TEXT-INDEPENDENT SPEAKER VERIFICATION: A COMPARATIVE ANALYSIS STUDY

Mohamed SOLTANE, Noureddine DOGHMANE and Noureddine GUERSI

Electronics Department, Faculty of engineering science
Bajdi Mokhtar University of Annaba, 23000 Annaba - Algeria
soltane@lri-annaba.net & xor99@hotmail.com, ndoghmane@univ-annaba.org, guersi54@yahoo.fr

Abstract: Gaussian mixture models (GMMs) remain the state of the art technique for modeling spectral envelope features for speech recognition systems. This paper presents a comparative analysis of the performance of three estimation algorithms Expectation Maximization (EM), Greedy EM Algorithm (GEM) and Figureiredo-Jain Algorithm (FJ) based Gaussian mixture models (GMMs) for text-independent speech biometrics verification. The simulation results are showed significant performance achievements. The test performance of, EER=0.26 % for "EM", EER=0.21 % for "GEM" and EER=0.16 % for "FJ", show that the behavioral information scheme of speech biometrics is more robust and have a discriminating power, which can be explored for identity authentication.

Key words: Biometric authentication, behavioral, speech, soft decision and Gaussian Mixture Modal, EM, GEM and FJ.

1. INTRODUCTION

BIOMETRIC is a Greek composite word stemming from the synthesis of bio and metric, meaning life measurement. In this context, the science of biometrics is concerned with the accurate measurement of unique biological characteristics of an individual in order to securely identify them to a computer or other electronic system. Biological characteristics measured usually include fingerprints, voice patterns, retinal and iris scans, face patterns, and even the chemical composition of an individual's DNA [1]. Biometrics authentication (BA) (Am I whom I claim I am?) involves confirming or denying a person's claimed identity based on his/her physiological or behavioral characteristics [2]. BA is becoming an important alternative to traditional authentication methods such as keys ("something one has", i.e., by possession) or PIN numbers ("something one knows", i.e., by knowledge) because it is essentially “who one is”, i.e., by biometric information. Therefore, it is not susceptible to misplacement or forgetfulness [3]. These biometric systems for personal authentication and identification are based upon physiological or behavioral features which are typically distinctive, although time varying, such as fingerprints, hand geometry, face, voice, lip movement, gait, and iris patterns. An identity verification system has to deal with two kinds of events: either the person claiming a given identity is the one who he claims to be (in which case, he is called a client), or he is not (in which case, he is called an impostor). Moreover, the system may generally take two decisions: either accept the client or reject him and decide he is an impostor.

Some works based on biometric speech identity verification systems has been reported in literature. Fortuna J. et al. [16] presents a comparative analysis of the performance of decoupled and adapted Gaussian mixture models (GMMs) for open-set, text-independent speaker identification (OSTISI) and concluded that the speaker identification performance is noticeably better with adapted-GMMs than with decoupled-GMMs. Their analysis is based on a set of experiments using an appropriate subset of the NIST-SRE 2003 database and various score normalization methods. They included a detailed description of the experiments and discuss how the OSTI-SI performance is influenced by the characteristics of each of the two modeling techniques and the normalization approaches adopted. Eduardo Sánchez-Soto et al. [17] present a new adaptation technique for speaker verification of models built using Bayesian Networks tested using the NIST 2002 data base and showed improvement in the verification performances. The adaptation problem of parameters of the conditional probability tables (CPTs) is treated in a specific manner. The model adaptation involves estimating the new vectors with a transformation that includes vectors in the world model and the speaker model and the combination of both models is based on a value computed using a measure of distance between vectors of both CPTs. Arnon Cohen et al. [18] describes an HMM based speaker verification system evaluated on a text-dependent database, which verifies speakers in their own specific feature space. The user feature space is determined by a Dynamic Programming (DP) feature selection algorithm, in which a suitable criterion, correlated with Equal Error Rate (EER) was developed and is used for this feature selection algorithm. A significant improvement in verification results was demonstrated with the DP selected individual feature space. An EER of 4.8% was achieved when the feature set was the “almost standard” Mel Frequency Cepstrum Coefficients (MFCC) space (12 MFCC + 12 ∆MFCC). Under the same conditions, a system based on the selected feature space yielded an EER of only 2.7%. Mijail Arcienega et al. [19] present a Bayesian network approach for modeling the pitch and spectral envelope and showed an increase in the performance of the speaker recognition system. In which the conditional statistical distributions (represented by GMMs) of the features are

