MFCC and Similarity Measurements for Speaker Identification Systems

A. MAAZOUZI, N. AQILI, A. AAMOUD, M. RAJI, A. HAMMOUCH
LRGE Laboratory, ENSET
Mohammed V University in Rabat
Rabat, Morocco

Abstract— **Identity of a person via voice is one of the most interesting techniques used for user identification. Almost of speaker identification systems are based on distance computation or likelihood. Accuracy of identification process depends on: (*i*) the number of feature vectors, (*ii*) their dimensionality, and (*iii*) the number of speakers. This paper aims to develop a system able to identify a person from a sample of his speech. Recognition relies on a text-dependent system using English words as a password. Speech features are extracted using Mel Frequency Cepstral Coefficients (MFCCs). The recognition is based on discrete to continuous algorithm. Experimental results demonstrated that the proposed system return good accuracy rate.**

Keywords— Speaker identification; verification; MFCC; discrete to continuous algorithm.

I. Introduction

Speaker identification is one of the most complicated tasks in speech recognition problems. It generally depends on the production of speech of the speaker with physiological and behavioral characteristics. These characteristics rely on the speech generation (voice source) and the envelope behavior (vocal and nasal tract) [1].

Speaker recognition can be arranged to speaker identification and speaker verification [2]. Speaker identification consists in comparing a vocal message with a set of registered utterances corresponding to different speakers and determines the one who spoke. It entails a multi-choice classification problem. Speaker verification consists in accepting or rejecting the speaker who claims to be. It entails a yes-no hypothesis testing problem.

All speaker recognition systems contain two main phases: (*i*) feature extraction, and (*ii*) recognition. During the first step, a training vector is generated from the speech signal of the word (password) spoken by the user. These training vectors are stored in a database for subsequent use in the recognition phase. During the recognition phase, the system tries to identify the unknown speaker by comparing the extracted features from password with the ones from a set of known speakers.

The reminder of this paper is organized as follows. Section II presents conventional algorithms and the proposed system followed by section III that explains the processing applied on spoken passwords. The procedure of feature extraction is detailed in section IV. Section V explains the matching algorithm. Section VI presents the experimental results followed by conclusion.

II. Speaker Recognition System

A. Conventional Algorithm

Speaker identification and speaker verification are two aspects of speaker recognition. Speaker identification determines the person who spoke the given utterance amongst a given set of speakers. Meanwhile, speaker verification process consists of accepting or rejecting the identity claimed by a speaker. Speaker recognition methods is divided into text-dependent (fixed passwords) and text-independent (no specified passwords) methods. Text-dependent systems require less training than text- independent systems [3].

The identity of speaker can be represented by Linear Predictive Cepstrum Coefficients (LPCC), delta LPCC, MFCC, delta MFCC and pitch parameters [4]. Prosody is also a parameter that relies on the speaker characteristics [5] Gaussian Mixture Model (GMM) is considered as robust system for speaker recognition [6].

B. Developed System's Architecture

The combination of algorithms and techniques improves the accuracy or the recognition rate of many speech applications.

In the proposed speaker recognition system, the signal is preprocessed using endpoint detection algorithm and then we used the discrete wavelet transform to generate the approximation coefficients of the speech signal, we notice that we chose Daubechies (DAUB) family. In the features extraction step, the MFCC method is applied to these approximation coefficients to determine the speech features. Finally, the input signal is compared with the stored models using similarity measurements and then the most similar model is chosen using the discrete to continuous algorithm. Figure 1 shows the architecture of our proposed system based on English password recognition.

