



Mixture linear prediction Gammatone Cepstral features for robust speaker verification under transmission channel noise

Ahmed Krobba¹ · Mohamed Debyeche¹ · Sid-Ahmed Selouani²

Received: 10 April 2019 / Revised: 27 December 2019 / Accepted: 17 February 2020 /

Published online: 9 March 2020

© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

In this paper, we present a Mixture Linear Prediction based approach for robust Gammatone Cepstral Coefficients extraction (MLPGCCs). The proposed method provides performance improvement of Automatic Speaker Verification (ASV) using i-vector and Gaussian Probabilistic Linear Discriminant Analysis GPLDA modeling under transmission channel noise. The performance of the extracted MLPGCCs was evaluated using the NIST 2008 database where a single channel microphone recorded conversational speech. The system is analyzed in the presence of different channel transmission noises such as Additive White Gaussian (AWGN) and Rayleigh fading at various Signals to Noise Ratio (SNR) levels. The evaluation results show that the MLPGCCs features are a promising way for the ASV task. Indeed, the speaker verification performance using the MLPGCCs proposed features is significantly improved compared to the conventional Gammatone Frequency Cepstral Coefficients (GFCCs) and Mel Frequency Cepstral Coefficients (MFCCs) features. For speech signals corrupted with AWGN noise at SNRs ranging from (-5 dB to 15 dB), we obtain a significant reduction of the Equal Error Rate (EER) ranging from 9.41% to 6.65% and 3.72% to 1.50%, compared with conventional MFCCs and GFCCs features respectively. In addition, when the test speech signals are corrupted with Rayleigh fading channel we achieve an EER reduction ranging from 23.63% to 7.8% and from 10.88% to 6.8% compared with conventional MFCCs and GFCCs, respectively. We also found that the combination of GFCCs and MLPGCCs gives the highest performance of speaker verification system. The best performance combination achieved is around EER from 0.43% to 0.59% and 1.92% to 3.88%.

Keywords Automatic speaker verification · Mixture linear prediction · Gammatone Frequency Cepstral Coefficients (GFCCs) · I-vector GPLDA · Transmission channel noise

✉ Ahmed Krobba
akrobba@usthb.dz; hkrobba@gmail.com

1 Introduction

The development of mobile devices such as: smartphones or tablets and the use of the voice biometrics (e.g. for access control, device personalization, transaction banking) have paved the way for a large number of new multimedia applications [7, 39]. Automatic Speaker Recognition (ASR) refers to recognizing a person based on his/her voice as a biometric feature. It consists of two tasks: Speaker Identification (SI) and Speaker Verification (SV). In speaker identification task, an unknown speaker is compared against a set of known speakers, and the best matching speaker gives us the identification. Speaker verification is the process of accepting or rejecting the identity claim of the speaker [16]. Speaker Recognition systems are broadly classified into two categories: Text Independent (TI-SR) and Text-Dependent (TD-SR). In TI-SR, the speaker can pronounce any sentence to be recognized, i.e. the system does not impose any constraint for training and test sessions. However, TD-SR systems use the same phrases/sentences for training and test sessions [2, 10]. Nowadays, many applications using speaker recognition have been deployed to improve the authentication procedure such as banking over wireless digital communication network, security control for confidential information, telephone shopping, database access services and voice mail [10]. Feature extraction is the crucial component in speaker recognition system where the speech signal is represented in a compact manner, in which the extracted features are capable of separating the speakers from each other in their space. The effects of additive noise and/or channel distortion have always been one of the most important problems in speaker recognition research. Various techniques have been proposed to improve the performance of speaker recognition systems in presence of noise. Speech enhancement methods include, for example, Spectral Subtraction (SS) or Nonlinear Spectral Subtraction (NSS), Wiener filtering and Kalman filtering. Moreover, other processing techniques are proposed to increase the robustness of ASR systems. Some of these techniques use feature normalization such as Cepstral Mean and Variance Normalization (CMVN), Relative Spectral (RASTA) processing of speech, or feature warping [24, 27]. In [22] they used multi-condition model and the missing features to compensate signal. The work in [31] proposed a soft spectral subtraction method that handles missing features in speaker verification. A recent work on robust speaker recognition based on the i-vector technique made a significant progress in reducing the channel effect and the additive noise [4]. Different channel compensation techniques were used, such as the Within-Class Covariance Normalization (WCCN), the Linear Discriminant Analysis (LDA) and the Nuisance Attribute Projection (NAP) [32]. In [26], the authors proposed a new variant of robust Mel Frequency Cepstral Coefficients (MFCC) that are extracted from the estimated spectral magnitude Bispectral-MFCCs. Score domain techniques such as H-norm, Z-norm, and T-norm have been studied in [2]. Recently, [40] proposed a new feature based on articulatory movement to characterize the relative motion trajectory of articulators in short-duration utterances. In [22], a multi-system fusion approach that uses multiple streams of noise-robust features for i-vector fusion is developed. The authors in [20] have analyzed the effects of multi-condition training on i-vector PLDA. In [6], the authors have proposed to use the Gammatone product-spectrum cepstral coefficients under noisy condition and speech codecs.

The great majority of past studies have addressed the effect of additive noise environment for speech and speaker recognition. However, only few studies have been reported the impact of Additive White Gaussian Noise (AWGN) and Rayleigh fading channels for speaker recognition performance. For instance, the work in [17] shows the effects of speech codecs, with AWGN and Rayleigh fading noises, on the performance of speaker recognition systems.

In [36], autoregressive MFCCs and Speech Activity Detection (SAD) algorithms have been applied for speaker recognition system over AWGN channel. In [13] a combination of modified LPC with Wavelet Transform (WT) in AWGN and real noise environments has been proposed. In [5], the authors proposed an approach that consists of using acoustic features, that are extracted directly from encoded bitstream, called ISF (Immittance Spectral Frequency), through a noisy channel (AWGN and Rayleigh).