Manuscript received February 11, 2010; revised March 21, 2010
simultaneously exploited for increasing the recognition score within the approach, and in particularly in noisy conditions. Driss Matrouf et al. [20] investigates the effect of voice transformation on automatic speaker recognition systems performance and showed an increase of about 2.7 time of the likelihood ratio, without a degradation of the natural aspect of the voice. It focuses on increasing the impostor acceptance rate, by modifying the voice of an impostor in order to target a specific speaker. Their work is inspired from the idea that in several forensic situations, it is reasonable to think that some organizations have a knowledge on the speaker recognition method used by the police department and could impersonate a given, well known speaker.

II. BIOMETRIC SPEECH VERIFICATION

1. Speech Analysis and Feature Extraction

Gaussian Mixture Models (GMMs), is the main tool used in text-independent speaker verification, in which can be trained using the Expectation Maximization (EM) algorithm [4]. In this work the speech modality, is authenticated with a multi-lingual text-independent speaker verification system. The speech trait is comprised of two main components as shown in figure 1: speech feature extraction and a Gaussian Mixture Model (GMM) classifier. The speech signal is analyzed on a frame by frame basis, with a typical frame length of 20 ms and a frame advance of 10 ms [5]. For each frame, a dimensional vector feature is extracted, the discrete Fourier spectrum is obtained via a fast Fourier transform from which magnitude squared spectrum is computed and put it through a bank of filters. The critical band warping is done following an approximation to the Mel-frequency scale which is linear up to 1000 Hz and logarithmic above 1000 Hz. The Mel-scale cepstral coefficients are computed from the outputs of the filter bank [6]. The state of the art speech feature extraction schemes (Mel frequency cepstral coefficients (MFCC)) is based on auditory processing on the spectrum of speech signal and cepstral representation of the resulting features [7]. One of the powerful properties of cepstrum is the fact that any periodicities, or repeated patterns, in a spectrum will be mapped to one or two specific components in the cepstrum. If a spectrum contains several harmonic series, they will be separated in a way similar to the way the spectrum separates repetitive time patterns in the waveform. The description of the different steps to exhibit features characteristics of an audio sample with MFCC is shown in figure 2.

The distribution of feature vectors for each person is modeled by a GMM. The parameters of the Gaussian mixture probability density function are estimated using three different estimation algorithms. The Expectation Maximization (EM) algorithm [8], Greedy algorithm (GEM) [8] and Figureiro-Jain (FJ) algorithm [8].

Given a claim for person C’s identity and a set of feature vectors $X = \{x_1, x_2, ..., x_N\}$ supporting the claim, the average log likelihood of the claimant being the true claimant is calculated using:

$$P(X|C) = \frac{1}{N} \sum_{i=1}^{N} \log p(C|x)$$

(1)

where

$$p(C|x) = \sum_{i=1}^{N} \frac{p(x_i|C) \pi_i}{\sum_{j=1}^{N} p(x_j|C) \pi_j}$$

(2)

and

$$\pi = \{\pi_1, ..., \pi_N\}$$

(3)

Here $\pi$ is the model for person C, $N$ is the number of mixtures, $p(x_i|C)$ is the weight for mixture $i$ (with constraint $\sum_{i=1}^{N} \pi_i = 1$), and $M(x_i|\theta, \Sigma)$ is a multivariate Gaussian function with mean $\theta$ and diagonal covariance matrix $\Sigma$. Given a set $L_{BG}$ of $B$ background person models for person C, the average log likelihood of the claimant being an impostor is found using:

$$L_{BG}(x) = \log \left[ \frac{1}{B} \sum_{i=1}^{B} \exp (C|x_i) \right]$$

(4)

The set of background person models is found using the method described in [15]. An opinion on the claim is found using:

$$v = L(x|C) - L_{BG}(x)$$

(5)

The opinion reflects the likelihood that a given claimant is the true claimant (i.e., a low opinion suggests that the claimant is an impostor, while a high opinion suggests that the claimant is the true claimant).