978-1-5386-1516-4/17/$31.00 ©2017 IEEE

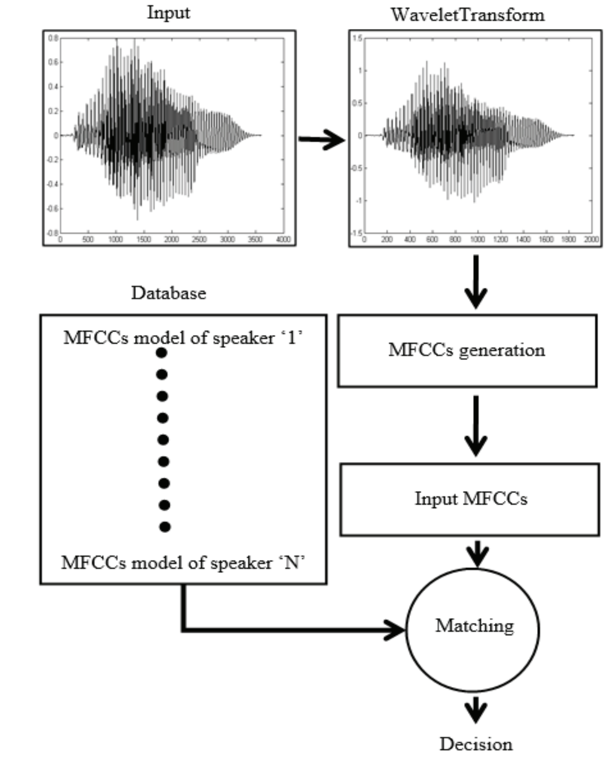

Fig. 1. Speaker identification system architecture

III. Signal processing

A. Signal Pre-processing

Speaker Pre-processing techniques of a signal are extremely important for an optimal recognition. Speech segments must be separated from non-speech to achieve good recognition accuracy in practical environment. The purpose of the endpoint detection algorithm [7] is to determine the beginning and the ending boundaries of each spoken word, and also to delete the noise region and silence. The algorithm relies on determining the energy level and zero-crossing rate. Figure 2 shows the endpoint detection for spoken digit "1" with its energy and zero-crossing rate.

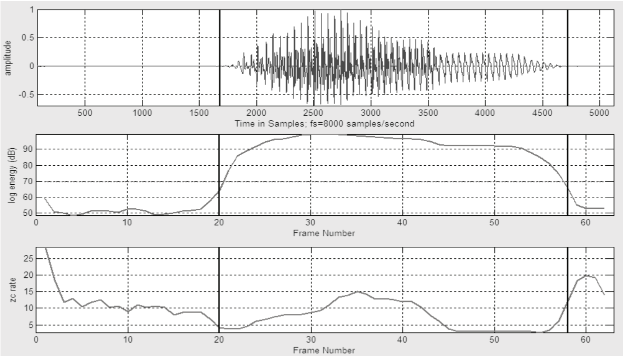

Fig. 2. Result of Endpoint Detection for spoken digit "1"

B. Wavelet Transform

Analyzing according to scale and translation is the prime idea behind the wavelet transform [8]. Wavelet decomposition satisfies certain mathematical requirements which are used to represent data or other functions. The Discrete Wavelet Transform (DWT) is a case of the Wavelet Transform for which the wavelet is sampled discretely. It provides a compact representation of a signal in time and frequency. The Discrete Wavelet Transform of a signal is computed by using a number of filters (Low-pass and high-pass filters). The filter outputs are then sub sampled by 2.

Different wavelet families have been presented such as Haar wavlet, Morlet wavelet, Daubechies (DAUB) wavelets, etc. In our work, we employed the wavelet family proposed by Daubechies [9].

IV. Features Extraction Using MFCC

Features extraction is the step of computing a sequence of feature vectors that characterize the word. They provide a compact representation of the given speech signal. MFCC is one of the most techniques used abundantly in automatic speech recognition systems (ASR). It is widely used for both speech and speaker recognition [10]. A Mel is a unit of measure based on human ear's perceived frequency. The Mel scale is approximately linear spacing below 1000Hz and a logarithmic spacing above 1000Hz. The Mel from frequency can be approximately expressed as:

$$mel(f) = 2595 \times \log(1 + f/700) \tag{1}$$

As illustrated in the block diagram below, a compact representation would be provided by a set of MFCC, which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale.

