RESEARCH ARTICLE OPEN ACCESS

Speaker Recognition Using MFCC and GMM with EM

Apurva Adikane, Minal Moon, Pooja Dehankar, Shraddha Borkar, Sandip Desai
Department of Electronics and Telecommunications, Yeshwantrao Chavan College of Engineering
Hingna Road, Wanadongri, Nagpur-441110
appuadikane@gmail.com
gudiyamoon18@gmail.com
pooh.dehankar09@gmail.com
shraddhab15@rediffmail.com
sad.ycce@gmail.com

ABSTRACT
This paper aims at showing the accuracy of a text dependent speaker recognition system using mel frequency cepstrum coefficient (MFCC) and Gaussian Mixture Model (GMM) accompanied by Expectation and Maximization algorithm (EM). The goal of speaker recognition is to determine which one of a group of known speakers best matches the input voice samples. Voice samples were taken, MFCC were extracted, and these coefficients were statistically analyzed by GMM in order to build each profile. We concentrate on the task of improving the performance of Gaussian Mixture Models for speaker identification by using Expectation-Maximization (EM) methods. These methods are together implemented resulting in decreased error rates.
Keywords **-** Automatic speaker recognition, access control, authentication, feature extraction, Gaussian Mixture Model (GMM), speaker verification.

I. Introduction

Speech is the primary means of communication between humans. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks which necessitate human-machine interactions, research in automatic speech and speaker recognition by machines has attracted a great deal of attention for five decades. Based on major advances in statistical modeling of speech, automatic speech recognition systems today find widespread application in tasks that require human-machine interface, such as automatic call processing in telephone networks and query-based information systems that provide updated travel information, stock price quotations, weather reports, etc.

The development of speaker recognition began in early 1960's. In 1960's the Bell Labs built experimental systems aimed to work over dialed-up telephone lines [1]. Text dependent and independent methods began to develop. In 1980's, speaker recognition systems based on Hidden Marcov Model (HMM) architecture were developed. Also Vector Quantization (VQ) algorithms were implemented along with HMM. In 1990's, research on increasing robustness became a central theme. Text prompted methods and score normalization were also developed. In 2000's, new score normalization methods were developed and high level features such as word idiolect, pronunciation, phone usage, prosody, etc have been successfully used in text independent speaker recognition systems.

Speech conveys several levels of information. On a primary level, speech conveys the words or message being spoken, but on a secondary level, speech also reveals information about the speaker. Given a speech signal there are two kinds of information that may be extracted from it. On one hand there is the linguistic information about what is being said, and on the other there is also speaker specific information. This report deals with the task of speaker recognition where the goal is to determine which one of a group known speakers best matches the input voice sample.

Given a speech sample, speaker recognition is concerned with extracting clues to the identity of the person who was the source of that utterance. Speaker recognition is divided into two specific tasks: verification and identification. In speaker verification the goal is to determine from a voice sample if a person is whom he or she claims. In speaker identification the goal is to determine which one of a group of known voices best matches the input voice sample. In either case the speech can be constrained to a known phrase (text-dependent) or totally unconstrained (text-independent) [2,3].

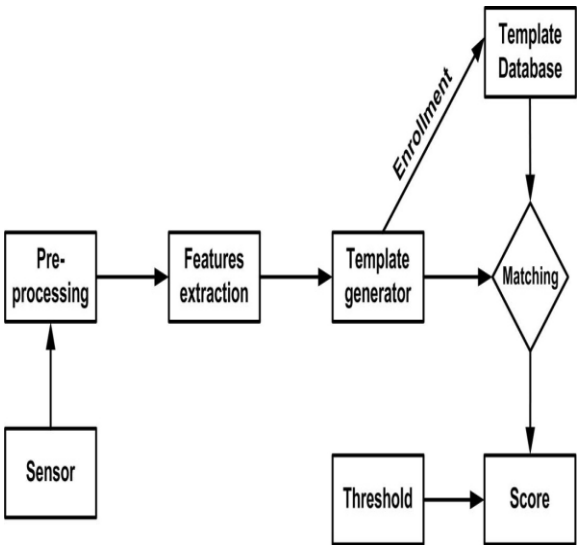


Fig (1): speaker recognition system

There are many algorithms and models that can be used for speaker recognition including Neural Networks, unimodal Gaussians, Vector Quantization, Radial Basis Functions, Hidden Markov Models and Gaussian Mixture Models (GMMs). These perform well under clean speech conditions, but in many cases performance degrades when test utterances are corrupted by noise, mismatched conditions or if there are only small amounts of training and testing data. Among these methods GMMs are usually preferred because they offer high classification accuracy while still being robust to corruptions in the speech signal.

In this paper, we propose text dependent 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.

The paper is organized as follows: Section 2 describes the feature extraction and introduces the MFCC technique, while Section 3 introduces the GMM models and Expectation and Maximization algorithm. Finally Section 4 concludes the work.

