Multimedia Tools and Applications (2023) 82:16195–16212
https://doi.org/10.1007/s11042-022-14068-4

A novel hybrid feature method based on Caelen auditory model and gammatone filterbank for robust speaker recognition under noisy environment and speech coding distortion

Ahmed Krobba[1] · Mohamed Debyeche[1] · Sid. Ahmed Selouani[2]

Received: 16 February 2021 / Revised: 27 September 2022 / Accepted: 9 October 2022 /
Published online: 27 October 2022

Abstract
Currently, the majority of the state-of-the-art speaker recognition systems predominantly use short-term cepstral feature extraction approaches to parameterize the speech signals. In this paper, we propose new auditory features based Caelen auditory model that simulate the external, middle and inner parts of the ear and Gammtone filter for speaker recognition system, called Caelen Auditory Model Gammatone Cepstral Coefficients (CAMGTCC). The performances evaluations of the proposed feature are carried by the TIMIT and NIST 2008 corpus. The speech coding represent by Adaptive Multi-Rate wideband (AMR-WB) and noisy conditions using various noises SNR levels which are extracted from NOISEX-92. Speaker recognition system using GMM-UBM and i-vector-GPLDA modelling. The experimental results demonstrate that the proposed feature extraction method performs better compared to the Gammatone Cepstral Coefficients (GTCC) and Mel Frequency Cepstral Coefficients (MFCC) features. For speech coding distortion, the features extraction proposed improve the robustness of codec-degraded speech at different bit rates. In addition, when the test speech signals are corrupted with noise at SNRs ranging from (0 dB to 15 dB), we observe that CAMGTCC achieves overall equal error rate (EER) reduction of 10.88% to 6.8% relative, compared to baselines.

✉ Ahmed Krobba
hkrobba@gmail.com

Mohamed Debyeche
mdebyeche@usthb.dz

Sid. Ahmed Selouani
selouani@umcs.ca

[1] Speech Communication and Signal Processing Laboratory, Faculty of Electrical engineering, University of USTHB, Algiers, Algeria

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

Keywords Speaker recognition · Caelen auditory model · Gammtone filter · Speech coding · Noise environment · GMM-UBM · I-vetcor G-PLDA

1 Introduction

The rapid growth of mobile phones today poses many challenges for voice technologies, in particular Speaker Recognition (SR). The major challenges in speaker recognition for mobile environment and effect of environmental noise are degradations introduced by the speech codec and transmission channel [10, 43]. Speaker recognition is a branch of biometry. It is defined by two tasks: speaker identification (SID) and speaker verification (SV) [18]. In speaker identification, the goal is to identify the speaker of an utterance from a given population and speaker verification is the process of authentication of a person's claimed identity by analyzing his/her speech signal. In general, robust SR system is based on three stages: the first is feature extraction where the speech signal is represented in a compact manner, the second is speaker modeling which can be defined as a process of describing feature properties for a given speaker, and the last stage is scoring or decision [14]. Speech codec can degrade the recognition performance in two different cases. In the first case, speech codec can be degraded by the compression itself, which degrades the speech quality and hence the recognition performance. The second case for performance reduction of the recognition system is given by the difference between the training and test conditions [5]. 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 and database access services [14]. Speaker recognition is easy to use, has low computation requirements (can be ported to cards and handhelds) and, given appropriate constraints, has high accuracy. Speaker recognition, as all biometrics, has limitations pour certain application. There are limitations in the software, it does not always work across all operating systems. Speaker recognition embedded is refers to a technique in which all speech coding processing, feature extraction, and recognition are performed in the mobile device. The most important disadvantage for the embedded system is that the resources are very limited on the mobile device [34].