MFCC coefficients are the most commonly used features in speaker and speech recognition systems. However, the MFCC features, which are computed by using a mel-scaled filter-bank are known to be very sensitive to additive noise. The auditory model based on mel scale in standard MFCC may not be optimal for speaker recognition [34] and the logarithmic nonlinearity used in MFCC to compress the dynamic range of filter bank energies does not possess noise immunity. In [42], the authors proposed a new front-end speech feature based on cochlear filter referred to as Gammatone Frequency Coefficient Cepstral (GFCCs). The work in [41] showed that the GFCCs features give superior performance of speaker recognition compared with other features such as MFCCs in noisy environments. Despite their relative robustness, it is important to mention that the GFCCs features are usually obtained by using the Fast Fourier Transform (FFT) [19]. Since the FFT requires the use of stationary signals within a given short-term frame; this may not analyze the non-stationary segments in transient state, which is not suitable in speaker recognition. Another popular feature extraction technique consists of the linear prediction (LP) filtering which is a well-known all-pole method for modeling the vocal tract by using a small number of parameters. The main drawback of conventional LP method is that the resulting spectral envelope may contain very sharp peaks for speakers with high pitch frequency.

Several modifications of LP method with an improved robustness against noise have been developed. One can cite the Weighted Linear Prediction (WLP), Stabilized Weighted Linear Prediction (SWLP), and regularization of linear prediction of spectrum analysis methods. Temporally weighted linear predictive [37] were studied in speaker verification under additive-noise condition. Extended Weighted Linear Prediction (XLP) [29] were evaluated for both channel distortion and additive noise. The study in [28] introduced a new algorithm based on linear predictive analysis utilizing an autoregressive (AR) Gaussian mixture model. In [30], the authors used an algorithm providing MFCCs features for speaker verification under vocal effort mismatch.

The use of a new approach of linear predictive modeling in this work is motivated by the ability of the linear predictive methods to capture relevant information from two major parts of the voice production mechanism that are the glottal excitation and the vocal tract. The LP signal analysis of this work uses a Gaussian mixture autoregressive model to compress the spectrum parameters. Besides this, it is showed in [11] that, even at low SNRs of environmental noise, the Gammatone filter bank and cubic root rectification provide more robustness to the features than the Mel-filter bank and log nonlinear.

In this paper, we propose a new feature extraction approach providing a Mixture of Linear Prediction Gammatone Cepstral Coefficients (MLPGCCs). The Mixture Linear Prediction MLP method is based on an autoregressive (AR) mixture model processed by Gammatone filter banks. This combination (i.e., MLP and Gammatone) is expected to take advantage of both MLP properties and Gammatone filtering to improve the robustness of speaker verification system under channel transmission noise [18]. The performance of speaker verification system is evaluated using i-vector and Gaussian Probabilistic Linear Discriminant Analysis GPLDA modeling.

The remainder of this paper is organized as follows. A brief introduction to the channel transmission noise is presented in Section.2. In Section.3, we describe the proposed MLPGCCs feature extraction algorithm. Block diagram of a MLPGCCs-based text-independent speaker verification system is presented in Section.4. Section.5 reports the performance evaluation carried out by comparing the proposed method with conventional extracted features. Finally, conclusions are summarized in Section 6.

2 Channel transmission noise

A communication system, as illustrated in Fig. 1, can be divided into two parts. The first part is digital and consists of the source encoder/decoder, the channel encoder/decoder and the digital modulator/demodulator. The second part is analog and is made of the transmitter, the receiver and the channel models. The modulation process involves the change of some parameters of a carrier wave, thus obtaining a set of signals suitable for a transmission channel. There are two main types of signal degradation introduced by the transmission channels: the first is attenuation and random variation of signal amplitude, and the second is distortion of the signal spectrum. Signal attenuation results from the degradation of the signal power level over distance while random variation of signal amplitude results from channel noise and multipath Rayleigh fading effects.

In order to implement the communication system, we used the standard speech codec, Adaptive Multi-Rate Wide Band (AMR-WB), codec, introduced by the European Telecommunication Standards Institute (ETSI). AMR provides better speech quality and more robustness for background noise. Binary Phase Shift Keying (BPSK) modulation and demodulation are simulated. We want to transmit symbols from an alphabet $\{m_i; i = 1, \dots, M\}$ and a signal $x_i(t)$, suitable for transmission and assigned to each symbol m_i . After transmission, we obtain a distorted version of the original $x_i(t)$ defined by $y_i(t)$. On the other hand, the distortion due to quantization and channel errors may make the received symbol \hat{m}_i different from the transmitted one m_i . We define the Additive White Gaussian Noise (AWGN) channel, which modifies the transmitted signal as

$$y(t) = x_i(t) + n(t) \quad (1)$$

where $n(t)$ is a white Gaussian-distributed noise of zero mean and variance $\sigma_n^2 = N_0/2$. In the AWGN channel, the noise is added to the transmitted signal by specifying the signal to noise

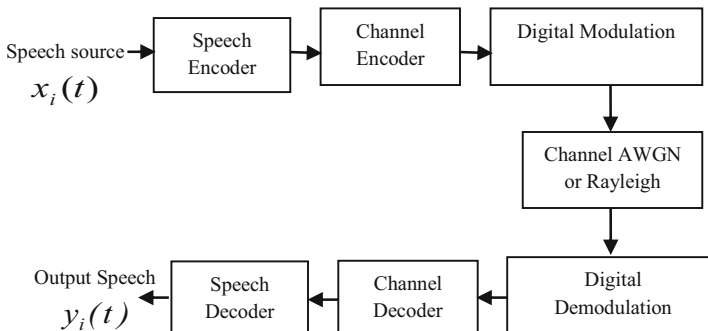


Fig. 1 General diagram of basic communication system [25]

ratio (SNR) value. To simulate the Fading channel, we apply a random signal envelope a and a random phase θ to the transmitted signal.

