Annals of
Reviews and Research

Juniper Publishers
key to the Researchers

Research Article
Volume 4 Issue 3 - November 2018

Ann Rev Resear
Copyright © All rights are reserved by Mustafa Yankayiş

Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition

Mustafa Yankayiş[1]*, Tolga Ensari[2] and Nizamettin Aydin[1]

[1]*Department of Computer Engineering, Yıldız Technical University, Turkey*

[2]*Department of Computer Engineering, Istanbul University, Turkey*

Submission: August 08, 2018; **Published:** November 19, 2018

***Corresponding author:** Mustafa Yankayiş, Department of Computer Engineering, Yıldız Technical University Istanbul, Turkey.

Abstract

In this study, the performance of the prominent feature extraction and modeling methods in speaker recognition systems are evaluated on the specifically created database. The main feature of the database is that subjects are siblings or relatives. After giving the basic information about speaker recognition systems, outstanding properties of the methods are briefly mentioned. While Linear Predictive Cepstral Coefficients (LPCC) and Mel-Frequency Cepstral Coefficients (MFCC) methods are preferred for feature extraction, Gaussian Mixture Model (GMM) and I-Vector methods are employed for modeling. The best results are tried to be obtained by changing the parameters of these methods. A number of features for LPCC and MFCC and number of mixture components for GMM are the parameters experimented by changing. The aim of this study is to find out which parameters of the most commonly used methods contribute the success and at the same time, to determine the best combination of feature extraction and modeling methods for the speakers having similar sounds. This study is also a good resource and guidance for the researchers in the area of speaker recognition.

Keywords: Speaker recognition; LPCC; MFCC; GMM-UBM; I-Vector

Introduction

The rapid development of information technologies has increased the importance of information security and led to the emergence of the new types of crimes (cyber-crime) accordingly. It also benefits from information technology for the detection of cyber-crimes. Reliable and valid digital evidence must be obtained for the elucidation of cybercrime. It is important for the reliability of the methods to obtain and examine digital evidence correctly since they can be easily damaged.

Recently, criminal cases have increased depending on the nature of the crime, the methods developed for the solution also have changed. The use of biometrics in forensics becomes a common solution increasingly. Speech and speaker recognition methods have been added to the popular biometric techniques such as fingerprint, face recognition, etc. The speech signal has significant energy up to around 4kHz. It has discriminative features to identify speakers. Dissemination of audio and video recording systems, increasing usage of mobile phones, the presence of speech in most criminal cases further enhance the importance of speech as evidence. The importance of forensic speech and speaker recognition system will increase by the assumption of increasing number of crimes in the future. Session variability, the voice similarities between people (especially siblings and relatives), voice imitation, the health of the speakers, psychological situations of the speakers, and environmental noise are some of the most important problems to overcome for the current speaker recognition systems [1]. An automated speaker recognition system mainly consists of the acquisition of the data, feature extraction, and similarity matching. A template matching is conducted between the voice acquired from a microphone or from a file and the previously recorded voice database. This system operates in two modes including training and testing. A reference model is created for each user in the training mode. A new input signal is compared with the generated reference models in the test mode [2].

Human voice carries information about the different characteristics of the speaker (health, age, gender, psychology, language, and identity etc.) and the environment in which speech is recorded. Thus, the voice signal is used as a reliable and distinctive feature in many sectors (forensics, telephone banking, telephone shopping, security control, voice control of computers, etc.). Voice processing technologies can be classified into two main categories such as speech recognition and speaker recognition. While speech recognition is related to what the talk is, speaker recognition is about who the speaker is. Speaker recognition systems can be categorized as speaker verification, speaker identification, and speaker diarization.

Speaker verification

The speaker claims to have a certain identity. It can be likened presenting your passport at border control. Voiceprint of a speaker is compared with the speaker claimed to be the person in the database. After the comparison, if the likelihood ratio is above a certain threshold, a result of acceptance is returned. If the likelihood ratio is below a certain threshold, a result of rejection

is returned. It is a vital issue to determine the threshold value. If a lower threshold value is taken, the system will respond to much false acceptance (false positive - I. type of error). It is not an acceptable situation for a system in which security is a concern. If a higher threshold value is taken, then the system will respond to many false rejections (false negative- II. type of error). In speaker verification, the performance is independent of the number of persons in the database since a one-to-one comparison is made.

Speaker identification

The speaker does not have any identity claims. In terms of operation, it can be likened comparing a sketch of the culprit with pictures in the database of criminals and finding the best match by a security guard. Voiceprint of a speaker is compared with the speakers in the database one by one. If all speakers are known by the system, it is called as "closed set" identification and the result is the best match. If all speakers are not known by the system, it is called as "open set" identification and sometimes there is a possibility of failure to a good match. In speaker identification, the performance is dependent on the number of persons in the database since a one-to-many comparison is made.

Speaker diarization

It is a kind of segmentation process that determining which speaker speaks in which part of the conversation in a speech signal belonging to two or more speakers. There are two stages in speaker diarization. Firstly, separating the speech signal into the segments in which a different speaker speaks. Secondly, finding which segment belongs to which speaker.

Speaker recognition can be realized in two different ways: text-dependent and text-independent. The text is previously known by the system in a text -dependent system. The speaker should speak a particular text (a word, a phrase, a password, digits, etc.) as an input to the system. Therefore, the texts in both the training and testing phases are the same. The risk is reduced and the system performance is affected positively. The text-dependent systems are mostly preferred in speaker verification. The text can be anything that a speaker speaks in a text-independent system which has a more flexible structure. The texts in the training and testing phases are different. Actually, the testing phase is realized without the knowledge of a person as in many forensic applications. The text-independent systems are mostly preferred in speaker identification.