1.1 Related work

The most recent research in speaker recognition performance focuses on the background noise and impact of speech coding. Many research works have been reported in the literature to reduce the impact of speech coding distortion for SR system McCree [26] and Vuppala et al. [48]. The effect of Global System for Mobile (GSM) coding is examined in Grassi.S et al. [13] and Krobba. A et al. [19]. In Dunn et al. [7], they used different standard speech coders (G.729, G.723, MELP) are used to evaluate speaker recognition performances under matched and mismatched conditions. Methods for extracting the features directly from the coded speech were proposed in Fedila.M and Amrouche.A [8]. McLaren et al. [27] analyzed several acoustic features to examine the robustness of speaker recognition. Krobba. A et al. [20] proposed a new framework based on Maximum Entropy (ME) and Probabilistic Linear Discriminate Analysis (PLDA) to improve the performance of speaker identification system in the presence of speech coding distortion. In noisy environments, the additive noises affect the signal spectrum. This results in the appearance of certain peaks that do not exist in the original

signal by the disappearance of certain peaks of the original signal and the flattening of the spectral envelope (loss of information). These noises result in the loss of speech intelligibility and quality, imposing great challenges on speaker recognition systems. Many different compensation strategies have been proposed to reduce the impact of noisy environment such as speech enhancement, feature compensation, robust feature extraction, robust modeling and score compensation. In the compensation of noise, the simplest solution would be to utilize speech enhancement (SE) technique as a pre-processing block for the SR system Olivier Bellot et al. [2]. Spectral subtraction (SS) method reduces the noise spectrum from the noisy speech spectrum to estimate the clean speech spectrum Chandra, M et al. [4], Minimum Mean Square Error (MMSE), and subspace-based speech enhancement techniques Yu, D et al. [49]. The robust feature extraction based Cepstral Mean Subtraction (CMS) is the most popular method employed to ameliorate the effects of channel variability Shabtai, N. R et al. [42]. RASTA filtering, feature warping [17], Mean and Variance Normalization (MVA) processing, and nonlinear spectral magnitude normalization used to improve the recognition performance in presence of convolution distortions and additive noise [28]. Samia Abd El-Moneim et al. [36] proposed a text- independent speaker recognition system based on Long-Short Term Memory Recurrent Neural Network (LSTM-RNN) classifier. MFCC extracted from the Discrete Wavelet Transform (DWT) of the speech signal, with and without feature warping were proposed in [1]. The spectrum estimation methods, for example Weighted Linear Prediction (WLP), Stabilized Weighted Linear Prediction (SWLP), Regularization of Linear Prediction (RLP) and Gaussian Mixture Linear Prediction (GMLP) are based on MFCC feature [31, 35]. In [22], the authors have demonstrated that Gammatone feature GFCC processing provided substantial improvements in recognition accuracy in the presence of various types of additive noise. In [9], the authors have proposed to use the Gammatone product-spectrum cepstral coefficients under noisy condition and speech codecs. In [21] used a mixed method based on the multitaper gammatone Hilbert envelope coefficients (MGHECs) and multitaper chirp group delay zeros-phase Hilbert envelope coefficients (MCGDZPHECs) is used. The great majority of past studies have addressed the effect of speech coding and additive noise environment for speaker recognition to develop the robust feature extraction. However, only few studies have been reported the impact of the complexity of human perception mechanism on the performance of speaker recognition systems.

1.2 Motivation and contribution

A large majority of speaker recognition systems is based on low-level features which convey physiological information about the speaker. This set of feature extractions can be modeled by two ways: modeling the human voice production or modeling the peripheral auditory system. The first way is generally based on the source-filter model, which leads to the extraction of features such as linear coding (LPC). The second takes the mechanism of the auditory, Mel-Frequency Cepstral Coefficients (MFCC) which have been the most widely used features for speaker and speech recognition tasks. The auditory model used in MFCC is not optimal for speaker recognition. The logarithmic nonlinearity used in MFCC feature to compress the dynamic range of filter-bank energies is not immune to distortions of speech spectra caused by a background noise. On the other hand, these acoustic feature extraction methods remain largely ineffective and fail to provide satisfactory robustness for speaker recognition system because spectral information includes a lot of redundant information and the complexity of the human perception mechanism.