$$y(t) = ae^{-j\theta}x_i(t) + n(t) \tag{2}$$

when there is no dominant received component, the envelope is Rayleigh-distributed and it is, defined by:

$$p(a) = \frac{a}{\sigma^2} \exp\left(\frac{-a^2}{2\sigma^2}\right) \tag{3}$$

where $2\sigma^2 = E[a^2]$ is the mean power of the fading and the phase is uniformly distributed. In our investigation, we use the Rayleigh fading channel that has been shown realistic to simulate fading channels. In mobile environments, the speech is compressed by a conventional speech codec, then it is transmitted to the server where the recognition is performed using the features extracted from the decoded signal. The Rayleigh fading channel is simulated based on the modified sum-of-sinusoids method. The quadrature components of Rayleigh fading process are given by:

$$u(t) = \sqrt{\frac{2}{E}} \sum_{i=1}^E \cos(\varpi_d t \cos\alpha_i + \phi_i) + j\sqrt{\frac{2}{E}} \sum_{i=1}^E \cos(\varpi_d t \sin\alpha_i + \phi_i) \tag{4}$$

where $\alpha_i = ((2\pi i - \pi - \theta_i)/4\pi)$, $i = 1, 2, \dots, E$, ϖ_d is the maximum angular Doppler frequency, ϕ_i and θ_s are statistically independent and uniformly distributed on $[-\pi, \pi]$, [1].

3 Mixture linear prediction Gammatone features

Feature extraction is a crucial component in the Automatic Speaker Verification (ASV) system. Generally speaking, the speech features extraction methods aim at extracting relevant information about the speaker. In this work, we have implemented different feature extraction techniques that have in common the modeling of peripheral auditory system, namely MFCCs, GFCCs and the new feature MLPGCCs. The block diagram of feature extraction is depicted in Fig. 2.

3.1 Mixture linear prediction

The linear prediction (LP) analysis is used to estimate the parameters of an autoregressive (AR) model by minimizing the prediction error. In speech processing LP model, each sample is predicted as a linear weighted sum of the past p samples, where p is the order of prediction. The predicted signal $\hat{s}(n)$ is defined as:

$$\hat{s}(n) = - \sum_{k=1}^p a_k s(n-k) \tag{5}$$

In the mixture autoregressive model, the signal s_n , $n \geq 0$ can be modeled as a mixture of J autoregressive processes with conditional density function defined by [30].

$$f(s_n/s_{n-1}, \dots, s_0, \lambda) = \sum_{i=1}^J \pi_{n,i} \frac{1}{\sigma_i} \varphi\left(\frac{u_{n,i}}{\sigma_i}\right) \tag{6}$$

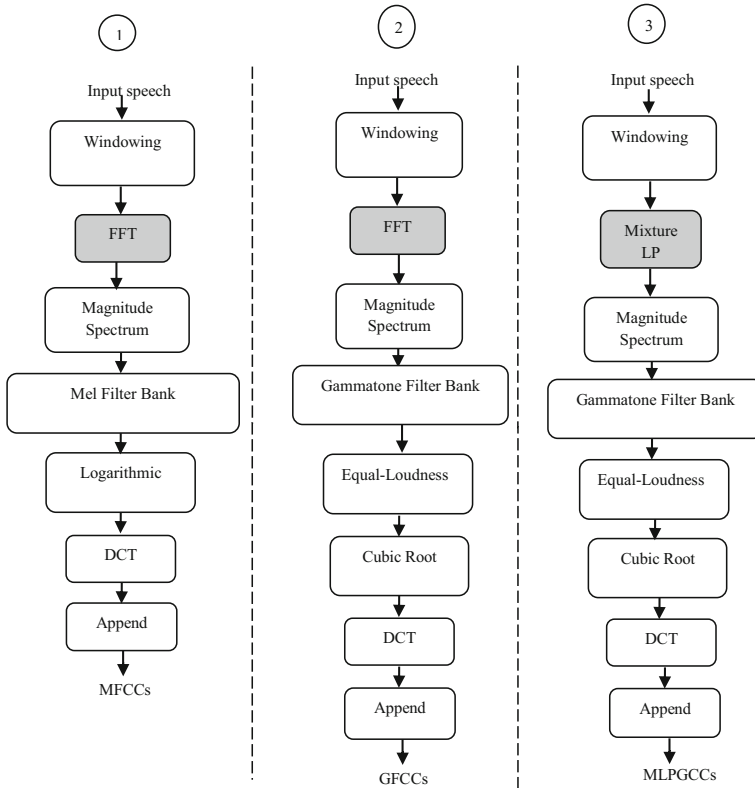


Fig. 2 Block diagram of (1) MFCCs, (2) GFCCs, and (3) MLPGCCs feature extraction

where λ is the model parameter set and $\varphi(\cdot)$ is the standard normal density function. The distribution of a hidden state variable is given by:

$$\pi_{n,i} = P(q_n = i / s_{n-1}, \dots, s_0, \lambda), 1 \leq i \leq J \tag{7}$$

where $q_n \in \{1, \dots, J\}$ determines the J of AR processes.