II. Feature Extraction

Feature extraction is the process that extracts a small amount of data from the voice signal that can later be used to represent each speaker.

The purpose of this module is to convert the speech waveform, using digital signal processing (DSP) tools to a set of features (at a considering lower information rate) for further analysis. This is often referred as the signal-processing front end.

There are a number of different speech features that have been shown to be indicative of speaker identity. These include pitch related features, Linear Prediction Cepstral Coefficients (LPCCs) and Maximum Autocorrelation Value (MACV) features. Although there are no exclusively speaker distinguishing features, the speech spectrum has been shown to be very effective for speaker recognition. Here we use Mel Frequency Cepstral Coefficients (MFCCs) extracted from the spectrum. The main reason for this is that in many applications speaker identification is a precursor to further speech processing, especially speech recognition, to identify what is being said. Among the possible features MFCCs have proved to be the most successful and robust features for speech recognition. So, to limit computation in a possible application, it makes sense to use the same features for speaker recognition.

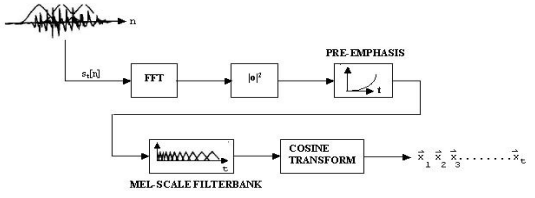
Figure 2.1 shows the block diagram of the procedure used for feature extraction in the front end. The speech signal is divided into 30 msec long segments overlapping by 15 msec using a Hamming window. The magnitude spectrum of this short time segment is passed through a simulated mel-scale filter bank consisting of 30 filters. The filter bank is similar to the one described in. The log of the output energy of each filter is calculated and collected into a vector. This is then cosine transformed into cepstral coefficients. The cepstral coefficients are truncated to obtain MFCCs.

2.1. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition [4].

We will give a high level intro to the implementation steps, then go in depth why we do the things we do. Towards the end we will go into a more detailed description of how to calculate MFCCs.

1. Frame the signal into short frames.
2. For each frame calculate the periodogram estimate of the power spectrum.
3. Apply the mel filter bank to the power spectra, sum the energy in each filter
4. Take the logarithm of all filter-bank energies.
5. Take the DCT of the log filter-bank energies.
6. Keep DCT coefficients 2-13, discard the rest.



The term "cepstrum" is a pun where the first letters of the term "spectrum" are reversed. Cepstrum is defined as in verse fourier transform of the logarithm of the spectrum of the signal.

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

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:

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

Mel-cepstrum estimates the spectral envelope of the output of the filter bank. Let Y_n 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:

$$c_k = \sum_{n=1}^{N} Y_n cos\left[k\left(n-\frac{1}{2}\right)\frac{\pi}{N}\right] €k = 0, \ldots., K \quad (3)$$

The simplified spectral envelope is rebuilt with the first k_m coefficients with k_m< k;

$$C(mel) = \sum_{k=1}^{K_m} c_k \cos\left(2\pi k \frac{mel}{B_m}\right) \quad (4)$$

where B_m is the bandwidth analyzed in Mel domain and *Km* = 20 is a typical value assumed by K_m. c_k is the mean value in dB 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.

III. Gaussian Mixture Models (GMM)

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities [4,5]. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm [6].

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 vectors $X = x_1,\ldots..,\ x_l,\ldots..$, each one of dimension D, containing data on the characteristics extracted from the voice of the subject, according to:

$$p(x_t|\lambda) = \sum_{i=1}^{M} w_i p_i(x_t) \quad (5)$$

$$p(X_t|\lambda) = \prod_{t=1}^{T} p(s > t|\theta) \quad (6)$$

where are the weights of the corresponding mixtures to the unimodal Gaussian densities with i=1, ... ,M and:

$$p_i(x_t) = \left(\frac{1}{\sqrt[D]{2\pi}\sqrt{\det(\Sigma_i)}}\right)^{\frac{-1}{2}\left((x_t-\mu_i)^T\Sigma^{-1}(x_t-\mu_i)\right)} \quad (7)$$

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

$$\lambda = \{w_i, \mu_i, \Sigma_i\}$$

Given a characteristic vector sequence of the speaker to be defined, 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:

$$\log P(X|\lambda) = \sum_{T} \log P(x_t|\lambda) \quad (8)$$

In order to decide, a similarity test is utilized obtained by the following ratio:

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

The final score of a certain subject over an vector X containing the voice features of the test is given by,

$$\log L(x) = \log p(X|S = S_c) - log \sum_{Sepop} p(X|S \neq S_c) \quad (9)$$

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

3.1. The EM algorithm for GMM

We define the EM (Expectation-Maximization) [7] algorithm for Gaussian mixtures as follows. The algorithm is an iterative algorithm that starts from some initial estimate of Θ (e.g., random), and then proceeds to iteratively update until Θ convergence is detected. Each iteration consists of an E-step and an M-step.

