjected to a log-based transform of the frequency axis (Mel-frequency scale)[6], and then use a modified Discrete Cosine Transform (DCT-II). Fig. 5 illustrates

the complete process to extract the MFCC vectors from the speech signal. It is to be emphasized that the process of MFCC extraction is applied over each frame of speech signal independently.

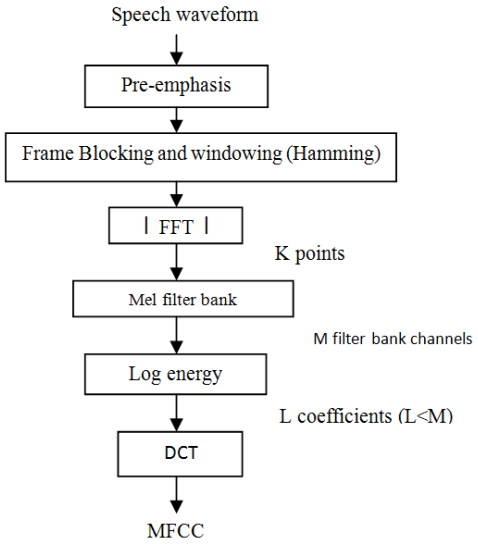

Figure 5. MFCC extraction process.

C. Perceptual Linear Prediction

Perceptual Linear Prediction (PLP) coefficients [7]. PLP analysis is based on LPC analysis incorporating a non-linear frequency scale and other psycho -physics properties of the human perception system.

LPC analysis is an effective method to estimate the main parameters of speech signals. The LPC coefficients are obtained for each frame independently. Fig. 6 illustrates the extraction process of the LPC coefficients.

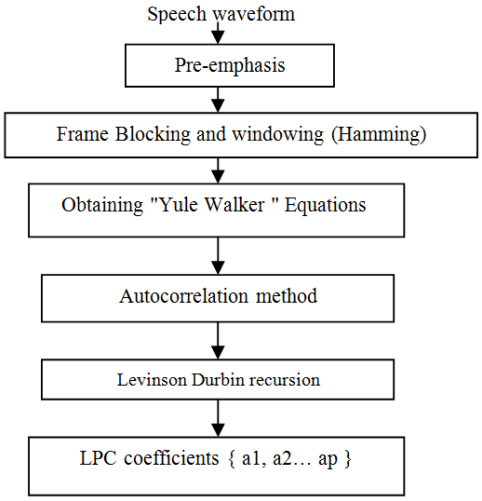

Figure 6. LPC coefficients extraction process.

PLP analysis is more similar to MFCC analysis, but the incorporation of more perceptual properties makes it more related to psycho-physical results. In Table I, the comparison between the properties of both methods can be seen.

The relative insensitivity of human hearing to slowly, varying stimuli may partially explain why human listeners do not seem to pay much attention to a slow change in the frequency characteristics of the communication environment or why steady background noise does not severely impair human speech communication. However, even when the experimental evidence from human perception may give us only limited support, the suppression of slowly varying components in the speech signal makes good engineering sense. Thus, to make speech analysis less sensitive to the slowly changing or steady-state factors in speech, a conventional critical-band short-term spectrum have been replaced

TABLE I. COMPARISON BETWEEN THE PROPERTIES OF MFCC AND PLP COEFFICIENTS

MFCC	PLP
Cepstrum – based spectral smoothing	LPC-based spectral smoothing
Pre-emphasis applied to speech waveform	Pre-emphasis applied to spectrum
Triangular Mel filter bank	Critical – band filter bank
Logarithmic amplitude compression	Cube root amplitude compression

D. Relative Spectral Technique - Perceptual Linear Predictive (RASTA - PLP)

The steps of RASTA-PLP [8] are as follows for each analysis frame, do the following.

1. Compute the critical-band power spectrum.
2. Transform spectral amplitude through a compressing static nonlinear transformation.
3. Filter the time trajectory of each transformed spectral component.
4. Transform the filtered speech representation through expanding static nonlinear transformation.
5. As in conventional PLP, multiply by the equal loudness curve and rise to the power 0.33 to simulate the power law of hearing.
6. Compute an all-pole model of the resulting spectrum, following the conventional PLP technique.

The key idea here is to suppress constant factors in each spectral component of the short-term auditory-like spectrum prior to the estimation of the all-pole model.

Speech is composed of excitation source and vocal tract system components. In order to analyze and model the excitation and system components of the speech independently, two Analysis methods are used.

- Spectral Analysis is very common for information to be encoded in sinusoids that form a signal. As well as those that has been created by humans. Many things oscillate in our universe. For example, speech is a result of vibration of the human vocal cords
- Cepstral analysis is to separate the speech into its source and system components without any prior knowledge about source and / or system.

IV. DATA SET AND EXPERIMENTAL RESULTS

Databases of 12 English words are spoken by 15 speakers as shown in Table II and have been used for

experiment of speaker identifications. The recorded Keywords have duration that varies from 60 ms to 14 seconds. The keywords are one, two, three, four, five, six, seven, eight, nine, ten, a half sentence {one, two, three, four, five} and a complete sentence {one, two, three, four, five, six, seven, eight, nine, ten}. One keyword is used for the train phase, whereas the remaining 11 keywords are used within the test phase, and the identification rate is calculated.