$$s_n = a_{0,i} + \sum_{k=i}^J a_{k,i} s_{n-k} + u_{n,i} 1 \leq i \leq J \tag{8}$$

where $a_{0,i}$ are the intercept (constant) terms. The mixture linear prediction is inspired by the principle of Gaussian Mixture Model (GMM), defined by the set of parameters:

$$\lambda_{GMM} = (P_1, \dots, P_J, \mu_1, \dots, \mu_j, \sigma_1^2, \dots, \sigma_j^2) \tag{9}$$

where P_i, μ_i and $\sigma_i^2, 1 \leq i \leq J$, are the component weights, Gaussian mean values and Gaussian variances, respectively. The mixture linear prediction (MLP) model is defined in as follows [28]:

$$\lambda_{MLP} = (P_1, \dots, P_J, a_{0,1}, a_{1,1}, \dots, a_{p,1}, a_{0,2}, \dots, a_{p,J}, \sigma_1^2, \dots, \sigma_J^2) \tag{10}$$

The parameters of this model are estimated by the Expectation - Maximization (EM) algorithm according to the following steps:

- In the E (expectation) step, estimate the excitations $u_{n,i}$ as a prediction residual

$$e_{n,i} = s_n - a_{0,i} - \sum_{k=1}^p a_{k,i} s_{n-k} \tag{11}$$

The hidden state posterior probabilities defined by

$$\gamma_{n,i} = P(q_n = i / s_n, \dots, s_{n-p}, \lambda_{GMLP}) = \max \left(0.01, \frac{P_i (1/\sqrt{2\pi\sigma_i^2}) \exp(-e_{n,i}^2 / (2\sigma_i^2))}{\sum_j P_j (1/\sqrt{2\pi\sigma_j^2}) \exp(-e_{n,j}^2 / (2\sigma_j^2))} \right) \tag{12}$$

- In the M (maximization) step, the component weights are re-estimated as $P_i = \frac{\sum_n \gamma_{n,i}}{\sum_n 1}$ and the noise variances as $\sigma_i^2 = \frac{\sum_n \gamma_{n,i} e_{n,i}^2}{\sum_n \gamma_{n,i}}$. To determine the AR parameters $a_{k,i}$ define $x_{n,0} = 1$ (for the intercept) and $x_{n,k} = s_{n-k}$, $k \geq 1$ and then solve the following normal equations:

$$\sum_{k=0}^p a_{k,i} \sum_n \gamma_{n,i} x_{n,k} x_{n,j} = \sum_n \gamma_{n,i} s_n x_{n,j}, 0 \leq j \leq p \tag{13}$$

3.2 Gammatone auditory filter bank

Gammatone filters are a popular way of modeling the auditory processing at the cochlea. The Gammatone function was first introduced in [12], characterizes physiological impulse-response data gathered from primary auditory fibers. The Gammatone filters were used for characterizing data obtained by reverse correlation from measurements of auditory nerve responses of the cat’s cochlea. The impulse response of a Gammatone filter centered at frequency f_c is defined as:

$$g(t) = K t^{(n-1)} e^{-2\pi B t} \cos(2\pi f_c t + \phi) \tag{14}$$

where K is the amplitude factor; n is the filter order; f_c is the central frequency in Hertz (Hz); ϕ is the phase shift; and B represents the duration of the impulse response. The Equivalent Rectangular Bandwidth (ERB) is a psychoacoustic measure of the auditory filter bandwidth at each point along the cochlea. The filterbank center frequencies are uniformly spaced on an equivalent rectangular bandwidth (ERB) scale between 200 and 3400 Hz (assuming a telephone bandwidth at a sampling rate of $F_s = 8$ kHz). The formula for calculating ERB (in Hz) at any frequency f (in Hz) is expressed by:

$$ERB = \frac{f}{Q_{ear}} + B_{min} \tag{15}$$

where $Q_{ear} = 9.26449$ and $B_{min} = 24.7$ are known as Glasberg and Moore parameters [8, 9]. The frequency response of the 64-channel Gammatone filter bank is illustrated in Fig. 3.

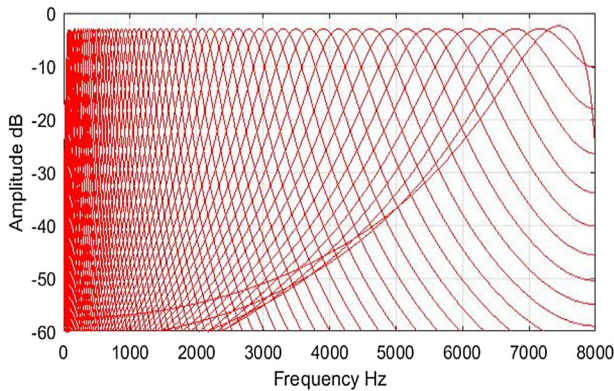


Fig. 3 A Gammatone filter bank with 64 filters

Herein, we used a bank of 64 filters whose center frequencies range from 50 Hz to 8000 Hz, relatively to the sampling frequency of the speech signal. The magnitudes of the down-sampled outputs are then loudness-compressed by a cubic root operation [42] such that:

$$G_m[i] = \left| |g|_{decimate} [i, m] \right|^{1/3} \quad (16)$$

$$i = 0, \dots, N-1, m = 0, \dots, M-1,$$

Here, $N = 64$ refers to the number of frequency (filter) channels. m is the frame index; M is the number of time frames obtained after decimation. The resulting responses $G_m[i]$ form a matrix representing the time-frequency (T-F) decomposition of the input signal. This T-F representation is a variant of cochleagram.

4 Speaker verification using mixture linear prediction Gammatone features

The investigated systems use the post feature extraction processing (MFCCs, GFCCs, and MLPGCCs) with i-vector extraction and channel compensation. The GPLDA technique is used to build the speaker model. The block diagram of the proposed speaker verification system is shown in Fig. 4.

4.1 Total variability i-vector modeling

Speaker verification based on i-vector approach involves different stages: i-vector feature extraction, GPLDA modeling and scoring using the batch likelihood ratio.