Speech production process

Speech can be defined as sequences of sound segments called phones. Phones have certain acoustic and articulatory properties. Phones refer to the instances of phonemes which are smallest structural units that comprise words. The main components of the human speech production system are the lungs, trachea, larynx, vocal and nasal tract. Speech production is similar to the acoustic filtering. Lungs are like the power supply of the system which provides air flows through the vocal cords into the vocal tract. The vocal tract is the section that begins at the vocal cords and ends at the lips. The nasal tract begins at the velum and ends at the nostrils. The important organs having contributions to speech production are vocal folds (or vocal cords), tongue, lips, teeth, velum, and jaw [3]. Speech sounds fall into two classes according to vocal cords behaviors: voiced speech (the vocal cords vibrate at the fundamental frequency and air flows through them into the vocal tract) and unvoiced speech (the vocal cords are held open and air flows continuously through them). /a/, /e/, and /i/ are examples of voiced sounds and /f/, /s/, and /t/ are examples of unvoiced sounds.

Automatic speaker recognition process

The aim of automatic speaker recognition is to extract features and to differentiate the speakers. Automatic Speaker Recognition is performed in three main steps: feature extraction, modeling, and testing. The system operates in two modes: training mode and testing mode. A reference feature model is developed in the training mode. The input signal is compared with the reference model(s) in the testing mode to verify or identify the speaker. The general structure of the process is shown in Figure 1.

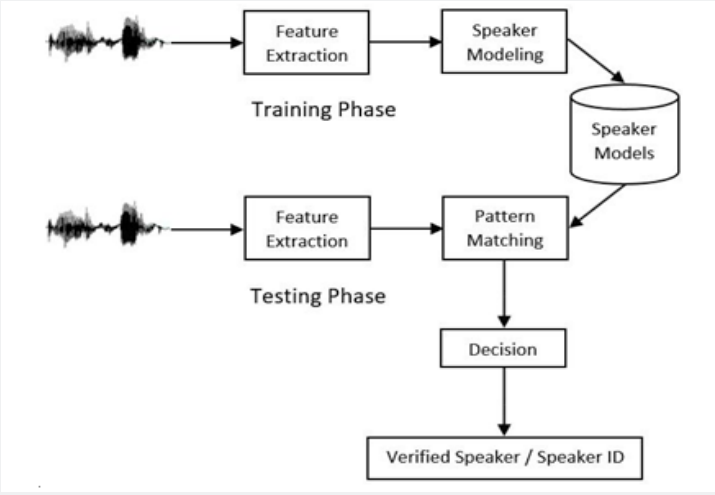

Figure 1: Block diagram of the speech recognition process.

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

The outline of the paper is as follows. Section 2 elaborates feature extraction and speaker modeling principles. Section 3 provides application details. Section 4 is devoted to the results for method pairs. Finally, we discuss the evaluation of speaker recognition performance with different parameter values of the methods in Section 5.

Methods

The aim of this study is to reveal the effects on the results by changing the parameters of the feature extraction and modeling techniques. Therefore, studies were performed on several techniques such as Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), Gaussian Mixture Model (GMM), and I-vector. The methods will be described briefly before the implementation details.

Feature extraction

Obtaining the values that characterize the speaker itself from a speech record is called feature extraction. It is a process for creating a small collection of data obtained from an audio signal. Feature extraction is vitally important for the performance of speech and speaker recognition systems. A feature should be easily accessible, and highly discriminative. It also needs to be protected. It should be trusted against imitation. A consistent feature is that the least affected by environmental factors (noise, microphone, telephone, etc.) and health conditions of the speaker (cold, flu, etc.) (Figure 1) [4].

Formant frequency is one of the short-term spectral features in speech and speaker recognition. A spectrogram of a speech signal displays audio components in three dimensions: time, frequency, and amplitude. The spectrum of vocal tract response consists of a number of resonant frequencies called "formants". Formants are the spectral peaks of each 1 kHz part of the sound spectrum. Three to four formants are present below 4 kHz of speech. Though formant frequencies of voiced speech are more explicit at lower frequencies, those of unvoiced speech is more explicit at higher frequencies. Formants differ from each other according to voiced speech and speakers, so they are very important to distinguish voiced speech. In Figure 2 the spectrum, spectrogram and formant frequencies of a speech signal are given.

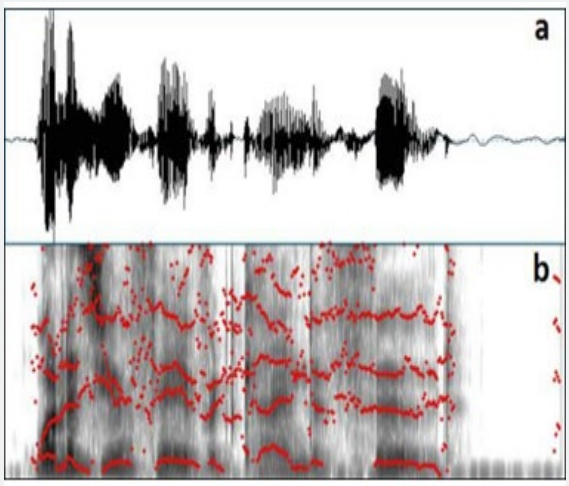

Figure 2: Spectrum of a speech signal (a), the spectrogram of the same speech signal and its formants (b).

Another important feature in speech processing is pitch frequency. The vibration frequency of vocal cords because of the air from the lungs during speech production is called pitch frequency (or the fundamental frequency of a particular person). It is very important for tonal languages because when a word is said in different tones (pitch frequency) it comes to different meaning. Many feature extraction methods are used in order to elucidate the distinctive features that characterize a speaker such as formant and pitch frequencies. Two of the prominent methods are summarized below.

Linear predictive cepstral coefficients (LPCC): Speech is a combination of signals representing voiced sounds, unvoiced sounds, and transitions between them. Voiced sounds are periodic with a speaker dependent fundamental frequency in a short time interval (Pitch Period). They are directly related to the vibration frequency of vocal cords because of the periodic structure. Although unvoiced sounds are similar to the noise, they are low amplitude sounds which are non-periodic in nature (vocal cords do not vibrate) and corresponds a specific meaning. In addition to voiced and unvoiced sounds, there are some intervals with voice inactivity during a conversation (Figure 2).