2. Maximum Likelihood Parameter Estimation

Given a set of observation data in a matrix $X$ and a set of observation parameters $\theta$, the ML parameter estimation aims at maximizing the likelihood $L(\theta)$ or log likelihood of the observation data $X = \{x_1, ..., x_N\}$

$$\theta = \arg \max L(\theta).$$

(6)

Assuming that it has independent, identically distributed data, it can write the above equations as:

$$L(\theta) = p(X|\theta) = p(x_1, ..., x_N) = \prod_{i=1}^{N} p(x_i|\theta).$$

(7)

The maximum for this function can be find by taking the derivative and set it equal to zero, assuming an analytical function.

$$\frac{d}{d\theta} L(\theta) = 0.$$

(8)

The incomplete-data log-likelihood of the data for the mixture model is given by:

$$L(\theta) = \log p(X|\theta) = \sum_{i=1}^{N} \log p(x_i|\theta)$$

(9)
which is difficult to optimize because it contains the log of the sum. If it considers $X$ as incomplete, however, and posits the existence of unobserved data items $Y = \{y_i\}$ whose values inform us which component density generated each data item, the likelihood expression is significantly simplified. That is, it assume that $y_i \in [1, \ldots, k]$ for each $i$, and $y_i = k$ if the $i$th sample was generated by the $k$th mixture component. If it knows the values of $Y$, it obtains the complete-data log-likelihood, given by:

$$L(c, \theta) = \log p(X|Y, \theta)$$

$$= \sum_{i=1}^{n} \log p(x_i|y_i, \theta)$$

$$= \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k f_k(x_i; \theta_k) \right)$$

$$= \sum_{i=1}^{n} \log \left( \sum_{k=1}^{K} \pi_k \phi_k(x_i; \theta_k) \right)$$

Which, given a particular form of the component densities, can be optimized using a variety of techniques [9].

### 2.1 EM algorithm:

The expectation-maximization (EM) algorithm [8, 10, 11, 12] is a procedure for maximum-likelihood (ML) estimation in the cases where a closed form expression for the optimal parameters is hard to obtain. This iterative algorithm guarantees the monotonic increase in the likelihood $L$ when the algorithm is run on the same training database.

The probability density of the Gaussian mixture of $k$ components in $\mathbb{R}^d$ can be described as follows:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \phi_k(x; \theta_k) \quad \forall x \in \mathbb{R}^d,$$

where $\phi_k(x|\theta_k)$ is a Gaussian probability density with the parameters $\theta_k = (m_k, \Sigma_k)$, $m_k$ is the mean vector and $\Sigma_k$ is the covariance matrix which is assumed positive definite given by:

$$\mu(k) = \mu(k|m_k, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2} (x-m_k)^T \Sigma^{-1} (x-m_k)},$$

and $\pi_k \in [0, 1]$ ($k = 1, 2, \ldots, K$) are the mixing proportions under the constraint $\sum_{k=1}^{K} \pi_k = 1$. If it encapsulate all the parameters into one vector: $\theta = (\pi_1, \pi_2, \ldots, \pi_k, m_1, m_2, \ldots, m_k, \Sigma_1, \Sigma_2, \ldots, \Sigma_k)$, then, according to Eq. (23), the density of Gaussian mixture can be rewritten as:

$$p(x|\theta) = \sum_{k=1}^{K} \pi_k \phi_k(x; \theta_k) = \sum_{k=1}^{K} \pi_k \phi_k(x; m_k, \Sigma_k).$$

For the Gaussian mixture modeling, there are many learning algorithms. But the EM algorithm may be the most well-known one. By alternatively implementing the E-step to estimate the probability distribution of the unobservable random variable and the M-step to increase the log-likelihood function, the EM algorithm can finally lead to a local maximum of the log-likelihood function of the model. For the Gaussian mixture model, given a sample data set $S = \{x_1, x_2, \ldots, x_n\}$ as a special incomplete data set, the log-likelihood function can be expressed as follows:

$$\log p(S|\theta, Y) = \log p(x_1, x_2, \ldots, x_n|\theta, Y) = \sum_{i=1}^{n} \log p(x_i|y_i, \theta)$$

Which can be optimized iteratively via the EM algorithm as follows:

$$p(\theta|S, Y) = \frac{1}{n} \sum_{i=1}^{n} p(\theta|x_i|y_i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k p(x_i|\theta_k)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(x_i; \theta_k)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi_k(x_i; m_k, \Sigma_k).$$

Although the EM algorithm can have some good convergence properties in certain situations, it certainly has no ability to determine the proper number of the components for a sample data set because it is based on the maximization of the likelihood.