4.1.1 I-vector extraction

The i-vector approach [4] is inspired from the Joint Factor Analysis (JFA). In JFA, speaker and channel effects are independently modeled using Eigen-voice (speaker subspace) and Eigen-channel (channel subspace) models:

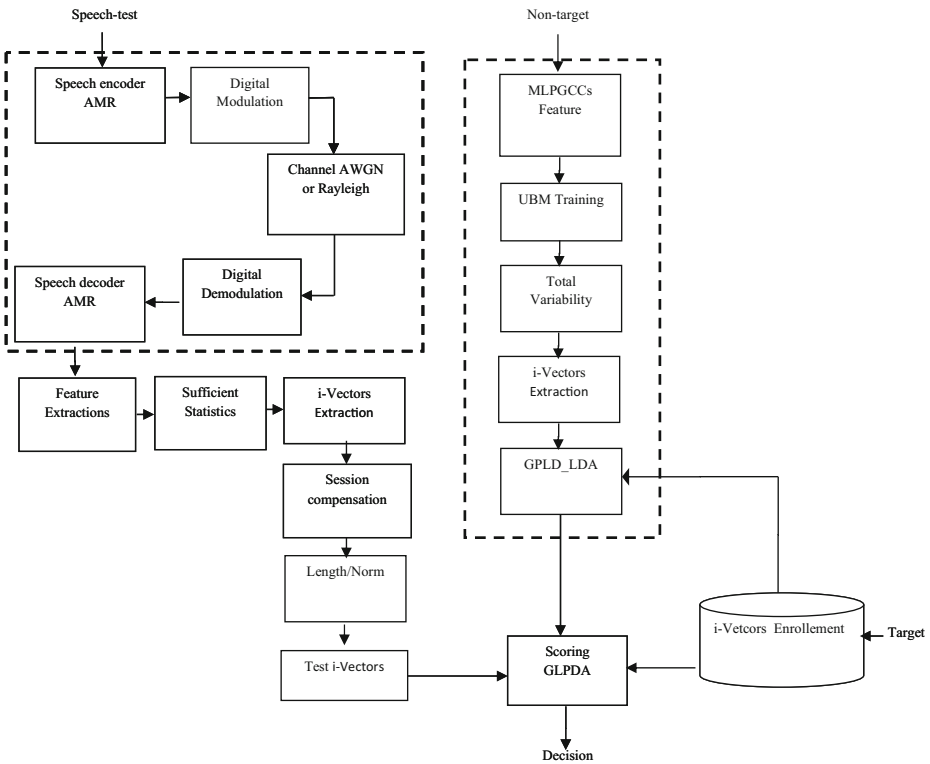


Fig. 4 Block diagram of i-Vector-GPLDA speaker verification system using MLPGCCs features employed in this study

$$M = m + Vy + Ux \tag{17}$$

where M is the speaker super-vector, m represents the speaker- and channel-independent super-vector, which can be taken to be the universal background model (UBM) super-vector. Both V and U are low rank transformation matrices. The variables x and y are assumed to be independent and have a standard normal distributions. In the i-vector extraction, the speaker and channel super-vector M is represented as:

$$M = m + Tw \tag{18}$$

where m is a speaker and channel independent super-vector, T is a low rank matrix representing the primary directions of variation across a large collection of development data, and w is a normally distribution with parameters $N(0, 1)$.

4.1.2 GPLDA modeling and scoring

The PLDA technique was originally proposed by [31] for face recognition, and later adapted to i-vectors for speaker verification by [15, 21]. This technique called Gaussian Probabilistic LDA (GPLDA), which divides the i-vector space into speaker and session variability subspaces, which has shown significant performance for intersession compensation for i-vector speaker verification [14]. In the GLPDA modeling approach, a speaker and channel dependent i-vector, $w_{s,r}$, can be defined as

$$w_{s,r} = \eta + Hz_s + \varepsilon_{s,r} \quad (19)$$

where η is the i-vectors' mean, H is the eigenvoice matrix, z_s is the speaker factor and $\varepsilon_{s,r}$ is the residual for each session.

The scoring in GPLDA is conducted using the batch likelihood ratio between a target and test i-vector [33]. Given two i-vectors, w_1 and w_2 , the batch likelihood ratio can be calculated as follows:

$$Score(w_1, w_2) = \log \frac{P(w_1, w_2 / \varphi_1)}{P(w_1, w_2 / \varphi_0)} \quad (20)$$

where φ_1 denotes the hypothesis that the i-vectors represent the same speakers and φ_0 denotes the hypothesis that they do not.

5 Evaluation experiments

The proposed features have been analyzed and evaluated by carrying out various experiments on the ASV. We use the NIST-2008 Speaker Recognition Evaluation (SRE) corpora containing a single channel microphone recorded conversational segments of 8 min or longer duration of the target speaker and an interviewer [23]. The speaker models were obtained from clean training speech data. The clean waveforms are transcoded by passing them through a coding and decoding AMRWB codec [35]. The mobile channel was simulated using two noise channels: AWGN and Rayleigh fading with different variances to make SNR within (-5, 0, 5, 10, 15 dB). In all experiments, the feature vectors contain 20 cepstral coefficients and log-energy/C0, appended with the first and second order time derivatives, thus providing 63 dimensional feature vectors, followed by cepstral mean and variance normalization (CMVN). The self-adaptive VAD (VQ-VAD) is employed to remove silence and low energy speech segments. We utilized three different acoustic features: (a) Mel frequency Cepstral coefficients (MFCCs), (b) Gammatone Frequency Cepstral Coefficients (GFCCs) as our baseline and (c) Mixture linear prediction Gammatone Cepstral Coefficients (MLPGCCs). The feature vector was extracted every 10 ms, using a Hamming window of 25 ms and the magnitude spectrum by FFT and MLP with (p : 8, 14, 20). After feature extraction, each speaker model is adapted from a 512-component in which the UBMs are trained using the entire database. For the total variability matrix training, the UBM training dataset is used. The EM training is performed throughout five iterations. We use 400 total factors (i.e., the i-vector size is 400) then LDA is applied to reduce the dimension of the i-vector to 200, and length normalization is then applied. In the process of variability compensation and scoring, a GPLDA model with adding noise is used. In practice, the MSR Identity Toolbox [38] was used to implement the i-vector-GPLDA processing. We evaluate the speaker verification accuracy using the equal error rate (EER).