The peak resonance points of a spectrum envelope are the results of articulators revealed from different acoustic pits through vocal tract. The resonance frequency locations vary because of changes in vocal tract shape and dimensions. The resonance frequencies that define the form of the whole spectrum are called formants. The model showing the production of the speech signal by applying a vocal tract filter to voiced and unvoiced sounds is depicted in Figure 3. Linear Predictive Coding (LPC) is a method for analyzing human voice and extracting features from the analysis. LPC analysis models the vocal tract as a filter (H(z)). Some fundamental speech parameters like formants and pitch frequencies can be obtained from the coefficients that are the results of the LPC method [5].

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

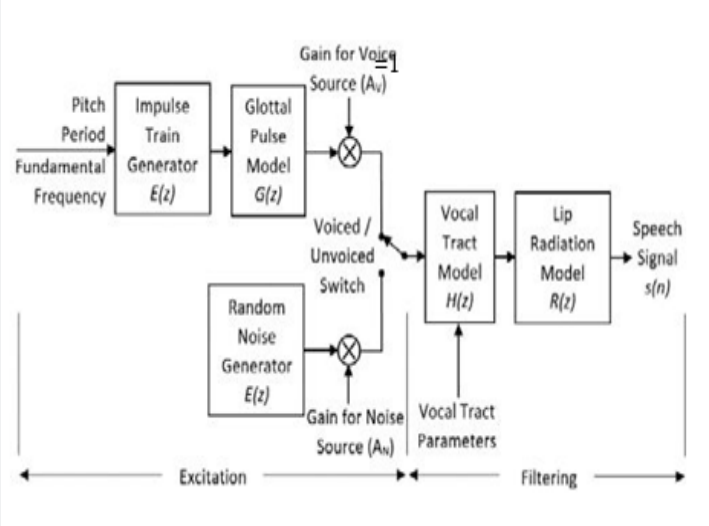

Figure 3: The model of the speech signal production.

The basic idea of this method is the obtainment of the sample in "nth" time of a speech signal as a linear combination of previous samples. The sound samples (s(n)) in "nth" time can be predicted by previous "p" samples with a certain margin of error(e[n]).

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n]$$

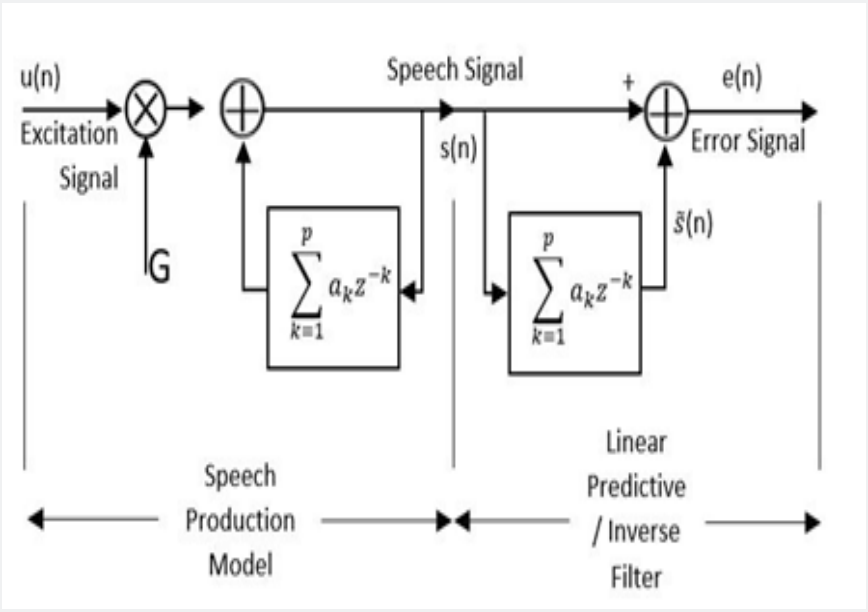

Figure 4: The LPC Model.

The error signal (e[n]) must be minimized. It must contain very little data in comparison with (s(n)). The system depicted in Figure 4 adjusts the coefficients so as to minimize the energy of (e[n]) (Figure 3,4).

$$e[n] = s[n] - s[n]$$

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k]$$

s[n] is the speech signal, $\mathbf{a}_k$ is the LPC coefficients, **s[n]** is the predicted signal.

Filter coefficients are assumed to be constant in certain ranges of the signal and LPC analysis is done in these frames that the signal is considered to be stationary. When LPC is applied to a frame of N samples length of a speech signal (N>>P), it gives linear predictive coefficients (a1, a2, a3, ap). These coefficients represent a significant portion of the signal.

Steps of the LPC model is shown in Figure 5. LPC method brings a linear approach to speech signal for all frequencies. But this is incompatible with human auditory perception. If the speech includes noise, LPC is inadequate for modeling the spectral characteristics of the speech [6]. Cepstral analysis is required to overcome the disadvantage of LPC. A method has been developed to convert LPC parameters to cepstral parameters (LPCC) (Figure 5,6).

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

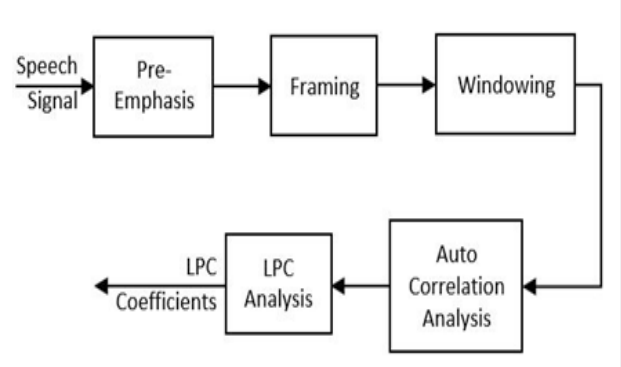

Figure 5: Obtaining LPC coefficients.