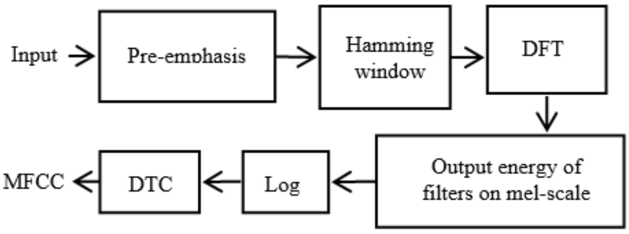

Fig. 3. MFCC computation

The calculation of MFCC is obtained by following contiguous steps: (*i*) Pre-emphasis**:** The pre-emphasis filter is used to improve the efficiency of the spectral analysis of the speech signal. (*ii*) Framing and windowing: The speech signal is usually segmented to frames, and the spectral and cepstral analysis is applied on each of these frames. Typically, the frame size is 25 milliseconds and the frames are overlapped by 10 milliseconds. After being separated into frames, each frame is multiplied by a window function before the spectral analysis to reduce the discontinuity introduced by framing. Windowing attenuates the values of the speech samples at the beginning and the ending of each frame. Typically, Hamming window is used. (*iii*) Spectral estimation: Fourier Transform

(FFT) algorithm is used to estimate the spectral coefficients of the speech frames. Only the magnitude of the spectral coefficients is retained and phase information is usually discarded. (*iv*) Mel filtering: To simulate the characteristics of human's ear, a set of triangular band pass filters are used to filter the spectrum of speech signal. It is utilized to model the human auditory system. (*v*) Logarithmic compression: deals with the loudness non-linearity. It approximately computes the relationship between the human's perception of loudness and the sound intensity. (*vi*) Discrete Cosine Transform (DCT): The cepstrum is defined as the inverse Fourier transform (IFFT) of the log magnitude of Fourier transform of the signal. Since the log Mel filter bank coefficients are real and symmetric, the IFFT operation becomes DCT to determine the cepstral coefficients.

V. Recognition Using Similarity Measurements

After a good feature extraction algorithm, the matching stage is a primary task in almost speaker identification system in term of reliability and also in term of time processing to achieve the right decision. In matching phase, we chose to implement the discrete to continuous algorithm [11] as a matcher algorithm. Where other algorithms try to find a good matching of some pairs of features by performing a comparison between the two features composition (point by point), the discrete to continuous algorithm traits this problem as point pattern matching problem. It brings the discrete representation of test signal onto the continuous representation (by interpolation) of the training dataset considering the issue in its entity and not point by point as is the case with the most algorithms dealing with this issue.

Let S_n and I_m represent the stored (training) template and the input (testing) one respectively, where m and n denote the number of features in S and I respectively. Both are extracted using MFCC. In our case $n = m$

1) The problem is to decide whether or not there is an allowed transformation T, such that: $T(I) \subset S$. This problem is knowing as finding the Largest Common Substructures(LCS).

The direct calculation of the transformation T has serious difficulties if it is performed on a discrete representation without prior knowledge of the correspondence. For this reason the discrete to continuous algorithm use an intermediate step in which the algorithm calculates a transformation T' that brings the discrete data (I) on the continuous representation of the model set (S).

Following, we describe the algorithm steps:

(*i*) Interpolation: Points (features) of S interpolated using a polynomial function P:

$$\forall (x, y) \;\; P(x_i) = y_i \tag{2}$$

Where x_i and y_i are the coordinates of ith element of S.

(*ii*) Search for T':

If the points of I are included in S, their transformation after applying T' must be included in the polynomial representation of S.

After transformation, we got $P(x_i') = y_i'$ where x_i' and y_i' are the coordinates of ith element of I.

To calculate the T' parameter the cost function QT is minimized as follows:

$$QT(ty) = \sum_{i=1}^{n} (P(x_j'') - y_j'')^2 \tag{3}$$

Where n is the number of the elements in I.

We notice that the algorithm is implemented using JAVA language programming.

VI. Simulations and Results

In the proposed speech recognition system, for feature extraction step, the MFCC are used to represent speech features. The MFCC parameters are presented in Table I.

