International Journal of Emerging Engineering Research and Technology
Volume 2, Issue 3, June 2014, PP 112-118
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Voice Recognition System for Machine Activation or Security Using MFCC

Lalita Palange[1], Sunita Sagat[2], Shashikant Hippragi[3], Raviraj Darekar[4]

[1]Department of E&TC, N.B. Navale College of engineering, Solapur, India
[2]Department of E&TC, N.B. Navale College of engineering, Solapur, India
[3]Department of E&TC, N.B. Navale College of engineering, Solapur, India
[4]Department of E&TC, A. G. Patil Institute of technology, Solapur, India

Abstract: *The aim of this thesis is to show the accuracy and time results of a text independent automatic speaker recognition (ASR) system, based on Mel-Frequency Cepstrum Coefficients (MFCC) and Gaussian Mixture Models (GMM), in order to develop a security control access gate. 450 speakers were randomly extracted from the audio data-base, their utterances have been improved using spectral subtraction, then MFCC were extracted and these coefficients were statistically analyzed by GMM in order to build each profile. For each speaker two different speech files were used: the first one to build the profile database, the second one to test the system performance. The accuracy achieved by the proposed approach is greater than 96% and the time spent for a single test run, implemented in Matlab language, is about 2 seconds on a common PC.*

1. INTRODUCTION

The speech signal of a person is unique and never changing. The signal taken as an input can be stored in template format and the stored templates can be compared with unknown or the input signal and exact match can be found out. The matched signal can be used for driving different mechanical machines, for machine activation in electronics or any industry and for determination of voice of unknown person for security purpose.

The concept of speech recognition which had poor result before few year is, now a days is emerged as a promising approach because of advancements in electronics developments such as digital signal processing, signal processors, coprocessors, Matlab software, separation of

Parameters and their analysis and uniqueness of speech signal for a person. Comparison of signal can also be done with using powerful C programming language or similar. The Matlab software can be used for separation of signal parameters efficiently. The concept of digital signal processing can be implemented successfully for parameter analysis.

The human voice is peculiar to each person and this is due to the anatomical apparatus of phonation. The vocal tract consists of three main cavities: the oral cavity, the nasal cavity and the pharyngeal cavity. The nasal cavity is essentially bony, hence static in time; further-more it can be isolated through the soft palate. The oral cavity is formed by the bony structure of the palate and soft palate; its conformation can be altered significantly by the movement of the jaw, lips and tongue. The pharyngeal cavity extends to the bottom of the throat and it can be compressed retracting the base of the tongue towards of the wall of the pharynx. In the lower part it ends with the vocal cords: a couple of fleshy membranes traversed by the air coming from the lungs. During the production of a sound, the space between the membranes (glottis) can be completely opened or partially closed. Due to the peculiarity of the voice formation apparatus, it can be possible to recognize a particular individual from its voice. Speaker recognition is classified as a hybrid biometric recognition approach, as it has two components: the physical one related to the anatomy of the vocal apparatus, and the behavioral component, pertinent to the mood of the speaker just in the recording moment.

There are several approach to ASR based on features, vector quantization, score normalization, pattern matching, etc, but the most of them are text dependent. Text independent ASR system based on Mel-Frequency Cepstrum Coefficients (MFCC) and Gaussian Mixture Models (GMM). Then the model parameters are estimated with the maxi-mum similarity making use of the Expectation and Maximization (EM) algorithm. The novel combination of these two techniques, allows the system to reach high recognition rates and high operative velocities, as shown in the following, allowing to use

the proposed system in real security context. In addition, unlike other works on ASR presented in literature, because the re-corded speaker signal could be corrupted by environ-mental additive noise, a spectral subtraction algorithm is also used. Comparisons with the state of the art demonstrate the effectiveness of the proposed approach in terms of accuracy rate. The data acquisition can be performed through simple microphones which are well spread and their cost is negligible. However cheap instrumentation may be more affected by disturbances such as background noise and the spectral subtraction algorithm could be no more sufficient for efficient noise suppression.