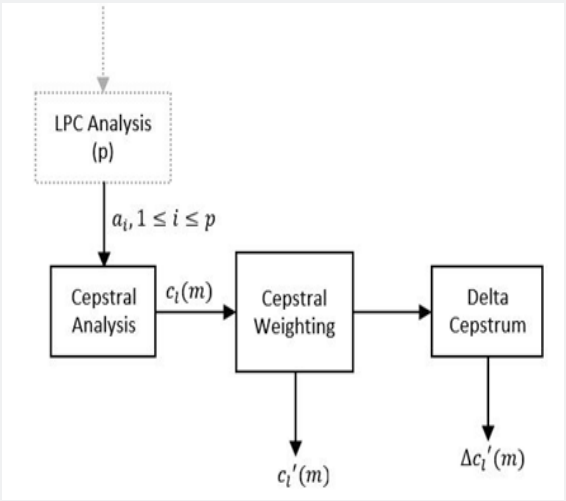

Figure 6: Derivation of LPCC from LPC.

Suppose that $c_{o=R(0);}$

$$cm = \begin{cases} am + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}, & 1 < m < p \\ \sum_{k=m-p}^{m-1} \frac{k}{m} c_k a_{m-k}, & m > p \end{cases}$$

Obtaining LPCC from LPC analysis is given in Figure 6.

Mel Frequency Cepstral Coefficients (MFCC)

The MFCC is the most common feature extraction method in speech and speaker recognition. It makes possible to obtain distinctive values for speakers by modeling the frequencies of the human auditory perception [7]. (Figure 7) shows the steps of the MFCC method [8]. In the first step, pre-emphasis, a filter boosts the high frequency of the speech signal is applied.

$$Y[n] = X[n] - 0.95.X[n-1]$$

Framing is to obtain stationary speech parts by dividing the non-stationary speech signal into short periods (~30 ms). Frames are placed by shifting ~10 ms and coinciding ~20 ms part of them. If the length of the frame is longer, it is less stationary. Besides shorter frames make it difficult to capture enough samples [9].

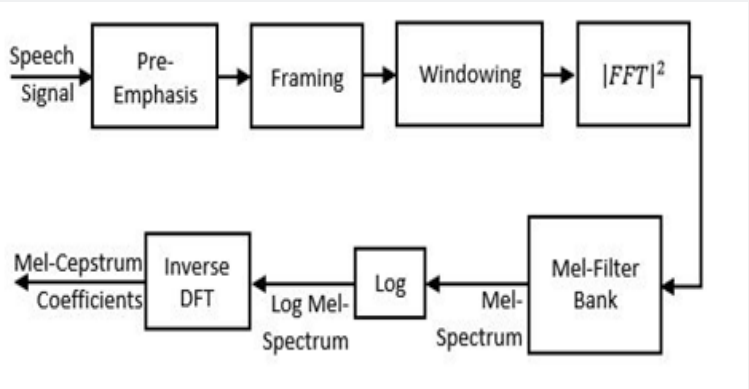

Figure 7: The steps of MFCC.

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

Framing introduces discontinuity into the signal resulting in a distortion. In order to minimize the effect of discontinuity usually, a windowing function such as (Hamming, Hanning, Gauss, Blackman, Rectangular etc.) is utilized. Each frame with N samples in the time domain is converted into the frequency domain by Fast Fourier Transform (FFT). Human auditory perception is approximately linear up to 1KHz, but it has logarithmic values over 1KHz (Mel-Frequency Scale). The structure of the human ear working as a filter is modeled by Mel-Filters [9]. Finally, logarithm compressing the dynamic range and inverse DFT allowing to return to time domain are applied [7].

A speaker-specific voiceprint is created after applying MFCC to the speech signal. Basically, the voiceprint has 12 coefficients. They can be increased up to 42 by taking the first and second derivatives (energy, delta MFCC features, double-delta MFCC features, etc.) [10]. Voiceprints created and modeled in the training phase compared with voiceprints created in the test phase and results obtained.

Modeling

Speaker models are used to represent speaker -specific features stored in feature vectors. The prominent speaker modeling techniques are listed below.

Dynamic time warping (DTW) [11]: Despite being a widely used classification technique in text-dependent speaker recognition systems, it has now left its place in statistical methods. It is useful for solving the timing problem in conversations at different speeds.

Vector Quantization (VQ) [12]: The dimension of the feature vector may increase if whole features of the speaker are used in text-independent speaker recognition systems. VQ is a method of reducing the dimension of the feature vector in which each speaker is represented by a codebook. The code book consists of code vectors that are the averages of feature vectors.

Hidden Markov Model (HMM) [13]: It is a statistical method applied successfully in speaker recognition. HMM creates a statistical model of how the speaker produces voice. Given the model parameters generated for the reference speakers, the possibility of hidden states forming an unknown output sequence is found using the Viterbi algorithm.

Support Vector Machines (SVM) [14]: It is a classification method which divides a space, positive and negative samples are known, into two, tries to find best hyper-plane and support vectors that make the distance between these samples the farther enough.

Gaussian Mixture Model (GMM) [6,15,16]: GMM is a model that works according to the principle of finding the Probability Density Function (PDF) representing the acoustic characteristics of a speaker. The PDF is found from the speaker's feature vectors by using multiple Gaussian Density Functions. In GMM, the probability of feature vector for the nth frame is obtained from the weighted sum of M multidimensional Gaussian Probability Density Function. Gaussian Mixture Model is expressed as a covariance matrix, mixture weights, and mean vector of each component.

In the training phase, the GMM parameters which are the most suitable for distribution of feature vectors are estimated. The Maximum Likelihood Estimation (MLE) is used for this estimation. MLE can be obtained in an iterative way using a special case of the Expectation Maximization (EM) algorithm. One of the parameters affecting the success and the performance of the GMM method used in text-independent speaker recognition systems is a number of mixtures. If the number of mixtures is low, features of the speaker will not be correctly modeled. If the number of mixtures is high, the performance of the process decreases during the training and testing phases. Experimentally, the optimal number of mixtures can be found for different cases.

Speaker modeling with GMM method consists of three main stages including development, enrollment, and test. During the development phase, a Universal Background Model (UBM) consisting the feature vectors of all training samples is created. The UBM values found during the development phase are used as the initial values to obtain the GMM in the enrolment phase. GMM values (w,μ,σ) for every person in the training set are found by adapting UBM. The models are compared with test values in the final phase. The steps of the GMM method are shown in (Figure 8).