5.1 ASV performance in additive white Gaussian noise (AWGN) channel

In this subsection, we investigate the effect of channel AWGN with different feature extractions (MFCC, GFCCs and MLPGCCs) on overall system performance. The first experiment aims to find the optimal number of MLP iterations to estimate the model of prediction.

Table 1 ASV performance in terms of EER (%) under AWGN channel using different SNR for the features(MFCCs, GFCCs) and proposed MLPGCCs with differents number of prediction ($p = 8, 14, 20$)

SNR (dB)	EER(%)				
	-5	0	5	10	15
MFCCs	18.8	14.62	10.5	7.55	5.45
GFCCs	13.31	10.14	7.19	4.9	2.85
MLPGCCs($p = 8$)	15.10	13.29	9.14	5.51	3.89
MLPGCCs($p = 14$)	11.30	9.03	5.54	4.28	2.96
MLPGCCs($p = 20$)	9.59	7.58	4.12	3.15	2.35

Figure 6 shows EER as a function number of MLP iterations. We can see an improvement of the accuracy when the iteration number is set to a specified values (5, 7). We found that the optimal number of MLP iterations is 5 compared to 7 as found in [30].

The goal of the next experiments is to evaluate the verification performance using MFCCs, GFCCs, MLPGCCs features and the combination of MLPGCCs-GFCCs features. Here, we consider the context of mismatched conditions where the test data is distorted with AWGN noise having a 5 dB of SNR level. The results obtained by using a development set and i vector GPLDA are displayed in Tables 1, 2 and Fig. 5. It can be observed from Table 1 that the proposed features perform better than GFCCs and MFCCs at almost all SNR levels and clean condition. We also note that the MLPGCCs with $p = 20$ slightly outperform the method presented in [29]. This can be explained by the fact that for the LP orders ranged in 8–20, the LP residual contains mostly the information about the excitation source.

From Fig. 5, it is clear that EER decreases with respect to signal to noise ratio for all the features. From the results it can be shown that MLPGCCs with (5-iterations) gives better correct recognition rate compared to other features at all SNR levels.

Results in Table 1 and Fig. 6 indicate that the proposed feature extraction method achieves a reduction in average equal error rate (EER) ranging from 9.41% to 6.65% and 3.72% to 1.50% compared with MFCCs and GFCCs features, when the test speech signals are corrupted with Additive White Gaussian Noise (AWGN) channel, at SNRs ranging from (-5 dB to 15 dB) respectively.

Furthermore, we combined two methods for estimating the short-term spectrum, namely FFT and mixture LP to improve the (EER). This combination is performed by the logistic regression technique where weights are trained using the BOSARIS Toolkit [3]. The effect of this combination of GFCCs and MLPGCCs is investigated under AWGN noise channel. The results are summarized in Table 2. In comparison

Table 2 ASV performance in terms of EER (%) under AWGN channel using different SNRs for the features (GFCCs, MLPGCCs) and combined features

Noise	Features	-5 dB	0 dB	5 dB	10 dB	15 dB
Awgn channel	GFCCs	13.31	10.14	7.19	4.9	3.85
	MLPGCCs	9.59	7.58	4.12	3.15	2.35
	Fusion	9.16	6.68	3.45	2.34	1.76

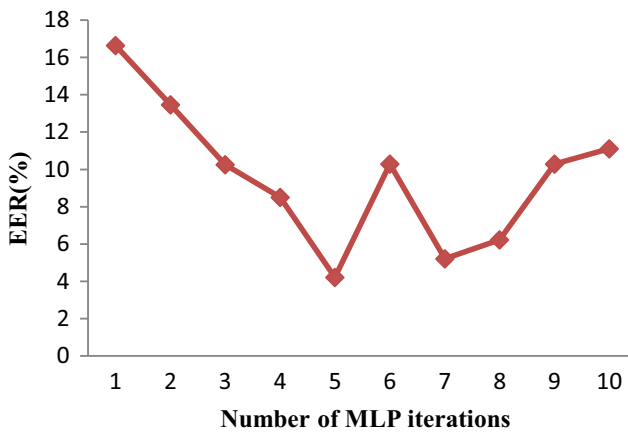


Fig. 5 The effect of varying the number of MLP iterations on the performance of speaker verification

with the results obtained by MLPGCCs based system as shown in Table 1 and Fig. 6, we observe significant EER reduction in all SNR levels between 0.43% to 0.59%.

5.2 ASV performance in Rayleigh fading channel

In the case of Rayleigh fading channel, we have carried out the same processing as in the case of AWGN channel distortion. The results obtained on the development set for different features using i-vector GPLDA are shown in Fig. 7. It can be seen that there is a drop in the accuracy of the verification system as the SNR increases. It can also be noticed that there is an accuracy improvement for MLPGCCs features compared to MFCCs and GFCCs features. Moreover, it can be shown that MLPGCCs with (7-iterations) gives better recognition rate compared with the (5-iteration) MLPGCCs.

As result, when the test speech signals are corrupted with Rayleigh fading channel indicate that the proposed feature extraction method achieves a reduction in the average equal error rate (EER) ranging from 23.63% to 7.8% and from 10.88% to 6.8% over conventional MFCCs and GFCCs features, at SNRs ranging from (-5 dB to 15 dB), respectively. In addition, Table 3 summarizes the results of GFCCs and MLPGCCs combination under Rayleigh fading channel.