E-Step: Denote the current parameter values as Θ.

$$w_{ik} = p(z_{ik} = 1|x_i, \Theta) = \frac{p_k(x_i|z_k, \theta_k).\alpha_k}{\sum_{m=1}^{K} p_m(x_i|z_m, \theta_m).\alpha_m}$$

Compute w_{ik} $(1 \le k \le K, 1 \le i \le N)$ for all data points x_i ; $1 \le i \le N$ and all mixture components $1 \le k \le K$. Note that for each data point x_i, the membership weights are defined such that $\Sigma_{k=1}^{K} w_{ik} = 1$. This yields an N × K matrix of membership weights, where each of the rows sum to 1.

where,

- x_i is a d-dimensional vector measurements.
- $p_x(x|z_k, \theta_k)$ are mixture components, $1 \le k \le K$.
- $z=(z_1, \ldots, z_K)$ is a vector of K binary indicator variable that are mutually exclusive and exhaustive.
- $\alpha_k = p(z_k)$ are the mixture weights, representing the probability that a random selected x was generated by component k, where $\sum_{k=1}^{K} \alpha_k = 1$.

The complete set of parameters for a mixture model with K components is,

$$\Theta = \{\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K\}$$

M-Step: Now use the membership weights and the data to calculate new parameter values. Let $N_k = \Sigma_{i=1}^{N} w_{ik}$, i.e., the sum of the membership weights for the k^{th} component—this is the effective number of data points assigned to component k.

Specifically,

$$\alpha_k^{new} = \frac{N_k}{N}, \ 1 \le k \le K.$$

These are the new mixture weights.

$$\mu_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \, . x_i, \quad 1 \le k \le K. \qquad (10)$$

The updated mean is calculated in a manner similar to how we could compute a standard empirical average, except that the i^{th} data vector xi has a fractional weight w_{ik}. Note that this is a vector equation since μ_k^{new} and x_i are both d-dimensional vectors.

$$\mu_k^{new} = \left(\frac{1}{N_k}\right) \sum_{i=1}^{N} w_{ik} \, . (x_i - \mu_k^{new})(x_i - \mu_k^{new})^t, \quad 1 \le k \le K \qquad (11)$$

Again we get an equation that is similar in form to how we would normally compute an empirical covariance matrix, except that the contribution of each data point is weighted by w_{ik}. Note that this is a matrix equation of dimensionality d ×d on each side.

The equations in the M-step need to be computed in this order, i.e., first compute the K new α's, then the K new $\mu_k's$, and finally the K new $\Sigma_k's$.

After we have computed all of the new parameters, the M-step is complete and we can now go back and recompute the membership weights in the E-step, then recompute the parameters again in the E-step, and continue updating the parameters in this manner. Each pair of E and M steps is considered to be one iteration.

IV. Conclusion

Automatic speaker recognition is the use of a machine to recognize a person from a spoken phrase. Speaker recognition systems can be used in two modes: to identify a particular person or to verify a person's claimed identity. The scope of this work is limited to speech collected from cooperative users in real world office environments and without adverse microphone or channel impairments.

The use of EM algorithm along with GMM and MFCC improves the system performance.

V. Acknowledgement

We are thankful to our project guide Mr. S.A. Desai, Lecturer, Electronics and Telecommunications, YCCE for his continuous guidance and support for this project.

References

[1] Sadaoki Furui, *50 years of progress in speech and speaker recognition*, Department of Computer Science, Tokyo Institue of Technology.

[2] Samudravijaya K, *Speech and Speaker Recognition: A Tutorial*, Tata Institute of Fundamental Research, Homi Bhabha Road, Mumbai.

[3] D. A. Reynolds, *An Overview of Automatic Speaker Recognition Technology*, *Acoustics, Speech and Signal Processing* (*ICASSP*), 2002, pp. 4072-4075.

[4] Alfredo Maesa, Fabio Garzia, Michele Scarpiniti, Roberto Cusani, *Text Independent Automatic Speaker Recognitio System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models*, Journal of Information Security, 2012, 3, 335-340, http://dx.doi.org/10.4236/jis.2012.34041 Published Online October 2012 (http://www.SciRP.org/journal/jis)

[5] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, *Speaker Verification Using Adapted Gaussian Mixture Models*, *Digital Signal Processing*, Vol. 10, No. 2, 2000, pp. 19- 41.

[6] R. Reynolds, *Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models*, *IEEE Transactions on Speech and Audio Processing*, 1995, pp. 72-83.

[7] *The EM Algorithm for Gaussian Mixtures*, Probabilistic Learning: Theory and Algorithms, CS274A: http://www.ccs.neu.edu/home/jaa/CS6140.13F/Homeworks/HW05/8-em.pdf