TABLE I. Parameter Setting for MFCC

Parameter	**value**	**unit**
window length	0.025	s
step between successive windows	0.016	s
number of cepstra	13	-
pre-emphasis	0.97	-

45 speakers are tested. The password is spoken twice by each speaker and total 90 utterances are collected in our experiments. We have used 45 utterances for training and 45 utterances for testing. In all our experiments, the speech signals are sampled at 8 kHz. We propose to identify speakers based on their vocal password. The system differentiates between speakers even if they use the same password **"one".**

A. First Experiment

In this first experiment, we will measure the accuracy of our proposed system applied only for 10 speakers. Each speaker's password is compared with the 10 speakers (identification problem).

After calculating the distances between the different subjects, the minimal distance is used in decision step.

TABLE II. Identification of System's Users

Speaker (test)	Minimal Distance	Decision
Speaker 1	0.00100	Speaker 1
Speaker 2	0.00049	Speaker 2
Speaker 3	0.00043	Speaker 3
Speaker 4	0.00037	Speaker 4
Speaker 5	0.00089	Speaker 5
Speaker 6	0.00037	Speaker 6
Speaker 7	0.00022	Speaker 7
Speaker 8	0.00043	Speaker 8
Speaker 9	0.00033	Speaker 9
Speaker 10	0.00082	Speaker 10

B. Second Experiment

In this scenario, we will test the thoroughness of the developed algorithm for 45 speakers. Each speaker's password is compared with the 45 speakers. The following table shows the speakers that the system failed to identify.

TABLE III. CONFUSED USERS WITH THEIR DISTANCES

Speaker	Confused with	Distance
Speaker 15	Speaker 26	0.0011
Speaker 19	Speaker 16	0.000975
Speaker 29	Speaker 5	0.00093
Total missed: 2 out of 45		

The results illustrate that among 45 test samples; around 95% samples are correctly classified. To determine the threshold so that we forbid a false acceptance, we must find the minimum distance of false detection (it is equal to 0.0011 according to the table above). We can fix the threshold to 0.001**.** In the case of the speaker 15 and regarding the used threshold (0.001), we can consider that the system didn't fail in the identification process because the distance between the speakers 15 and 26 exceeds the threshold. Thus, the speaker will be claimed to spell the password again.

It is worthy to note that the recognition accuracy will be improved when we use a remarkable number of trials in training and testing phases.

VII. CONCLUSION

In this paper, a method based on cepstral analysis using MFCCs to extract speech features was developed. A word-dependent identification system was proposed. Based on the used database, the recognition rate exceeded 95%. We believe that in addition of the accuracy obtained by the system, the time running is also optimized by using the mean value of MFCCs. Also, more training data and good preprocessing methods can increasingly enhance the accuracy of the system.

REFERENCES

[1] D. Yu and L. Deng, *Automatic speech recognition: a deep learning approach*. London: Springer, 2015.

[2] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 33–41, Oct. 1994.

[3] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, Jan. 2010.

[4] Li Liu, Jialong He, and G. Palm, "Signal modeling for speaker identification," 1996, vol. 2, pp. 665–668.

[5] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust prosodic features for speaker identification," 1996, vol. 3, pp. 1800–1803.

[6] B. Xiang, U. V. Chaudhari, J. Navratil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time Gaussianization for robust speaker verification," 2002, p. I-681-I-684.

[7] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," *Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, Feb. 1975.

[8] H.-G. Stark, *Wavelets and Signal Processing: An Application-Based Introduction*. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg, 2005.

[9] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, pp. 909–996, Oct. 1988.

[10] S. Engelberg, *Digital signal processing: an experimental approach*. London: Springer, 2008.

[11] N. Aqili, M. Raji, A. Jilbab, S. Chaouki, and A. Hammouch, "PPM Translation, Rotation and Scale in D-Dimensional Space by the Discrete to Continuous Approach," *International Review on Computers and Software (IRECOS)*, vol. 11, no. 3, p. 270, Mar. 2016.