2. LITERATURE REVIEW

In the literature survey it is found that the results for different approach are as follows:

For Gaussian Mixer speaker model Accuracy rate is 96.80% with 49 Speakers, Iterative Clustering approach Accuracy rate is 91% with 50 Speakers, Fused MFCC & IMFCC feature based on Gaussian Filter Accuracy rate is 97.42% with 130 Speakers, Novel parametric Neural Network Accuracy rate is 94% with 40 Speakers, Performance Evaluation of Statistical Approach Accuracy rate is 94% with 40 Speakers, MFCC feature based on Gaussian Filter Accuracy rate is 97.98% with 450 speakers.

2.1 Applications of Speech Recognition:

Various applications of speech recognition domain as follows:

- Speech/Telephone/Communication Sector/Recognition
- Education Sector
- Domestic sector
- Military sector
- Physically Handicapped
- Medical sector
- Translation

3. PROBLEM STATEMENT

This study introduces a new method for speaker verification system by feature extraction method to improve the recognition accuracy and security.

Approach: The proposed system uses Mel frequency cepstral coefficients for speaker identification and Modified MFCC for verification.

Results: The proposed system was investigated the effect of the different length segmental feature as well as speaker modeling for speaker recognition. The performance is to be evaluated against 450 speakers for with duration of 2 sec.

Conclusion**:** Experimental results of the proposed system showed that higher recognition accuracy of 97.98% is achieved by increasing the number of filter banks used for feature extraction method, the system efficiency may further be improved using other speaker modeling techniques like Gaussian Mixer Model.

4. OBJECTIVES & SCOPE

The storing and comparing of the signal is used many times by old theories. But those techniques had implemented the concept by the analysis using frequency parameters or storing the complete recorded signal. As the amplitude, frequency parameters are not reliable for comparison purposes, results were so poor that it could not be implemented practically. Today the separation of speech signal parameters such as pitch , formants can give better results as these parameters are not changing time to time or due to change in atmospheric conditions. It is found that these methods i.e. use of DSP concept and software's like matlab can give results better than 95 % can be used efficiently in practical applications.

The proposed work includes the following activities -

The speech signal is used as stored reference signal and parameters like pitch formant are separated and analyzed using software techniques and they are designed in such a way that direct comparison can be made with the help of programming techniques. Speech Recognition can be dealt with Speaker

Identification & Speaker Verification. It may be the best Security System if used in combination with some other security system as human voice cannot be stolen or lost. It can be used in law court, Verification of rights, financial Transactions and Entrance control. Various Techniques can be implemented as Template matching, Vector Quantization, MFCC Cepstrum and Neural Network.

Traditional approaches had Limited Success due to noise, large No. of speakers, unlimited text etc. Any System extract features like Cepstrum, power spectrum, LPC, pitch, formants & process them. LPC & Cepstrum arc found to give best Performance. It may also be accomplished with use of single digits instead of sentences. Inputs may be considered in the form of Power Spectrum, LPC, Mel Scaled Cepstrum, Reflection Coefficients, Auto Correlation Coefficients and Mel scaled power spectrum etc. Of the above Power spectrum is found to be best as it is efficient having lower error rate, least response time.

The analysis of signals can be done for this purpose using software like Matlab. Preference method for the separation of parameters by signal processing approach. Pre-processing deals with taking of input signal and storing. Post processing involves the storing of parameters in ready format for comparison. Stored samples compared with input, unknown signal. Use of the recognized signal can be implemented for activation of desired purpose like machine activation, security system or similar.

5. Methodology

A sensor which makes acquisition of data and its subsequent sampling: in the specific case the sensor is a microphone, possibly with a high Signal to Ratio (SNR) value. Since the input signal is essentially speech, the sampling rate is usually set to 8 kHz. A step of preprocessing that in the voice context is constituted by the signal cleaning: simply de -noising algorithm can be applied to recorded data after a normalization procedure. In order to clean recorded speech signal from environmental additive noise, a spectral subtraction algorithm.

The extraction of the peculiar characteristics (feature extraction): in this stage Mel frequency cepstral coefficients are evaluated using a Mel filter bank after a transformation of the frequency axis in a logarithmic one. The generation of a specific template for each speaker: in this work we have decided to use the Gaussian Mixture Models (GMM) where model parameters are estimated with the maximum similarity making use of the Expectation and Maximization (EM) algorithm. In case of the user is registering (enrollment) for the first time to the system, this template will be added to the database, using some database programming techniques.