### 2.2 Greedy EM Algorithm:

The greedy algorithm (GEM) [8, 10, 12, 13] starts with a single component and then adds components into the mixture one by one. The optimal starting component for a Gaussian mixture is trivially computed, optimal meaning the highest training data likelihood. The algorithm repeats two steps: insert a component into the mixture, and run EM until convergence. Inserting a component that increases the likelihood the most is thought to be an easier problem than initializing a whole near-optimal distribution. Component insertion involves searching for the parameters for only one component at a time. Recall that EM finds a local optimum for the distribution parameters, not necessarily the global optimum which makes it initialization dependent method.

Given $\theta_c$, a $C$-component Gaussian mixture with parameters $\theta_c$, the general greedy algorithm for Gaussian mixture is as follows:

1. Compute the optimal (in the ML sense) one-component mixture $\theta_1$ and set $C = 1$.
2. Find a new component $\varphi(\mu_2, \Sigma_2)$ and corresponding mixing weight $\pi_2$ that increase the likelihood the most:

$$\{\varphi, \pi_2\} = \arg \max_{\varphi, \pi} \{\varphi, \pi\}$$

while keeping $\theta_1$ fixed.
3. Set $\theta_2 = \theta_1 - \theta_2$ and then $C = C + 1$.
4. Update $\theta_2$ using EM (or more other method) until convergence.
5. Evaluate some stopping criterion; go to step 2 or quit.

The stopping criterion in Step 5 can be for example any kind of model selection criterion, wanted number of components, or the minimum message length criterion.
The crucial point is of course Step 2. Finding the optimal new component requires a global search, which is performed by creating $N_{	ext{send}}$ candidate components. The number of candidates will increase linearly with the number of components $C$, having $N_{	ext{send}}$ candidates per each existing component. The candidate resulting in the highest likelihood when inserted into the (previous) mixture is selected. The parameters and weight of the best candidate are then used in Step 3 instead of the truly optimal values.

The candidates for executing Step 2 are initialized as follows: the training data set $\mathbf{X}$ is partitioned into $C$ disjoints data sets $\{X_1, \ldots, X_C\}$ according to the posterior probabilities of individual components; the data set is Bayesian classified by the mixture components. From each $A_i$, number of $N_{\text{send}}$ candidates are initialized by picking uniformly randomly two data points $x_1$ and $x_r$ in $A_i$. The set $A_i$ is then partitioned into two using the smallest distance selection with respect to $x_1$ and $x_r$. The mean and covariance of these two new subsets are calculated with only the points in the respective set $A_i$. The lower-bound log-likelihood is overcome three major weaknesses of the basic EM one component for every single training sample.

When CEM converges, it is not guaranteed that the Figueiredo-Jain (FJ) [8,10,12,13] algorithm tries to EM components), and then select the "best" model. The idea behind the FJ algorithm is to abandon such hierarchy and to find the "best" overall model directly. Using the minimum message length criterion and applying it to mixture models leads to the objective function:

\[ a(\theta, x) = \frac{1}{N_{\text{send}}} \ln{\frac{1}{N_{\text{send}}}} + \ln{\mathbb{E}[Z_n]} + \ln{\mathbb{E}[Z]} - \ln{\mathbb{E}[Z_n]} \]  \hspace{1cm} (27)

Where $N$ is the number of training points, $V$ is the number of free parameters specifying a component, and $\mathcal{C}_{nz}$ is the number of components with nonzero weight in the mixture ($C > 0$). $\theta$ case in the Gaussian mixture is the same as in (Eq. 3) the last term $\ln{\mathbb{E}[Z_n]}$ is the log-likelihood of the training data given the distribution parameters (Eq. 13).

The EM algorithm can be used to minimize Eq. 27 with a fixed $\mathcal{C}_{nz}$. It leads to the M-step with component weight updating formula:  \[ \alpha_{j+1} = \frac{\max\{0, \mathbb{E}[Z_{j+1} w_{nz}] - \mathbb{E}[Z]\}}{\mathbb{E}[Z] \max\{0, \mathbb{E}[Z_{j+1} w_{nz}] - \mathbb{E}[Z]\}} \]  \hspace{1cm} (28)

This formula contains an explicit rule of annihilating components by setting their weights to zero.

The above M-steps are not suitable for the basic EM algorithm though. When initial C is high, it can happen that all weights become zero because none of the components have enough support from the data. Therefore a component-wise EM algorithm (CEM) is adopted. CEM updates the components one by one, computing the E-step (updating $W$) after each component update, where the basic EM updates all components "simultaneously". When a component is annihilated its probability mass is immediately redistributed strengthening the remaining components.

