International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 04, April-2015

GMM Based Speaker Verification System

Privacy Preserving System

Sayana P Babu
M.Tech Student
Dept. of Electronics and Communication
College of Engineering
Cherthala, India

Jayadas C. K.
Associate Professor
Dept. of Electronics and Communication
College of Engineering
Cherthala, India

***Abstract*—Speaker verification system is a popular biometric system and is widely used in application such as speaker authentication. This paper presents a framework for speaker verification by preserving the speaker privacy. A small speech sample of a speaker is a private form of communication and contains information such as a message via words, gender, language being spoken, emotional state etc. In this, the verification system does not have a direct access to the voice input provided by the speaker. The system also provides privacy of the speaker model saved by the system and thus preventing it to verify the speaker elsewhere. By using Gaussian mixture model and features extracted from the speech signal we build a unique identity for each speaker enrolled within the system for later verification with privacy criteria.**

Keywords—Speaker Recognition; Feature extraction; Mel Frequency Cepstral Coefficients; Statistical model; Gaussian Mixture Model;

I. INTRODUCTION

The underlying premise for speaker recognition is that given speech sample is a unique characteristic of an individual, each person's voice and manner of speaking are different and makes it uniquely distinguishable. The speaker recognition system is mainly divided into two: (1) Speaker Verification system, and (2) Speaker Identification system. Speaker verification is the authentication of a person based on a speech input. Speaker identification system is an extension of speaker verification is a related problem of identifying the speaker from a given set of speaker that best corresponding to a given speech sample.

Each speaker recognition system has two distinct phases, training phase and test phase. During training phase a user enrolls the system by providing enrollment recordings and typically a number of features are extracted to form a voice print, template, or model for each speaker. In the verification phase, a speech sample or "utterance" is compared against a previously created speaker model. Speaker recognition systems can be divided into two categories: text-dependent and text-independent. In a text-dependent system, text is same during enrollment and verification phase. In text-independent systems the speaker is allowed to say anything, i.e. the speech during training and testing are different.

In a conventional speaker verification system, the speaker models are stored without any obfuscation and the system matches the speech input obtained during authentication with these models. If the speaker verification system is compromised, an adversary can use these models to later impersonate the user. Also the system requires complete access of the speech input provided by the user without any privacy. This will be a problem in audio based surveillance applications like listening conversations of innocent individuals. It is therefore important to develop privacy preserving speech verification system that enables to verify the speaker by their speech input, while simultaneously providing privacy to speech input and speaker models of the user. Clearly in order to ensure privacy, the system should not have clear access to speech input and also not possess a model of the user's speech that it could use to identify the speaker elsewhere.

II. FEATURE EXTRACTION

Feature extraction transforms the speech signal to a set of feature vectors. Although there is no exclusively speaker distinguishing speech features, the speech spectrum has been shown to be very effective for speaker recognition. This is because spectrum reflects a person's vocal tract structure, the predominant physiological factor which distinguishes one person's voice from others.

A. Mel-Frequency Cepstral Coeffcients(MFCC)

The voice generated by a speaker is filtered by the shape of the vocal tract, and this shape gives an accurate representation of the phoneme being produced. This shape manifests itself in the envelope of the short time power spectrum of speech. MFCC will accurately represent this envelope and are widely used in speech processing.

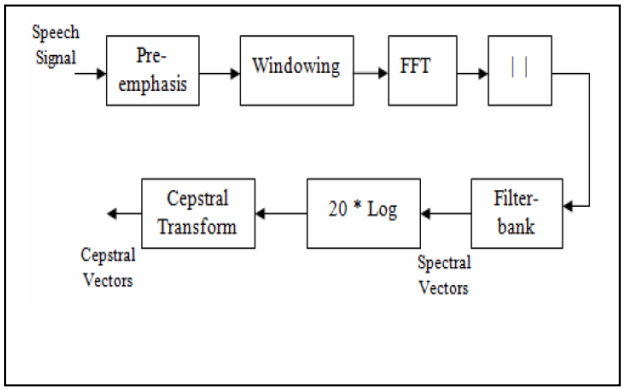

Fig. 1. Filter bank based cepstral parameterization

MFCC are obtained by filter bank-based cepstral parameterization. Fig. 1 shows a modular representation of a filterbank-based cepstral representation.