Otherwise, in case of test among users already present in the database, a comparison (matcher) determines which profile matches the generated template of the test speech. The matcher utilizes a similarity test, obtaining by a ratio value that can be accepted if it is higher than a decision threshold. The typical ASR system is shown in Figure 1. The technologies used for the development of the biometric system are the MMFCC for the extraction of the characteristics and the GMM for the statistical analysis of the data obtained, for the templates generation and for the comparison.

6. Mel Frequency Cepstral Coefficient

The term "Cepstrum" is a pun where the first letters of the term "spectrum" are reversed. It was described in 1963 by Bogert *et al*. Cepstrum is defined as the inverse Fourier transform of the logarithm of the spectrum of a signal,

$$x_c(n) = DFT^{-1}\left\{\log\left|DFT\left\{x(n)\right\}\right|\right\} \qquad (1)$$

The Cepstrum transform the signal from the frequency domain into the quefrency domain.

When Cepstrum is applied to the voice, its strength is to be able to divide excitation and transfer function. In a signal $y(n)$ based on the source-filter model, in this specific context, respectively the vocal cords and the vocal tract, Cepstrum allows separation in $y(n)=x(n)*h(n)$, where the source $x(n)$ passes through a filter described by the impulse response $h(n)$. The spectrum of $y(n)$ obtained by the Fourier transform is $y(k)=X(k)\,H(k)$, where k index of discrete frequencies, *i.e*. the product of two spectra, respectively the source and the filter one. Separating these two spectra is complicated. On the contrary, it is possible to separate the real envelope of the filter from the remaining spectrum by formulating all the phase at the beginning. The Cepstrum is based on the properties of the logarithm that can transform the product of the argument in sums of logarithms. Starting from the logarithm of the modulus of the spectrum:

$$\log|Y(k)| = \log\left(|X(k)H(k)|\right) = \log\left(X(k)\right) + \log\left(H(k)\right) \quad (2)$$

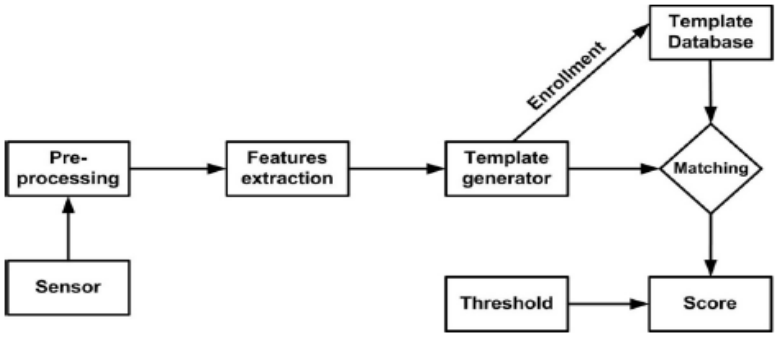


Figure 1. *A typical ASR system.*

It is possible to separate the fast oscillating component from the slow one, respectively by means of a high and low pass filter, obtaining:

$$\begin{aligned} c(n) &= DFT^{-1}\left(\log|Y(k)|\right) \\ &= DFT^{-1}\left(\log|X(k)|\right) + DFT^{-1}\left(\log|H(k)|\right) \end{aligned} \quad (3)$$

That is the signal Cepstrum in the quefrency domain. In the low quefrencies are described the transfer function information, in the high quefrencies there is data about excitation. Hence the initial wave of percussion created by the vocal cords and shaped by the throat, nose and mouth can be analyzed as a sum of a source function (given by the excitation of the vocal cords) and a filter (throat, nose, mouth). The separation between high and low quefrency, can be obtained by a high pass lifter (filter) for the fast oscillation and a low pass lifter for the slow one. Psychoacoustic studies have shown that the mind perception of the frequency content of the sound follows a nearly logarithmic scale, the Mel scale, which is linear up to 1 kHz and logarithmic there after:

$$\text{mel}(f) = \begin{cases} f \text{ if } f \le 1 \text{ kHz} \\ 2595 \log\left(1 + \dfrac{f}{7000}\right) \text{ if } f > 1 \text{ kHz} \end{cases}$$