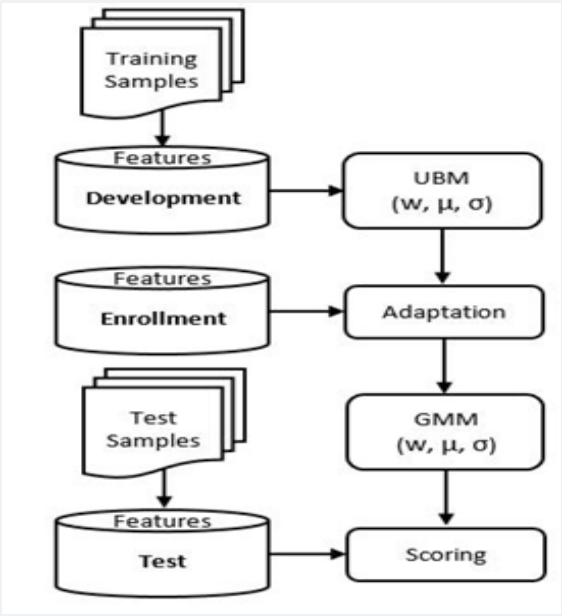

Figure 8: The steps of GMM method.

a. Global means (*gm*) and global variance (gv) are calculated in order to develop a UBM,

b. (*gm*) and (*gv*) are the initial values of the EM algorithm,

c. The sample signal is divided into 2^N mixtures to find the Log-Likelihood (LL),

d. LL is used to measure how a model is fitting a data,

$$L_L = \frac{1}{N}\log\prod_{K=1}^{N} p\left(X_K\right) = \frac{1}{N}\sum_{K=1}^{N}\log p\left(X_k\right)$$

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

e. Frames are distributed to related mixtures,

f. All the steps above are applied to whole training samples (E-Step),

g. The means and variances of mixtures are calculated for each iteration. These values are the initial values of E-Step in proceeding iterations (M-Step),

h. A UBM consisting of weights (w), means (μ), and variances (σ) of each mixture is obtained,

i. Then, GMM values (w, μ, and σ) of each speaker are calculated by adapting UBM values as initial values,

j. Ultimately, the values obtained in the training phase are compared with test values. The posterior probability of UBM and GMM for each frame in test data are calculated. The posterior probability of UBM is subtracted from the posterior probability of GMM and the mean of this subtraction is taken as Log-Likelihood-Ratio (LLR).

I-Vectors [17-20]: I-vectors are an instance of sub-space modeling approach. They are used to decrease the dimension of data before applying classifiers and training. The thought of I-vector emerges from Joint Factor Analysis (JFA) model used in speaker verification. The features of speakers are produced from a multivariate Gaussian Model in JFA. A speech sample is represented by a super vector (M) containing additional components from a speaker and channel (session) subspace. The super vector depending on a speaker is defined as

$$M = m + Vy + Ux + Dz$$

In the equation; $x, y,$ and z are low dimensional random variables with a normal distribution (with zero mean and unit diagonal covariance – N (0, I)). These vectors are factors depending on speakers and channels (sessions) in their respective subspaces. m is a speaker and session independent mean distribution super vector that can be produced from UBM. U (eigenchannel matrix) models channel variability (session subspace). V (eigenvoice) and D (diagonal residual) define a speaker subspace.

Firstly, it is necessary to estimate the subspaces (i.e.,) from appropriately labeled development corpora and then estimate the speaker and session factors (i.e.,) for a given new target utterance to apply JFA to speaker recognition. If m, V *and* D are known for all utterances speaker-dependent features are isolated in low dimensional y *and* z vectors. The speaker-dependent supervector is given by $s = m + Vy + Dz$

Test results are obtained by computing the likelihood of the test utterance feature vectors against a session-compensated speaker model. The block scheme of the I-vector based speaker recognition systems is shown in (Figure 9). The aim of using Linear Discriminant Analysis (LDA) is to maximize inter-speaker variances and to minimize intra-speaker variances.

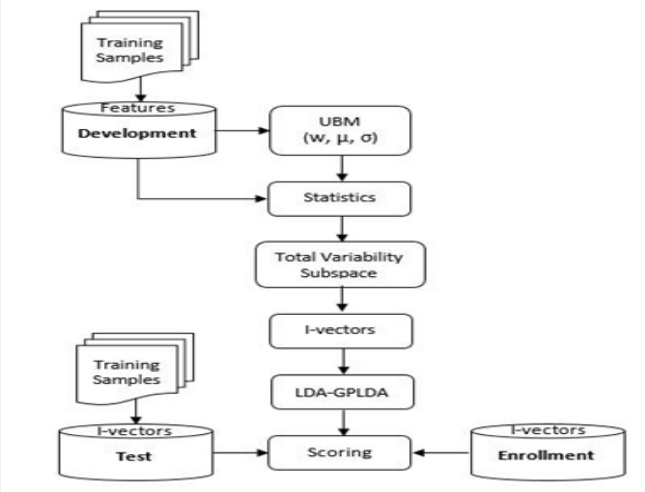

Figure 9: I-vector based systems.

Evaluation

There are various standards for measuring the performance of the biometric systems. One of them is the Equal Error Rate (EER). The False Acceptance Rate (FAR) and the False Rejection Rate (FRR) terms must also be known for EER to be understood.

a. The False Rejection Rate (FRR): It is also called "Type I" error. It indicates the possibility of inadvertent rejection of a person who should be able to access to the biometric system.

b. The False Acceptance Rate (FAR): It is also called "Type II" error. It shows the likelihood that someone who does not have access to a biometric system has access and misidentification as a registered person.

c. Equal Error Rate (EER): In biometric verification systems, the most critical error is expressed as the "Type II" error, but it is desirable that both of the above-mentioned error rates FAR and FRR are low. The low rate of both errors indicates the point where they are equal and it is called Equal Error Rate (EER) or Crossover Error Rate (CER). (Figure 10) demonstrates FAR, FRR, and EER on the same graph. High secure applications require lower false acceptance and higher false rejection rates, while high compatible and user-friendly

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

applications require higher false acceptance and lower false rejection rates.