A number of electrical analogues of the auditory model have been developed to estimate the displacement basilar membrane Seneff [40], Lyon [25] and Ghitza [11]. Zhao X et al. In [50] proposed novel auditory feature based gammatone (GT), inspired by Auditory Scene Analysis (ASA) research, Computational Auditory Scene Analysis (CASA). Li et al. [23] proposed an auditory-based feature, Cochlear Filter Cepstral Coefficients (CFCC), based on time-frequency transform plus a set of modules to simulate the signal processing functions in the cochlea. Xavier. V and Francesc. A [46] introduced a novel feature extraction method, the Gammatone cepstral coefficients (GTCC) are a biologically inspired modification employing Gammatone filters with equivalent rectangular bandwidth bands. Our work is inspired by previous works that suggested the Gammatone Cepstral Coefficients (GTCC) which is based on an auditory periphery model of the speech features to noisy environments that is significantly better than MFCC. In this paper, we extend that work by integrating in the front-end of the speaker recognition system based on auditory model to incorporate both hearing/perception and phonetic/phonological knowledge, the auditory model which was first proposed by Caelen. J [3] and adapted to be used as a front-end module in speech and speaker recognition systems by Selouani [37, 38] and Kamil Lahcene Kadi et al., [16]. Our contributions can be concluded as follows.

- Based on the GTCC feature, we design novel feature extraction methods based on Caelen auditory model and cochlear gammatone filterbank.
- We introduced the speaker recognition system in the client-server architecture of the mobile network and simulation of mobile environment by noisy environment and speech coding distortion.

- We provide an experimental evaluation with the proposed feature and the total variability i-vector G-PLDA modeling to improve speaker recognition system performance

The paper is structured as follows. Section 2 gives an overview of speaker recognition over mobile communication. In Section.3, we describe the proposed CAMGTCC feature. Experimental setup is given in Section 4. Results and discussion are presented in Section 5. Conclusions and future work directions are provided in Section 6.

2 Overview of speaker recognition over mobile communication

In mobile communication, the SR system is developed in two architectures such as Network speaker recognition (NSR) and Embedded speaker recognition (ESR) [44]. In NSR, where speech is transmitted over the communication channel and the recognition is performed on the remote server (Fig. 1). This technique makes it possible to consider the use of much more powerful servers and therefore provide more diverse and generally better quality services. In ESR both front-end and back-end are implemented on the terminal. The most important disadvantage for the embedded system is that the resources are very limited on the mobile device.

2.1 Speech coding

Speech coding has been used in digital communication system; mainly to remove the maximum redundancy in the speech signal while maintaining a quality in the decoded