TABLE II. MORE SPEAKERS INFORMATION.

Ages	Male	Female
Child	1	1
Adult	11	2

Within the experiment, sampling rate is 44100 Hz, Hamming window is used, window duration is 25 ms with overlapping of 10 ms. Vector Quantization is used for the recognition, and number of centroids is 16. Various features are employed for the identification problem. Order in PLP Method is twelve and twenty filter banks are employed for MFCC calculation.

Table III shows the identification rate using each type of extracted feature. Best identification rate is obtained for PLP, whereas RASTA-PLP gives the worst identification rate. Fig. 7 shows the average identification rate for the first three speakers (S1, S2, S3), where the train keyword used is the complete sentence. As shown, PLP is speaker-sensitive method, but it is still the method that gives the highest identification rate.

TABLE III. RESULTS OF IDENTIFICATION RATE BY USING DIFFERENT FEATURE

Feature Type	Identification Rate
RASTA–PLP Spectral Analysis	20:46%
RASTA-PLP Cepstral Analysis	22:54%
MFCC	42:63%
PLP	52:73%

Fig. 8 shows the identification rate versus the train keywords used (e.g. train keywords One, Two, and Three). As shown, the identification rate is almost independent of the spoken keyword that is used within the train phase, but it depends only on the feature extraction method.

Fig. 7 and Fig. 8 show that PLP provides the highest identification rate.

V. CONCLUSIONS

Vector Quantization is used for speaker identification. The identification rate differs according to the speech features extracted. Various techniques are used for feature extraction such as RASTA-PLP, MFCC, and PLP. The features are employed for the identification problem within the train and test phases .The paper specifies PLP technique as it provides the highest identification rate.

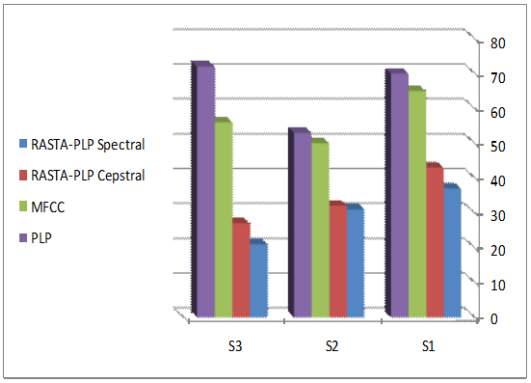

Figure 7. The Overall of the correctly identified for first three speakers using different feature.

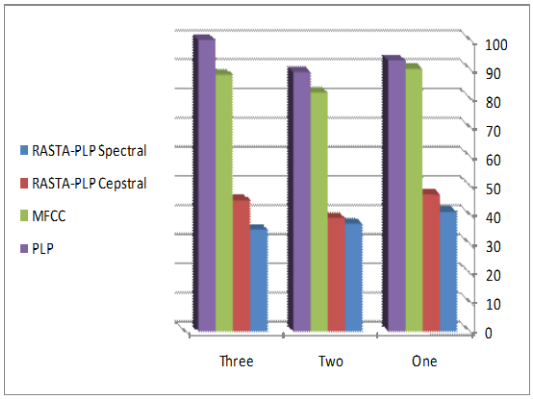

Figure 8. The overall of the correctly identified for the first train keywords using different feature.

REFERENCES

[1] P. Joseph and JR. Campbell, "Speaker recognition: A tutorial," *Invited Paper Proceedings of the IEEE*, vol. 85, no. 9, September 1997.

[2] A. Gersho and R. Gray, *Vector Quantization and Signal Compression*, Boston: Kluwer Academic Publishers, 1991.

[3] H. B. Kekre and V. Kulkarni, "Speaker identification by using Vector Quantization," *International Journal of Engineering Science and Technology*, vol. 2, no. 5, pp. 1325-1331, 2010.

[4] K. Saeed, "A speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 2, April 2007.

[5] T. Kinnunen and A. Haizhou, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Conmunication*, vol. 52, no. 1, pp. 12-40, January 2010.

[6] F. Bimbot, J. Francois, *et al*., "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal 1Processing*, vol. 4, pp. 430–451, 2004.

[7] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, April 1990.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech", *IEEE Transactions on Spech and Audio Processing*, vol. 2, no. 4, October I994.

Marwa T. Elkholy was born in Alexandria, Egypt, 1985. She received Bachelor Electronics and communications engineering (Graduated in July 2007) from Faculty of Engineering, Electrical Engineering Department, Communications and Electronics Section, Alexandria University. Her Graduation project: Satellite Remote Sensing and

Applications. MCIT 2011 Track (Graphics Design and Development). Graduation Project: 2D Plat-former Game. She is currently work as Computer Instructor, Database Administration, Web Developer in Electrical Engineering Department, Faculty of Engineering, Alexandria University. Her Fields of research include acoustics, communications, and programming.

Noha O. Korany was born in Alexandria, Egypt. She received Doctor of philosophy in Electrical Engineering, Alexandria, Egypt. She is currently an Assistant Professor at Electrical Engineering Department, Alexandria University, Alexandria, Egypt. Her main research fields are acoustics and communications.