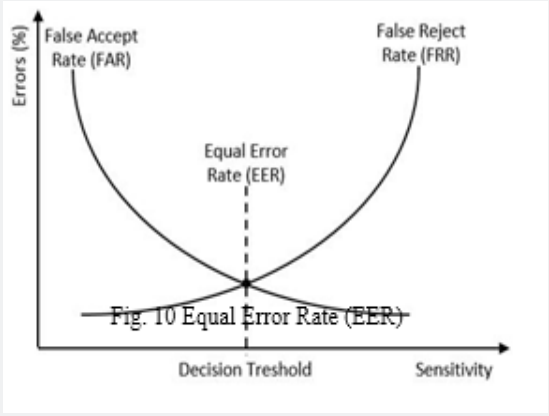

Fig. 10 Equal Error Rate (EER)

Figure 10: Demonstrates FAR, FRR, and EER on the same graph.

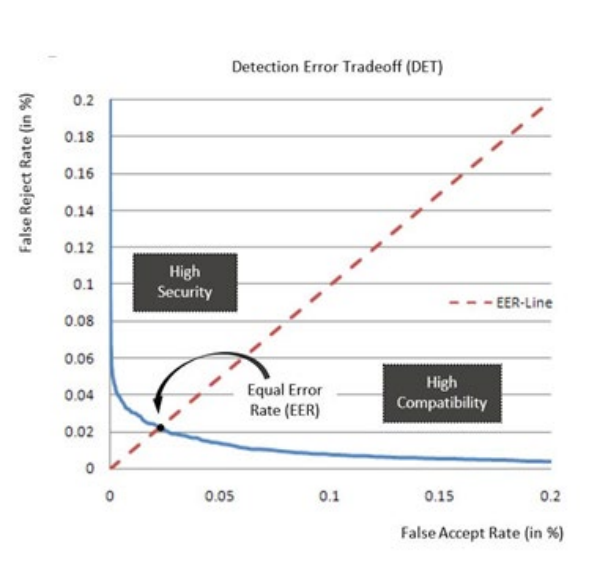

Figure 11: Detection Error Tradeoff (DET).

d. Detection Error Trade-offs (DET) Curve: In order to be able to more easily observe the system performance, the EER is represented by Detection Error Trade-offs (DET) Curve as in the figure. The closer the EER is to zero point, the better the system performance as shown in (Figure 11).

Application Details

LPCC and MFCC, the two most prominent feature extraction methods, and GMM and I-vector, the two most powerful text-independent modeling techniques, are applied on our dataset in the scope of this study. The methods are applied to 460 samples obtained from 46 people. There are 10 samples for each speaker. Five of them are used for training and the remaining five samples are used for testing. In the coding step of the methods [10,21,22] are very helpful guides.

In the feature extraction step, a number of the features obtained are 8, 12, 16, and 20 from the LPCC method and 12, 13, 14, 28 and 42 from the MFCC method. The LPCC coefficients are determined by the "order" variable which represents the number of features desired to be obtained. The MFCC coefficients are determined by the parameter "w":

a. w = ''; % Hamming window, without parameter: 12 features,

b. w = '0EdD'; % Hamming window parameters,

c. % '0' "0th degree" includes cepstral coefficients: 13 features,

d. % 'E' includes log energy: 14 features,

e. % 'd' includes delta coefficients (dc/dt): 28 features,

f. % 'D' includes delta-delta coefficients (d^2c/dt^2): 42 features.

The models for each speaker are created after getting UBM models in GMM-UBM based applications. Then, the models are compared with test data. The most suitable one is identified by changing the number of mixture components. In the application, some experiments are carried out with 32, 64, 128, 256, 512 and 1024 mixture components. The influence of the number of mixture components on the speaker recognition is observed.

Experimental Results

By using two prominent feature extraction methods LPCC (8, 12, 16, 20 features) and MFCC (12, 13, 14, 28 and 42 features), two different modeling techniques (GMM-UBM and I-vector) are examined. The parameters used in MFCC method are frame length (512 samples), frame shifting ratio (½, 256 samples), number of mel-filters (30), and window type (Hamming). In the GMM Model, the influence of the number of mixtures on speaker recognition is observed. The number of the mixture in GMM is taken as 32, 64, 128, and 256 and the results are evaluated accordingly. The results of feature extraction methods are directly compared without using a different parameter in the I-Vector Model.

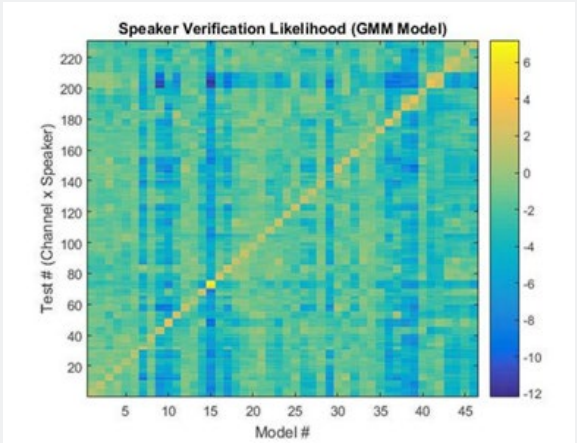

Figure 12: A likelihood matrix in the GMM Model (number of mixtures = 32).

As a result of the application, speaker verification likelihood was found by comparing the models to be used for the test and the

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

models obtained after the training. An example of these likelihood matrices is seen in Figure 12. All the results are also plotted as DET curves. The examples of the DET curves for the GMM Model with a different number of mixtures (128 and 256) are shown in Figure 13. In the following sections, results will be given in tabular form according to the feature extraction-modeling method pair (Figure 12,13).

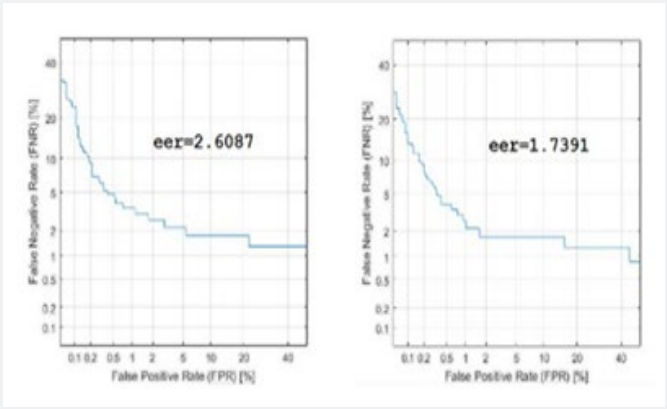

Figure 13: DET curves for the GMM Model with Different Number of Mixtures (128 and 256).