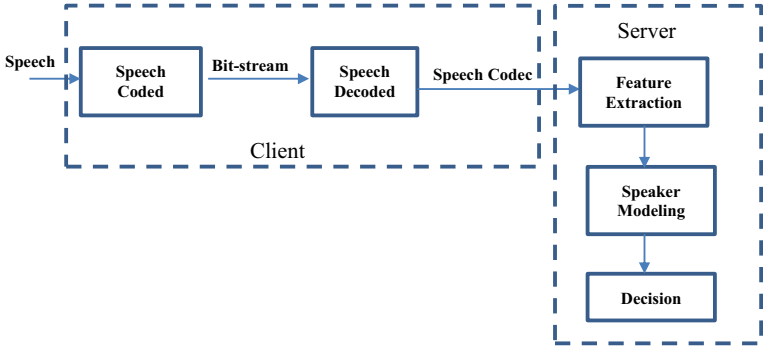

Fig. 1 NSR System architecture

speech signal that is acceptable for the applications [6]. Reconstitution of the speech signal is done from the parameters of the model production speech generally (source / excitation). We have two steps in the speech coding, the first step: analysis of speech for extraction of LPC (Liner- prediction) and pitch parameters. The second step: the speech synthesis using these parameters for speech signal encoding. Fig. 1 represents a simple block diagram of speech codec, this coder constitutes of two main blocks: A speech coder represents the analysis of speech using the input speech signal to produce the Bit-stream and another speech coder represents the speech synthesis. The bit-stream is used as input to the block speech decoder to produce the output speech signal. The codec used in this work is Adaptative Multi- Wideband Rate (AMR-WB G.722.2) speech coding standard based on ACELP speech [5]. It was selected as ITU-T recommendation G.722.2 and it operates on speech of extended bandwidth ranging from 50 hz to 7 khz.with bit-rates of 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.60 kbp. AMR-WB codec characterized by a Voice Activity Detector (VAD) and Discontinuous Transmission (DTX) function to improve channel capacity and provides better speech quality [33].

2.2 Background noise

In practical applications of speaker recognition in mobile communication, noise is defined as a phenomenon that prevents the transmission of a message from a source to its destination or anything that deteriorates the quality and intelligibility of the transmitted message. Noise directly affects the signal spectrum, which results in the appearance of certain peaks that do not exist in the original signal, by the disappearance of certain peaks of the original signal and the flattening of the spectral envelope (loss of information) [30].

The most common source of noise is the background noise and noise can be classified into a number of categories such as.

1. Noise from industrial systems: these correspond to noise emitted by machines with poor sound insulation.
2. Noise from means of transportation: these correspond to the noise that can be observed in various vehicles such as cars, trains or planes.
3. Noise from administrative and urban environments: the type of noise present in offices, homes or in urban concentrations

3 The feature extraction method based on Caelen auditory model and Gammatone filterbank

Feature extraction is a crucial component in the ASR 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 MFCC, GTCC and the new feature CAMGTCC. The block diagram of feature extraction is depicted in Fig. 4. The proposed feature extraction methods are obtained from gammatone filter-bank. Gammatone filters are a popular way of modeling the auditory processing at the cochlea. The Gammatone function was first introduced by Johannesma.P [15]. Gammatone filters were used for characterizing data obtained by reverse correlation from measurements of auditory nerve responses of cats [12]. The impulse response of a Gammatone filter centered at frequency f_c is:

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

where K is the amplitude factor; n is the filter order; f_c is the central frequency in Hertz; ϕ is the phase shift; and B represents the duration of the impulse response. The filters are placed in equal distance in frequency according to the Equivalent Rectangular Bandwidth (ERB).

The ERB filter models the spectral integration derived from the channeling effectuated by the inner hair cells, the ERB is defined by

$$ERB = \left[\left(\frac{f_c}{EarQ}\right)^p + (\min BW)^p\right]^{1/p} \tag{2}$$

where $EarQ$ is the asymptotic filter quality at high frequencies, bandwidth at low frequencies and p is commonly 1 or 2, min$BW$is the minimum.

3.1 Caelen auditory model (CAM)

Caelen Auditory Model (CAM) consists of three parts which simulate the behavior of the ear, [3]. The extern and middle ear are modeled using a band pass filter, which can be expressed as follows

$$s'(k) = s(k) - s(k-1) + \alpha_1 s'(k-1) + \alpha_2 s'(k-2) \tag{3}$$

where $s(k)$ is the speech signal, $s'(k)$ is the filtered signal, $k = 1,\ldots, K$ is the time index and K is the number of frame-samples. The coefficients $\alpha 1$and α_2 depend on the sampling frequency Fs, the central frequency of the filter and its Q-factor [39]. The next part of the model simulates the behavior of the basilar membrane (BM), the most important part of the inner ear that acts substantially as a non-linear filter bank. The output of each filter is given by:

$$y_i(k) = \beta_{1,i} y_i(k-1) + \beta_{2,i} y_i(k-2) + G\left[s'(k) - s'(k-2)\right] \tag{4}$$