The Mel scale is shown in Figure 2, where it is clear the compression of the Mel scale (reported in *y*-axis) with respect the Hertz scale (in *x*-axis) for frequencies greater than 1 kHz. In this scale pitches are judged by listeners to be equal in distance from one another. Mel-Cepstrum estimates the spectral envelope of the output of the filter bank. Let *Yn* represent the logarithm of the output energy from channel n, applying the discrete cosine transform (DCT) we obtain the cepstral coefficients MFCC through the equation:

$$c_k = \sum_{n=1}^{N} Y_n \cos\left[k\left(n - \frac{1}{2}\right)\frac{\pi}{N}\right] \forall k = 0, \cdots, K \quad (4)$$

The simplified spectral envelope is rebuilt with the

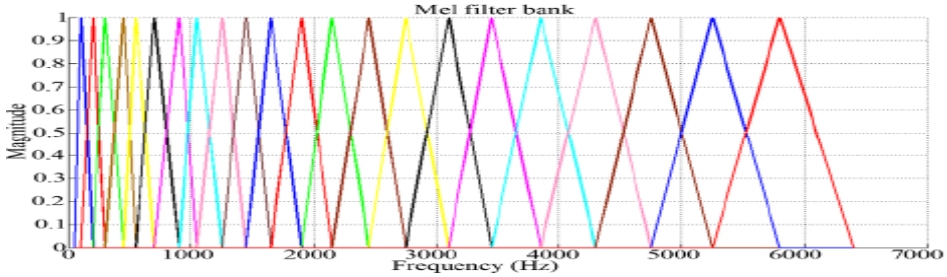


Figure 2. *Mel filter bank.*

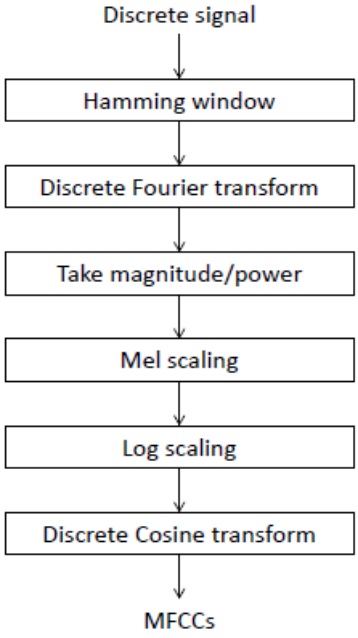
first *Km* coefficients, with *Km* < *K*:

$$C(\text{mel}) = \sum_{k=1}^{K_m} c_k \cos\left(2\pi k \frac{\text{mel}}{B_m}\right) \qquad (5)$$

where *Bm* is the bandwidth analyzed in Mel domain and *Km* = 20 is a typical value assumed by *Km*. *c*0 is the mean value in d*B* of the energy of the filter bank channels, hence it is in direct relation with the energy of the sound and it can be used for the estimation of the energy. Schematically, the coefficients are derived in the following way: the spectrum of the original signal is computed with the Fourier transform; the obtained spectrum is mapped in Mel making use of appropriate overlapping windows; for each obtained function the logarithm is calculated; the discrete cosine transform is calculated (DCT); the coefficients are the amplitudes of the resulting spectrum. In order to emphasize the low quefrencies DCT is chosen.

7. MEL-FREQUENCY CEPSTRUM COEFFICIENTS PROCESS

The speech input is typically recorded at a sampling rate above 10000 Hz. This sampling frequency was chosen to minimize the effects of aliasing in the analog-to-digital conversion. These sampled signals can capture all frequencies up to 5 kHz, which cover most energy of sounds that are generated by humans. The main purpose of the MFCC processor is to mimic the behavior of the human ears. In addition, rather than the speech waveforms themselves, MFFC's are shown to be less susceptible to mentioned variations.