When CEM converges, it is not guaranteed that the minimum of $A(\theta, X)$ is found, because the annihilation rule (Eq. 28) does not take into account the decrease caused by decreasing $\mathcal{C}_{nz}$. After convergence the component with the smallest weight is removed and the CEM is run again, repeating until $\mathcal{C}_{nz} = 1$. Then the estimate with the smallest $A(\theta, X)$ is chosen. The implementation of the FJ algorithm uses a modified cost function instead of $A(\theta, X)$:

\[ A(\theta, X) = \frac{1}{2} \sum_{j=1}^{V} \ln{w_{nz}} + \mathbb{E}[Z] \ln{N} - \ln{\mathbb{E}[Z]} \]  \hspace{1cm} (29)
III. EXPERIMENTS AND RESULTS
The experiments were performed using audio database extracted from video, which is encoded in raw UYVY AVI 640 x 480, 15.00 fps with uncompressed 16bit PCM audio; mono, 32000 Hz little endian. The capturing devices for recording the video and audio data were: Allied Vision Technologies AVT marlin MF-046C 10 bit ADC, 1/2" (8mm) Progressive scan SONY IT CCD; and Shure SM58 microphone. Frequency response 50 Hz to 15000 Hz. Unidirectional (Cardiod) dynamic vocal microphones. The extracted 16 bit PCM audio files (with wav header), were sampled at 16000 Hz, mono little endian. Thirty subjects were used for the experiments in which twenty-six are males and four are females. For each subject, six multi-lingual (.wav files) of one minute each recording were used for each subject. The database obtained from eNTERFACE 2005 [14]. For the experts, four speech recording samples of one minute each one were used for the modeling (training); two samples were used for the subsequent validation and testing. Three sessions of the speech database were used separately. Session one was used for training the speech experts. Each expert used ten mixture client models. To find the performance, Sessions two and three were used for obtaining expert opinions of known impostor and true claims.

Performance Criteria:
The basic error measure of a verification system is false rejection rate (FRR) and false acceptance rate (FAR) as defined in the following equations:

**False Rejection Rate (FRR):** is an average of number of falsely rejected transactions. If \( n \) is a transaction and \( x(n) \) is the verification result where 1 is falsely rejected and 0 is accepted and \( N \) is the total number of transactions then the personal False Rejection Rate for user \( i \) is

\[
FRR_i = \frac{1}{N} \sum_{n=1}^{N} x(n)
\]  

(30)

**False Acceptance Rate (FAR):** is an average of number of falsely accepted transactions. If \( n \) is a transaction and \( x(n) \) is the verification result where 1 is a falsely accepted transaction and 0 is genuinely accepted transaction and \( N \) is the total number of transactions then the personal False Acceptance Rate for user \( i \) is

\[
FAR_i = \frac{1}{N} \sum_{n=1}^{N} x(n)
\]  

(31)

Both FRR and FAR are usually calculated as averages over an entire population in a test. If \( P \) is the size of populations then these averages are

\[
FRR = \frac{1}{P} \sum_{i=1}^{P} FRR_i
\]  

(32)

\[
FAR = \frac{1}{P} \sum_{i=1}^{P} FAR_i
\]  

(33)

**Equal Error Rate (EER):** is an intersection where FAR and FRR are equal at an optimal threshold value. This threshold value shows where the system performs at its best (see Figure 3).

IV. CONCLUSIONS
As a common starting point, classifier parameters were selected to obtain performance as close as possible to EER on clean test data (following the standard practice in the biometric verification area of using EER as a measure of expected performance). A good decision is to choose the decision threshold such as the false accept equal to the false reject rate. In this paper it uses the Detection Error Tradeoff (DET) curve to visualize and compare the performance of the system.

The paper has presented a human authentication method of behavioural biometrics speech information. Simulation results show that state-of-the art finite mixture modal...
(GMM) is quite effective in modelling the genuine and impostor score densities. The (EM), (GEM) and (FJ) estimation algorithms achieve a significant performance rates, EER=0.26 % for "EM", EER=0.21 % for "GEM" and EER=0.16 % for "FJ". Hence, the behavioral information scheme based speech biometrics is robust and have a discriminating power, which can be explored for identity authentication.

REFERENCES