and its transfer function can be written as

$$H_i(z) = \frac{G_i[1 - z^{-2}]}{1 - \beta_{1,i} z^{-1} + \beta_{2,i} z^{-2}} \tag{5}$$

where $y_i(k)$ represents the vibration amplitude at position x_i of the BM and constitutes the BM response to a mid-external sound signal $s(k)$. The G_i, $\beta_{1,i}$ and $\beta_{2,i}$, parameters represent the gain and coefficients, respectively, of the filter i. Nc is the number of overlapping cochlear filters or channels and is set to 24 in our case. The absolute energy in the output of each channel was calculated as follows:

$$W_i^{'}(T) = 20\log \sum_{k=1}^{K} |y_i^{'}(k)| \text{ where } i = 1, 2.\ldots\ldots N_c \tag{6}$$

T refers to the frame index i refers to the channels and N_c is the total number of channels that is 24; k denotes samples and therefore K is the frame length. A smoothing function is applied in order to reduce the energy variations:

$$W_i(T) = c_0 W_i(T-1) + c_i W_i^{'}(T) \tag{7}$$

where Wi(T) is the smoothed energy, the coefficients c0 and c1 averaging Wi(T − 1) and Wi'(T). The acoustic features based on the Caelen ear model were calculated after performing linear combinations of energies of the channel outputs. Each feature is computed based on the output of the 24 channel filters of the above-mentioned ear model.

In this work, we extracted seven acoustic features which are: Acute/grave (AG), open/closed (OC), diffuse/compact (DC), sharp/flat (SF), mat/strident (MS), continuous/discontinuous (CD) and tense/lax (TL) given in Table 1 [38]. In the Figs. 2 and 3 examples of the acoustic feature based on clean auditory model computed from speech coding and noisy speech are given.

We conclude from Figs. 2 and 3 that the acoustic feature derived from the Caelen auditory model with a noisy speech give the best representation of auditory model compared to the acoustic feature derived from speech coding. Figure 4 illustrates the

Table 1 Descriptions of the acoustic feature based clean auditory model

Acoustic feature	Description
(G/A)	Measures the difference of energy between low frequencies (50–400 Hz) and high frequencies (3800–6000 Hz): $(W1+\cdots+W5)-(W20+\cdots+W24)$
(O/C)	A phoneme is considered closed if the energy of low frequencies (230–350 Hz) is greater than that of the middle frequencies (600–800 Hz). Hence, the O/C cue is calculated by: $W8+W9-W3-W4$
(D/C)	Compactness reflects the prominence of the central formant region (800–1050 Hz) compared with the surrounding regions (300–700 Hz) and (1450–2550 Hz): $W10+W11-(W4+\cdots+W8+W13+\cdots+W17)/5$
(F/S)	A phoneme is considered sharp if the energy in (2200–3300 Hz) is more important than the energy in (1900–2900 Hz): $W17+W18+W19-W11-W12-W13$
(M/S)	Strident phonemes are characterized by a presence of noise because of a turbulence at their articulation point which leads to more energy in (3800–5300 Hz) than in (1900–2900 Hz): $W21+W22+W23-W16-W17-W18$
(C/D)	Quantifies the variation of the spectrum magnitude by comparing the energy of current and preceding frames. $N\,i{=}c1\lvert Wi(T)-Wa(T)-Wi(T-1)+Wa(T-1)\rvert$ $Wi(T)$ is the energy of channel i $Wa(T)$ is the energy average over all channels of current frame T
(T/L)	Measures the difference of energy between middle frequencies (900–2000 Hz) and relative high frequencies (2650–5000 Hz): $(W11+\cdots+W16)+(W18+\cdots+W23)$