The First step is to pre-emphasis the speech signal where the signal is sent through a filter which emphasizes higher frequencies. The analysis of the signal over a sufficiently short period of time is done by the application of a window. The duration of window in time is shorter than the whole signal. Hamming window is used for this purpose. Each frame from the time domain is converted into the frequency domain by the application FFT to obtain spectral vector. Then the modulus of the FFT is extracted and a power spectrum is obtained, which contains a lot of fluctuations. To obtain the envelope of the spectrum, multiply the spectrum with a previously obtained Mel filterbank. Filterbank is a series of triangular filters with Mel scale for the frequency localization. The localization of the central frequencies of the filter is given by

$$f_{\text{MEL}} = 2595 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

Spectral vectors are obtained after taking log of this spectral envelope. Each coefficient is multiplied by 20 in order to obtain spectral envelope in dB. Finally discrete cosine transform is applied to the spectral vectors to yield cepstral coefficients

$$c_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \qquad n = 1,2, \ldots . L \quad (2)$$

where K is the number of log-spectral coefficients S_k , and L is the number of cepstral coefficients.

III. STASTICAL MODELING

Feature extraction converts the speech signal into a D-dimensional feature vector. From this the system forms a speaker model using some statistical model like Gaussian Mixture Model (GMM) for each speaker during the enrollment. Subsequently during verification the system compares the incoming speech signal with the stored model of the claimed user and determines if the speaker is indeed who they claim to be.

A. Speaker Verification using GMMs

GMM is a classic parametric method and has been successfully employed in several text-independent speaker recognition applications [4]. We use speaker verification method based on adapted Gaussian mixture models (GMM) as the underlying technique, because this mode has good ability of recognition. One of the powerful attributes of the GMM is its ability to form smooth approximation to the underlying long-term sample distribution of observations obtained from utterances by a given speaker.

A GMM is represented as weighted sum of M Gaussian component densities as

$$P(x|\lambda) = \sum_{i=1}^{M} w_i \, p_i(x) \quad (3)$$

Where x is a D-dimensional feature vector, p_i (x) is the component densities and w_i is the mixture weights. The Gaussian Function can be defined as

$$p_i(x) = \frac{1}{2\pi^{D/2}|\sum_i|^{1/2}} e^{-(1/2)(x-\mu_i)'\sum_i^{-1}(x-\mu_i)} \quad (4)$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weight satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weight from all component densities. These parameters can collectively represented by the notation.

$$\lambda = \{ w_i , \mu_i , \sum_i \} \text{ , for i = 1,2,……..,M} \quad (5)$$

In speaker verification system, each speaker can be represented by such a GMM and is referred to by the above model λ.

The most common technique for text-independent speaker verification treats the problem as one of hypothesis testing performed using a Likelihood ratio test [2]. To authenticate a recording X given by a speaker, i.e.to check if it is likely to be uttered by the enrolled speaker or by an imposter, the system computes the probability of X using a model λ_S for the speaker and compares it to the probability computed from a Universal Background Model (UBM) λ_U representing generic speech. Verification uses the following rule

$$\frac{P(\mathbf{X}|\lambda S)}{P(\mathbf{X}|\lambda U)} = \begin{cases} \geq \theta \text{accept speaker} \\ \leq \theta \text{reject speaker} \end{cases} \quad (6)$$

Where θ pre-calibrated the decision threshold for accepting or rejecting speaker.