Experiments and results with LPCC & GMM pair

EERs obtained with GMM are tabulated according to a number of LPCC features and mixture components. (Table 1). Increasing the number of mixture components in this method seems to have a positive effect on speaker recognition in general. However, the increase in the number of LPCC features adversely affect the performance. The best result (EER=3,8841) is obtained with LPCC (12 features) and GMM (256 mixtures).

Table 1: Equal Error Rates for LPCC & GMM Pair.

GMM & LPCC	Number of Mixture Components			
	32	64	128	256
8 Features	4,589	4,348	4,483	10,145
12 Features	5,507	5,217	4,174	3,884
16 Features	17,382	11,894	10,812	6,087
20 Features	23,478	20,869	19,488	12,348

Experiments and results with LPCC & I-Vector pair

Table 2: Equal Error Rates for LPCC & I-Vector Pair.

LPCC &I-Vector	LDA DIM = 100UBM (nmix) = 256
8 Features	6,528
12 Features	4,309
16 Features	6,956
20 Features	3,478

EERs obtained with I-vector are tabulated according to a number of LPCC features (Table 2). The number of mixture components in the UBM calculation is set to 256. LDA dimension in Linear Discriminant Analysis is taken as 100. No significant effect of the increase in the number of LPCC features is detected. The best result (EER = 3,4783) is obtained with LPCC (20 features).

Experiments and results with MFCC & GMM pair

EERs obtained with GMM are tabulated according to a number of MFCC features and mixture components (Table 3). Increasing the number of mixture components in this method seems to have a positive effect on speaker recognition. However, no significant effect of the increase in the number of MFCC features is detected. The best result (EER=3, EER=1,7391) is obtained with MFCC (12, 13, 14 and 42 features) and GMM (256 mixtures).

Table 3: Equal Error Rate for MFCC & GMM Pair.

GMM & MFCC	Number of Mixture Components			
	32	64	128	256
12 Features	3,044	2,174	1,961	1,739
13 Features	2,927	2,609	2,174	1,739
14 Features	3,043	3,043	2,174	1,739
28 Features	3,043	3,043	2,579	2,174
42 Features	3,043	3,014	2,609	1,739

Experiments and results with MFCC & I-Vector pair

Table 4: Equal error rates for MFCC & I-Vector pair.

MFCC & I-Vector	LDA DIM = 100 UBM (nmix) = 256
12 Features	1,739
13 Features	1,739
14 Features	1,353
28 Features	2,609
42 Features	3,130

EERs obtained with I-vector are tabulated according to a number of MFCC features. (Table 4) The number of mixture components in the UBM calculation is set to 256. LDA dimension in Linear Discriminant Analysis is taken as 100. No significant effect of the increase in the number of MFCC features is detected. The best result (EER=1,3527) is obtained with MFCC (14 features). At the same time, this result is the best one among the methods used.

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

Discussion

The Equal Error Rates of the methods examined are given in (Table 5). It is clear that more successful results are obtained from the experiments done with MFCC for feature extraction and I-vector for modeling than the other methods used. The reason why the MFCC method gives more successful results is that obtaining the features better distinguish the speakers by modeling of human ear's frequency selectivity. The reason for the success of the I -Vector Model is that putting forward the most discriminating features by linear discriminant analysis.

Table 5: The Equal Error Rates (EER).

MODELING & FEATURE EXTRACTION		GMM (Number of Mixture Components)				I-vector UBM (nmix)
LPCC	8	32	64	128	256	256
		4,589	4,348	4,483	10,145	6,522
	12	5,507	5,217	4,174	3,884	4,309
	16	17,382	11,894	10,812	6,087	6,956
	20	23,478	20,869	19,488	12,348	3,478
MFCC	12	3,043	2,174	1,961	1,739	1,739
	13	2,927	2,609	2,174	1,739	1,739
	14	3,043	3,043	2,174	1,739	1,353
	28	3,043	3,043	2,579	2,174	2,609
	42	3,043	3,014	2,609	1,739	3,130

It is seen that MFCC (14 features) and I-Vector (LDA DIM=100 UBM (nmix) = 256) pair give the most successful result (EER=1,3527) in the experiments. When Table 5 is examined in more detail, it is observed that the change in the number of features in LPCC and MFCC does not have a meaningful effect on the results. The fact that the result is not influenced by the change in the number of features occurs also in some different studies on this topic [23] [24]. ΔMFCC in the MFCC method reflects the dynamic features of the speaker. The dynamic features show the variation between successive frames. ΔMFCC is generated by taking the first-order derivative of the cepstrum coefficients. These dynamic features seem to have no positive effect on speaker recognition both in this study and in some other studies examined in this area [15], [16].

Besides that, the increase in the number of mixture components from the GMM parameters has generally positive effects. The results of experiments in which the number of mixture components is increased to higher levels (512, 1024...) are not included in (Table 5) because the performance does not change. The reason why the number of mixture components does not affect the result after a certain level is that the number of mixtures and the highest number of phonemes people can make while speaking Turkish are correlated.

Conclusion

According to the results given in Table 5 the Equal Error Rates with which the highest success rates are achieved and the studies of the researchers who have important publications in this area are compared (the results in bold lines are from our studies). It is seen in (Table 6) (for I-Vector) and (Table 7) (for GMM) that the Equal Error Rates in our experiments are on a competitive level with some other prominent studies in this subject. These results encourage us that better results can be achieved if research and experimentation on methods continue [25,26].

Table 6: Comparison of the results (for I-Vector).