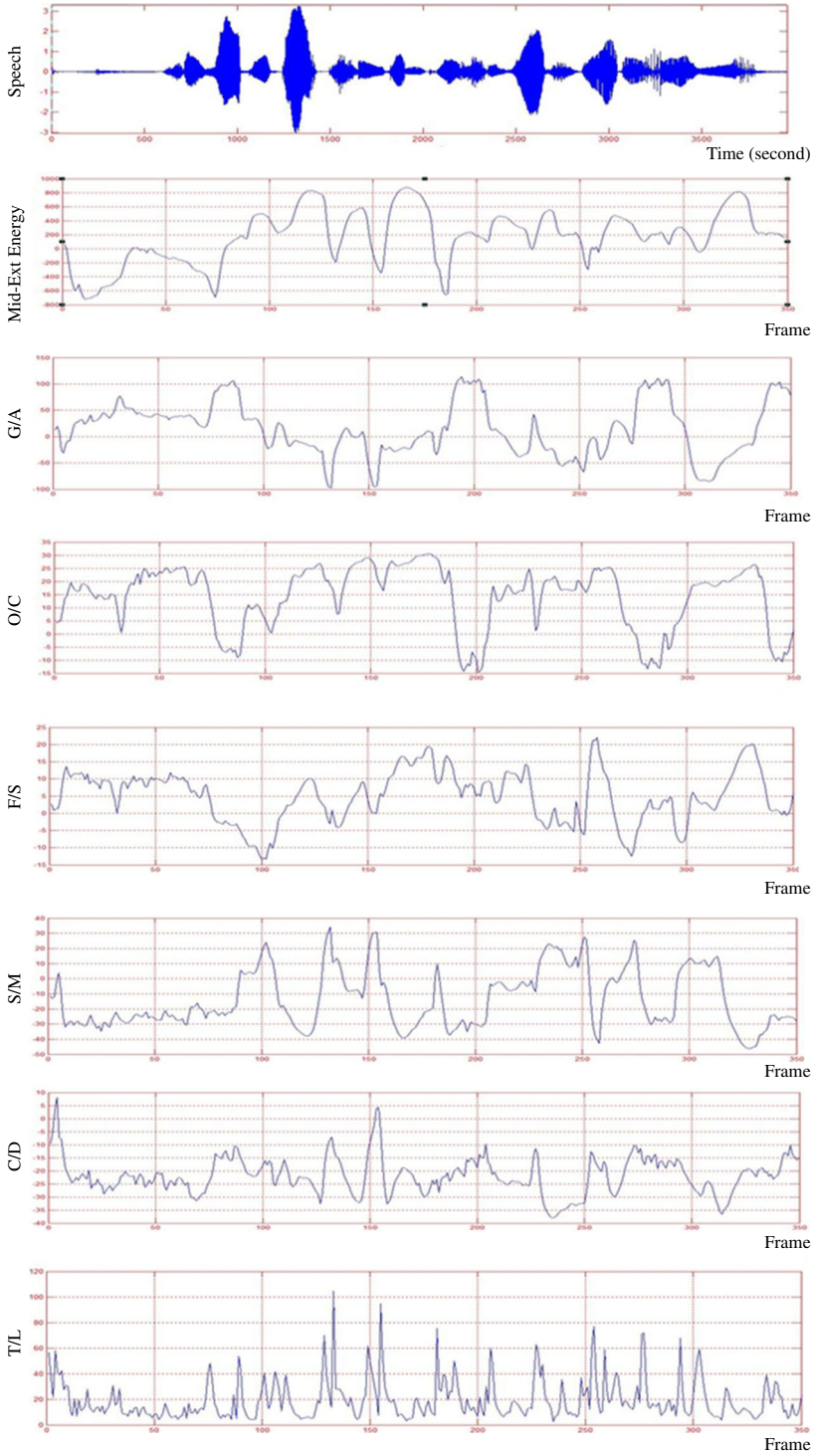

Fig. 2 Acoustic feature based on clean auditory model derived from the speech coding