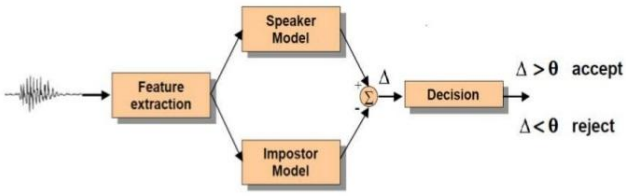

Fig. 2. Likelihood-ratio-based speaker verification system.

Fig. 2 shows the basic components found in speaker detection systems based on likelihood ratios. The parameters of the UBM are learned from a collection of speech recordings from a large number of speakers, to represent the characteristics of a generic speaker. These parameters are learned using the expectation-maximization (EM) algorithm. The parameters of the model for a speaker are obtained by adapting the UBM to the speaker.

B. Model Adaptation

The UBM parameters are adapted to individual speakers using maximum a posteriori (MAP) estimation to obtain each speaker model. These models obtained by MAP estimation significantly outperform the models trained directly on the enrollment data.

The MAP adaptation procedure comprises estimation of a sample estimate of the speaker's parameters, followed by interpolation with the UBM. Given set of enrollment speech samples $x_1, x_2, \ldots . x_n$, we first compute the *a posteriori* probabilities of the individual Gaussians in the UBM . For the i^{th} mixture component of the UBM,

$$P(i|x_t) = \frac{w_i^U \, N\,(x_t;\, \mu_i^U \,;\, \sum_i^U)}{\sum_j w_j^U \, N\left(x_t;\, \mu_j^U \,;\, \sum_j^U\right)} \quad (7)$$

Similar to the EM algorithm, we use the *a posteriori* probabilities to compute new weights, mean, and second moment parameters.

$$w_i' = \frac{1}{T}\sum_t P(i|x_t)$$

$$\mu_i' = \frac{\sum_t P(i|x_t)x_t}{\sum_t P(i|x_t)}$$

$$\sum_i' = \frac{\sum_t P(i|x_t)x_t\, x_t^T}{\sum_t P(i|x_t)} \quad (8)$$

Finally, the parameters of the adapted model λs from the convex combination of the above parameters and the UBM parameters are obtained as follows,

$$\hat{w}_i^S = \alpha_i w_i' + (1-\alpha_i)w_i^U$$

$$\hat{\mu}_i^S = \alpha_i \mu_i' + (1-\alpha_i)\mu_i^U$$

$$\hat{\sum}_i^S = \alpha_i \sum_i' + (1-\alpha_i)[\sum_i^U + \mu_i^U \mu_i^{UT}] - \hat{\mu}_i^S \hat{\mu}_i^{ST} \quad (9)$$

The adaptation coefficients α_i control the amount of contribution of the enrollment data relative to the UBM.

IV. Privacy Preseving Speaker Verification

The process of speaker verification is mainly divided into two phases. The first phase is enrollment, in which speech samples are collected from all speakers, and they are used train models. The collection of enrolled speaker model is also called speaker database. In second phase identification or verification phase, the system compares a test speech sample from an unknown speaker against the stored speaker database. Both the phases contain the same first step, feature extraction, which extracts speaker dependent characters from speech samples. The main purpose of this step is to reduce the amount of test data while retaining the speaker discriminative information. After feature extraction and modeling the features are encrypted using public key cryptosystem and these are modeled and stored in the database to protect from an adversary who can break the system. This process is represented in the Fig. 3.

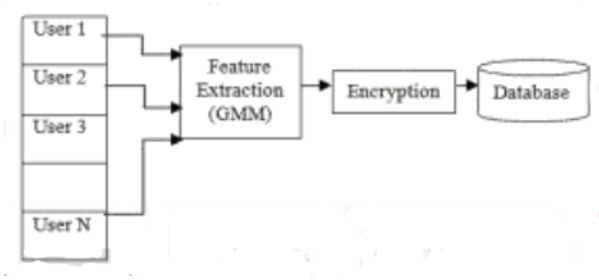

Fig. 3 Enrolment phase

In identification or verification step, the extracted features are compared against the models stored in the speaker database. Based on these comparisons the final decision about the speaker identity is made. This process is represented in Fig. 4