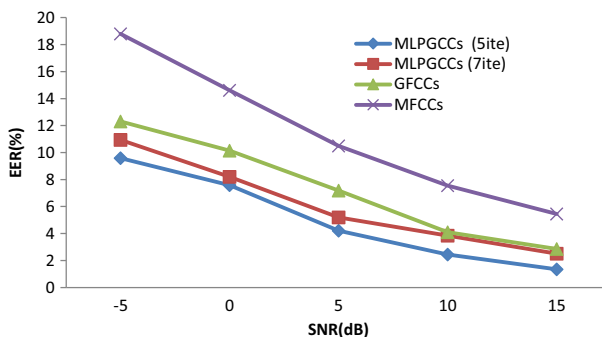


Fig. 6 Performance comparison of alternative noise-robust features considered in this study against MFCCs on the ASV task under AWGN noise channel at different SNR

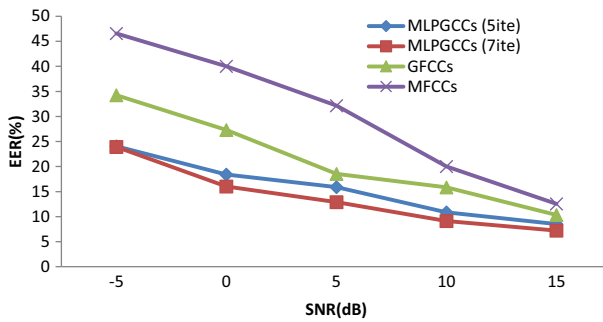


Fig. 7 Performance comparison of alternative noise-robust features considered in this study against MFCCs on the ASV task under Rayleigh fading channel using different SNR

These results have shown a significant EER(%) reduction at all SNR levels between 1.92% to 3.88%.

6 Conclusion

In this paper, a new feature extraction method based on the mixture linear prediction Gammatone is proposed. The MLPGCCs features are evaluated on a speaker verification system using i-vector GPLDA modeling in mobile communications with considering the impact of transmission channel distortion. The key point of our idea is to take advantage of the characteristics of the linear prediction approach by using the iterative parameter re-estimation of a mixture autoregressive (AR) model, instead of using standard spectrum estimation performed by FFT. The new features are evaluated on NIST 2008 dataset by considering the effects of noisy transmission channel (AWGN and Rayleigh fading). Experimental results show that the proposed MLPGCCs outperform the conventional MFCCs and GFCCs features in speaker verification task. The best performance is obtained in the context of AWGN channel (vs. Rayleigh fading channel). The combination of the proposed and conventional features achieves better performance when compared with each system alone and data corrupted by transmission channel noise. The results have shown that the proposed MLPGCCs considerably improved the robustness in all types of channel distortions. We have also demonstrated that the algorithm that uses Gammatone filter bank and mixture linear prediction is suitable in the context of transmission channel noise compared to the FFT and the Mel-filterbank. Future research includes the study of MLPGCCs system performance under other types of noise and degradations such as: convolutive noise and reverberation.

Table 3 ASV performance in terms of EER (%) under Rayleigh fading channel using different SNRs and features (GFCCs, MLPGCCs) and combined features

Noise	Features	-5 dB	0 dB	5 dB	10 dB	15 dB
Rayleigh fading channel	GFCCs	34.25	27.03	18.55	15.86	14.00
	MLPGCCs	23.37	16.02	12.91	9.12	7.2
	Fusion	21.45	12.56	10.64	7.34	3.32

References

- Al-Momani O, Gharaibeh KM (2014) Effect of wireless channels on detection and classification of asthma attacks in wireless remote health monitoring systems. *Int J Telemed Appl*:1–8
- Apsingekar VR, De Leon PL (2011) Speaker verification score normalization using speaker model clusters. *Speech Communication*, Elsevier Science vol 53, pp 110–118
- Brummer N, Villiers ED (2011) The BOSARIS toolkit: theory, algorithms and code for surviving the new DCF. In: NIST SRE11 Analysis Workshop, Atlanta (USA), Dec. 2011, pp:1–23 [Online]. Available : <https://sites.google.com/site/nikobrummer/bosaris/toolkit/fullpaper.pdf>
- Dehak N et al (2011) Frontend factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing* 19(4):788–798
- Fedila M, Amrouche A (2012) Automatic speaker recognition for mobile communications using AMR-WB speech coding. *IEEE, information science, signal processing and their applications*, ISSPA, pp 1034–1038
- Fedila M, Bengherabi M, Amrouche A (2017) Gammatone filterbank and symbiotic combination of amplitude and phase-based spectra for robust speaker verification under noisy conditions and compression artifacts. *Multimedia Tools Appl*:1–19
- Gallardo LF (2016) Human and automatic speaker recognition over telecommunication channels. Springer Science + Business Media, Singapore
- Glasberg BR, Moore BCJ (1986) Auditory filter shapes in subjects with unilateral and bilateral cochlear impairments. *J Acoust Soc Am* 79:1020–1033
- Glasberg, Moore (1990) Derivation of auditory filter shapes from notched-noise data. *Journal of Hearing Elsevier*, vol 47, issues 1–2, pp 103–138
- Hansen JHL, Hasan T (2015) Speaker recognition by machines and humans: a tutorial review. *IEEE Signal Process Mag* 32(6):74–99. <https://doi.org/10.1109/MSP.2015.2462851>
- Jeevan M, Dhingra A, Hanmandlu M, Panigrahi BK (2017) Robust speaker verification using GFCC based i-vectors. In: Proceedings of the international conference on signal, networks, computing, and systems. Springer, New Delhi, pp 85–91
- Johannesma PIM (1972) The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: Symposium on hearing theory (IPO, Eindhoven, The Netherlands), pp 58–69
- Kaled Dagrouq A, Alkhateeb (2013) Wavelet LPC with neural network for speaker identification system. *Wseas Transactions on Signal Processing* 9:216–226
- Kanagasundaram A (2018) Improving the performance of GPLDA speaker verification using unsupervised inter-dataset variability compensation approaches. *Int J Speech Technol* 21:533–544
- Kenny P, Stafylakis T, Ouellet P, Alam J, Dumouchel P (2013) PLDA for speaker verification utterances of arbitrary duration. In: Proceedings of IEEE international conference on acoustics, speech signal processing, pp 7649–7653
- Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: From features to supervectors. *Speech Commun* 52:12–40
- Krobba A, Debyeche M, Amrouche A (2010) Evaluation of speaker identification system using GSM-EFR speech data. In: proc. of int. conf. on design and technology of integrated systems (nanoscale era), Hammamet, DTIS, Tuins, IEEE, pp 1-5
- Krobba A, Debyeche M, Selouani SA (2018) Feature extraction using mixture linear prediction Gammatone filter for robust speaker verification over AWGN Channel. 4th International Conference on Signal, Image, Vision and their Applications, Guelma – Algeria, 26–27 November
- Li Z, Gao Y (2015) Acoustic feature extraction method for robust speaker identification. *Multimed Tools Appl* 75(12):7391–7406
- Li X, Wang L, Zhu J (2017) SNR-multicondition approaches of robust speaker model compensation based on PLDA in practical environment. In Proceedings on the international conference on artificial intelligence (ICAI) (pp. 146-150). The Steering Committee of the World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp)
- Mak M-W, Pang X, Chien J-T (2016) Mixture of PLDA for noise robust I-vector speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(1):130–142
- Ming J, Hazen TJ, Glass JR, Reynolds DA (2007) Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing* 15(5):1711–1722
- NIST Year (2008) Speaker recognition evaluation plan, Technical report, NIST. <http://www.itl.nist.gov/iad/mig/yest/ser/2008>
- Padilla M, Quatieri T, Reynolds D (2006) Missing feature theory with soft spectral subtraction for speaker verification. CSLP, ninth international conference on spoken language processing, Pittsburgh, PA, USA
- Pahlavan K, Krishnamurthy P (2011) Principles of wireless networks: a unified approach. Prentice Hall PTR