Modeling	Researchers	Year	Feature Extraction	Equal Error Rate (EER)
I-Vector	Yankayis et. al	2017	MFCC (14features)	1,353
I-Vector	Dehak N, et al. [25]	2011	MFCC (60 features)	1,12 (male)
I-Vector	Kanagasundaram A, et al. [26]	2011	MFCC (13 features)	3,13

Table 7: Comparison of the Results (for GMM).

Modeling	Researchers	Year	Feature Extraction	EqualError Rate (EER)
Gaussian Mixture Model (GMM)	Yankayis et. al	2017	MFCC (12,13, 14, 42features)	1,739
Gaussian Mixture Model (GMM)	Douglas A, Reynolds [27]	1995	-	0,24 (TIMIT)
Gaussian Mixture Model (GMM)	Douglas A, Reynolds [27]	1995	-	7,19 (NTIMIT)

References

1. Hollien H, Bahr RH, Harnsberger JD (2014) Issues in Forensic Voice. Journal of Voice 28(2): 170-184.
2. Chauhan T, Soni H, Zafar S (2013) A Review of Automatic Speaker Recognition System. International Journal of Soft Computing and Engineering (IJSCE) 3(4).

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.

3. Aggarwal RJ (2012) A Thesis Submitted in fulfillment of the requirement of the degree of Doctor of Philosophy, Improving Hindi Speech Recognition Using Filter Bank Optimization and Acoustic Model Refinement. Department of Computer Engineering National Institute of Technology, Kurukshetra-136119 India.
4. Wolf JJ (1972) Efficient Acoustic Parameters for Speaker Recognition. The Journal of the Acoustical Society of America. 51(6B): 2044.
5. ÖCAL k, Seema Deoghare (2005) Application of Automatic Speech Recognition. Ankara University, Institute of Science and Technology, Department of Electronics Engineering, Ankara.
6. Reynolds DA, Rose RC (1995) Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models". IEEE Transactions on Speech and Audio Processing 3(1): 72-83.
7. Muda L, Begam M, Elamvazuthi I (2010) Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing 2(3).
8. Eskidere O, Ertaş F (2009) The Effects of Variabilities in Mel Frequency Cepstrum Coeffcients On Speaker Recognition." Uludağ University, The Journal of Faculty of Engineering and Architecture 14(2).
9. Chang WW Time-Frequency Analysis for Voiceprint (Speaker) Recognition. Graduate Institute of Communication Engineering National Taiwan University, Taipei, Taiwan.
10. Seyed Omid S, Malcolm S, Larry H (2013) MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker-Recognition Research." Speech and Language Processing Technical Committee Newsletter.
11. Segarceanu S, Zaharia T (1985) Speaker Verification Using the Dynamic Time Warping. UPB Sci Bull 75(1).
12. Soong F, Rosenberg A, Rabiner L, Juang BH (1985) A vector quantization approach to speaker recognition." Acoustics Speech and Signal Processing 10: 387-390.
13. Abdallah SJ, Osman IM, Mustafa ME (2012) Text-Independent Speaker Identification Using Hidden Markov Model." World of Computer Science and Information Technology Journal (WCSIT) 2(6): 203-208.
14. Campbell WM, Campbel JP, Gleason TP, Reynolds DA, Shen W (2007) Speaker Verification Using Support Vector Machines and High-Level Features. IEEE Transactions on Audio Speech and Language Processing 15(7): 2085-2094.
15. Hanilçi C (2007) Comparative Analysis of Speaker Recognition Methods, Bursa: Uludağ University, Institute of Science and Technology.
16. Karasartova S (2011) Investigation and Implementation of Text-Independent Speaker Identification Systems. Ankara University, Institute of Science and Technology, Ankara.
17. Tokheim AEH (2012) I Vector Based Language Recognition." Norwegian University of Science and Technology Department of Electronics and Telecommunications, Trondheim.
18. Kenny P, Boulianne G, Ouellet P, Dumouchel P (2007) "Joint factor analysis versus eigenchannels in speaker recognition." IEEE Trans Audio Speech Lang 15(4): 1435-1447.
19. Kenny P, Boulianne G, OuelletP, Dumouchel P (2007) "Speaker and session variability in GMM-based speaker verification." IEEE Trans Audio Speech Lang Process 15(4): 1448-1460.
20. Kenny P, Ouellet P, Dehak N, Gupta V, Dumouchel P (2008) A study of inter-speaker variability in speaker verification." IEEE Transactions on Audio Speech and Language Processing 16(5): 980-988.
21. Brookes M VOICEBOX, Department of Electrical & Electronic Engineering, Imperial College.
22. "Multimodal Information Group, the National Institute of Standards and Technology (NIST).
23. Kumar P, Lahudkar SL (2015) Automatic Speaker Recognition using LPCC and MFCC. International Journal on Recent and Innovation Trends in Computing and Communication 3(4): 2106-2109.
24. Reynolds DA (1994) Experimental evaluation of features for robust speaker identification." Speech and Audio Processing IEEE Transactions on 2(4): 639-643.
25. Dehak N, Kenny P, Dehak R, Dumouche P, Ouellet P (2011) Front-end factor analysis for speaker verification. Audio Speech and Language Processing IEEE Transactions 19(4): 788-798.
26. Kanagasundaram, Vogt R, Dean D, Sridharan S, Mason M (2011) I-vector Based Speaker Recognition on Short Utterances. Speech and Audio Research Laboratory, Brisbane, Australia.
27. Reynolds DA (1995) Automatic Speaker Recognition Using Gaussian Mixture Speaker Models. The Lincoln Laboratory Journal 8(2): 173-192.

This work is licensed under Creative Commons Attribution 4.0 License

Your next submission with Juniper Publishers will reach you the below assets

- Quality Editorial service
- Swift Peer Review
- Reprints availability
- E-prints Service
- Manuscript Podcast for convenient understanding
- Global attainment for your research
- Manuscript accessibility in different formats

(Pdf, E-pub, Full Text, Audio)

- Unceasing customer service

Track the below URL for one-step submission

https://juniperpublishers.com/online-submission.php

How to cite this article: Mustafa Y, Tolga E, Nizamettin A. Performance Evaluation of Feature Extraction and Modeling Methods for Speaker Recognition. Ann Rev Resear. 2018; 4(3): 555639.