Steps for computing MFCCs of a discrete signal

MFCC is chosen for the following reasons

1. MFCC is the most important features, which are required among various kinds of speech applications.

2. It gives high accuracy results for clean speech.

3. MFCC can be regarded as the "standard" features in speaker as well as speech recognition.

4. In the MFCC frequency bands are positioned logarithmically which approximates the human auditory systems response more closely than the linearly spaced frequency bands of FFT or DCT

Gaussian Mixture Model

Each arbitrary probability density function (pdf) can be approximated by a linear combination of unimodal Gaussian density. Under this assumption, Gaussian mixture models have been applied to model the distribution of a sequence of vectors $X = x_1, x_2, \cdots, x_t, \cdots, X_T$ each one of dimension *D*, containing data on the characteristics extracted from the voice of a subject, according to:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i p_i(x_t) \quad (6)$$

$$p(X_t|\lambda) = \prod_{t=1}^{T} p(s > t|\theta) \quad (7)$$

Where *wi* are the weights of the corresponding mixtures to the unimodal Gaussian densities *pi* with *i=1, 2,…….M* and:

$$p_i(x_t) = \left(\frac{1}{\sqrt[D]{2\pi}\sqrt{\det(\Sigma_i)}} \right)^{-\frac{1}{2}\left((x_t-\mu_i)^T \Sigma^{-1}(x_t-\mu_i)\right)} \quad (8)$$

The weights of the mixtures satisfy the constraint:

$$\sum_{i=1}^{M} w_i = \quad (9)$$

Each speaker is identified by a λ model obtained from GMM analysis. In particular lambda is defined as:

$$\lambda = \{w_i, \mu_i, \Sigma_i\} \quad (10)$$

where μ_i is the mean vector and is the covariance matrix Σ_i

Given a characteristic vector sequence of the speaker to be identified, the model parameters are estimated with the maximum similarity λ making use of the Expectation and Maximization algorithm. The λ model is compared with a characteristic vector *X* by calculating the log-likelihood similarity. In order to decide, it is utilized a similarity test obtained by the following ratio:

$$\frac{P(X|\text{ Speaker})}{P(X|\text{ Other Speaker})} > \sigma \quad (12)$$

where σ is the dec on the contrary, a collection of models of different speakers. The final score of a certain subject S_c over an *X* vector containing the voice features of the test is given by:

$$\log L(x) = \log p(X|S = S_c) - \log \sum_{S \in pop} p(X|S \neq S_c) \quad (13)$$

where L(X) represents the similarity value of *X* vector with respect to *c* compared with the characteristics of other individuals in the database (pop), excluding the one taken into account.

8. CONCLUSION

A new speaker recognition scheme is proposed and the proposed system uses MFCC features for identification. The proposed system is suitable for highly secured environments, because of zero false rejection rates. Even with this high population the system performed well since it has produced comparatively good performance than the existing algorithms. The proposed system can be extended to multilingual text independent speaker recognition system.

In this paper we introduced an Automatic Speaker Recognition system based on MFCC and GMM. The accuracy of the proposed system is to be greater than 96% and with 450 speakers.

A high recognition rate on a wide number of subjects, together with a high operative velocity, make it useful for real security access control applications.

REFERENCES

[1] S. Chakroborty and G. Saha, "Improved Text-Independent Speaker Identification Using Fused MFCC and IMFCC feature Sets Based on Gaussian Filter," International Journal of Signal Processing, Vol. 5, No. 1, 2009, pp. 11-19.

[2] Alfredo Maesa, Fabio Garzia, Michele Scarpiniti, Roberto Cusani, "Improved Text-Independent Speaker Identification Using MFCC feature Sets & Gaussian Mixer Model," Journal of information security, 2012.

[3] A. Revathi, R. Ganapathy and Y. Venkataramani, "Text Independent Speaker Recognition and Speaker Independent Speech Recognition Using Iterative Clustering Approach," International Journal of Computer Science & Information Technology, Vol. 1, No. 2, 2009, pp. 30-42.

[4] "Voxforge Database." www.voxforge.org

[5] Nguyen Viet Cuong, Vu Dinh, Lam Si Tung Ho, "Mel-frequency Cepstral Coefficients for Eye Movement Identification", IEEE 2012.

[6] Shivanker Dev Dhingra, Geeta Nijhawan , Poonam Pandit, "ISOLATED SPEECH RECOGNITION USING MFCC AND DTW" , International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, Aug 2013