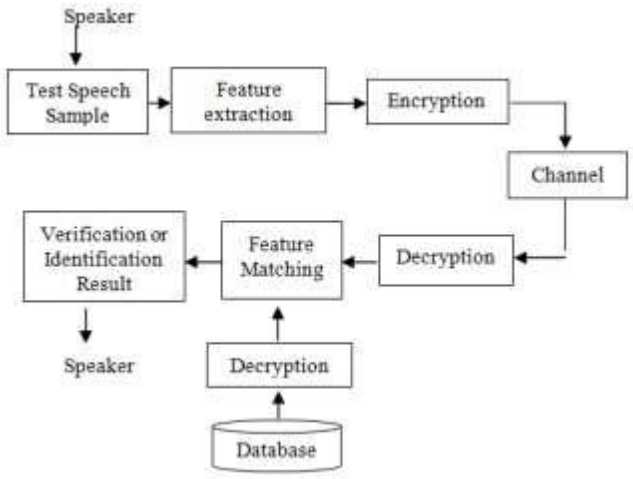

Fig. 4 Identification or Verification phase

However, these two phases are closely related. For instance, identification algorithm usually depends on the modeling algorithm used in the enrollment phase. We use GMM for speaker model generation.

One of the main block involved both in the enrollment and verification phase is Public key Cryptosystem. Public key Cryptography is also known as asymmetric cryptography, is a class of cryptographic algorithms which requires two separate keys, one of which is secrete or private key and the other one is public key. Although different, the two parts of the key pair are mathematically linked. The public key is used to encrypt plaintext, whereas the private key is used to decrypt the ciphertext. The term "asymmetric" stems from the use of different keys to perform these opposite functions, each the inverse of the other. We developed a framework for privacy preserving speaker verification using GMM and RSA cryptosystem.

V. Conclusion

With increasing use of speech-based services, the problem of the privacy of the speakers and their speech data is considered in this work. In this we developed GMM-based privacy-preserving speaker verification system in Matlab using homomorphic RSA cryptosystem. The system observes only encrypted speech data, and hence, cannot obtain any information about the user's speech. We also developed a framework for privacy preserving speaker identification using GMM and RSA cryptosystem. In this model, the system is able to identify a speaker from a given set of speakers which best corresponding a given speech input provided by the client without being able to observe the input. The GMM provides a simple and effective speaker representation which is computationally inexpensive and provides high recognition accuracy. Using this probabilistic speaker model, the recognition systems were defined as implementations of maximum likelihood classification and hypothesis testing rules.

REFERENCES

[1] Manas A. Pathak and Bhiksha Raj, "Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models", IEEETransaction on Audio, Speech, And Language Processing, Vol. 21 NO. 2,PP.397-406, 2013

[2] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D.Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independentspeaker verification," *EURASIP J. Appl. Signal Process.*, vol.

[3] D. A. Reynolds, "Speaker identification and verification usingGaussian mixture speaker models," *Speech Commun.*, vol. 17, no.1–2, pp. 91–108, 1995.

[4] D. A. Reynolds and R. C. Rose, "Robust text-independent speakeridentification using Gaussian mixture speaker models," *IEEE Trans.Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[5] M. Pathak and B. Raj, "Privacy preserving speaker verification usingadapted GMMs," in *Proc. Interspeech*, 2011.

[6] P. Smaragdis and M. Shashanka, "A framework for secure speechrecognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no.4, pp. 1404–1413, May 2007..

[7] M. Pathak, S. Rane, W. Sun, and B. Raj, "Privacy preserving probabilisticinference with hidden Markov models," in *Proc. ICASSP*, 2011,pp. 5868–5871.

[8] Pathak, M., Raj, B.: Privacy-Preserving Speaker Verification as Password Matching.In: Proc. ICASSP (2012)