26. Pawan K. A, Navnath S. Nehe · Dattatray V. Jadhav · Raghunath S. H, (2012). Robust feature extraction from: spectrum estimated using bispectrum for speaker recognition. *Int J Speech Technol* 15, pp:433–440.
27. Pelecanos J, Sridharan S (2001) Feature warping for robust speaker verification. In: Proc. ISCA speaker recognition workshop Odyssey, Chania, Crete, pp 213–218
28. Pohjalainen J, Alku P (2014) Gaussian mixture linear prediction. *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp 6285–6289
29. Pohjalainen J, Saeidi R, Kinnunen T, Alku P (2010) Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions. In: Proc Interspeech, Japan
30. Pohjalainen J, Cemal H, Kinnunen T, Alku P (2014) Mixture linear prediction in speaker verification under vocal effort mismatch. *IEEE Signal Process Lett* 21(12):1516–1520
31. Prince Simon JD, Elder JH (2007) Probabilistic linear discriminant analysis for inferences about identity. *IEEE 11th international conference on computer vision. ICCV'07*, pp 1–8
32. Rahman MH, Kanagasundaram A, Himawan I, Dean D, Sridharan S (2018) Improving PLDA speaker verification performance using domain mismatch compensation techniques. *Comput Speech Lang* 47:240–258
33. Rao W, Mak MW (2013) Boosting the performance of i-vector based speaker verification via utterance partitioning. *IEEE Trans Audio Speech Lang Process* 21(5):1012–1022
34. Ravindran S, Anderson DV, Slaney M (2006) Improving the noise robustness of mel-frequency cepstral coefficients for speech processing. In: Proc. ISCA SAPA. Pittsburgh, PA, pp 48–52
35. Recommendation G (2003) 722.2: wideband coding of speech at around 16 kbit/s using adaptive MultiRate wideband (AMR-WB)
36. Riadh A, Salim S, Said G, Ali CA, Taleb-A (2014) An efficient approach for MFCC feature extraction for text Independent speaker identification system. *Int J Commun* 9:114–122
37. Saeidi R, Pohjalainen J, Kinnunen T, Alku P (2010) Temporally weighted linear prediction features for tackling additive noise in speaker verification. *IEEE Signal Process Lett* 17(6):599–602
38. Seyed OS, Malcolm S, Heck L (2013) MSR identity toolbox v.1.0.A MATLAB toolbox for speaker recognition research In: Proc, IEEE Signal Process, Speech and Language Processing Technical Committee Newsletter
39. Sreenivasa R K, Vuppala AK (2014) *Speech processing in mobile environments*. Springer, ISBN: 978–319–03116-3
40. Y. Zhang, Y. Long· X. Shen, H. Wei, M. Yang, H. Ye, H. Mao, (2017). Articulatory movement features for short-duration text dependent speaker verification. *Int J Speech Technol* 20, 753–759.
41. Zhao X, Wang D (2013) Analyzing noise robustness of MFCC and GFCC features in speaker identification. *ICASSP'13*, pp 7204–7208
42. Zhao X, Shao Y, Wang DL (2012) CASA-based robust speaker identification. *IEEE Transactions on Audio, Speech, and Language Processing* 20(5):1608–1616

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Ahmed Krobb¹ · Mohamed Debyeche¹ · Sid-Ahmed Selouani²

¹ Université des Sciences et de la Technologie Houari Boumediene (USTHB), LCPTS, Algiers, Algeria

² LARIHS Laboratory, Campus Shappaing, University of Moncton, Moncton, Canada