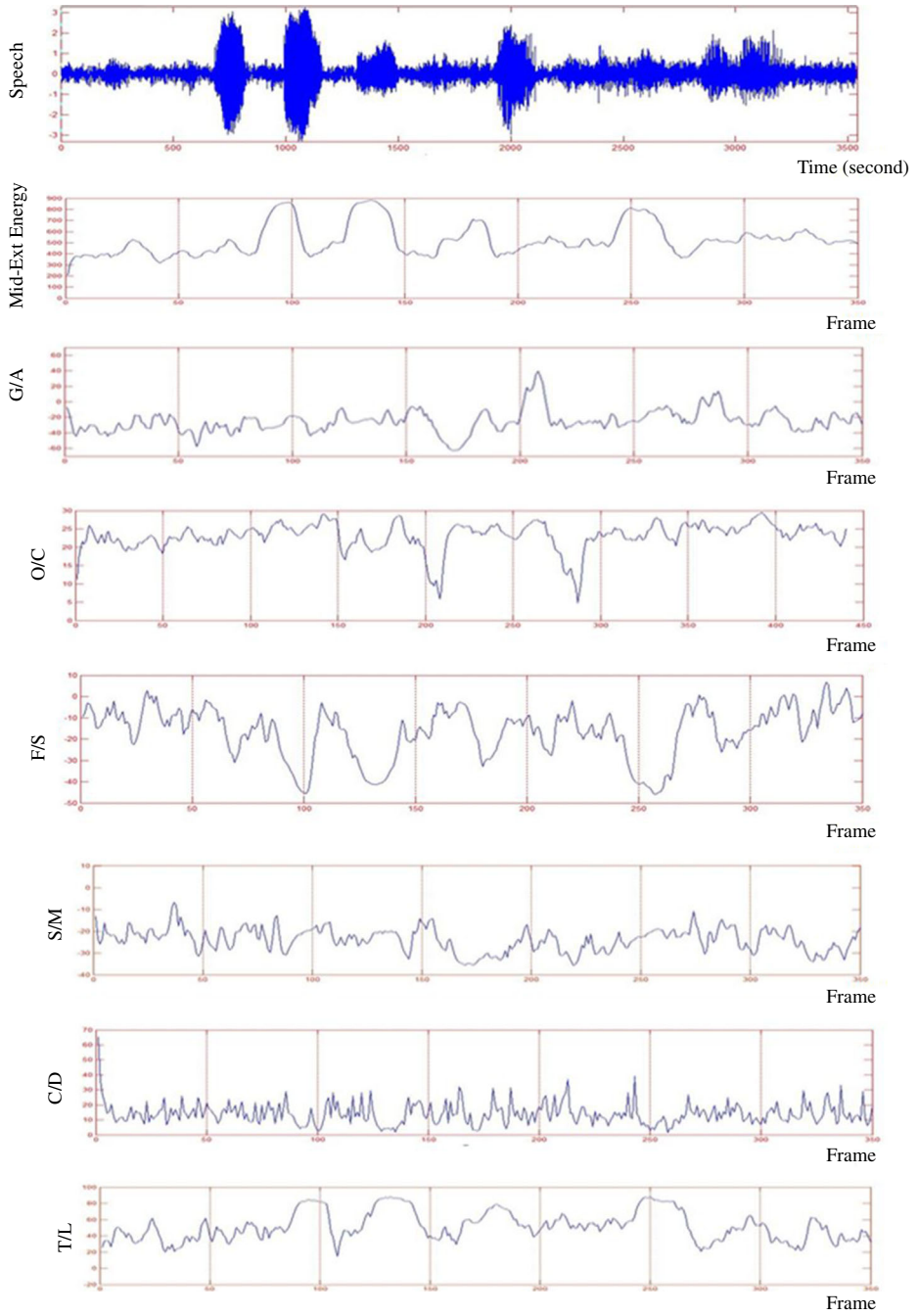

Fig. 3 Acoustic feature based on clean auditory model derived from the noisy speech

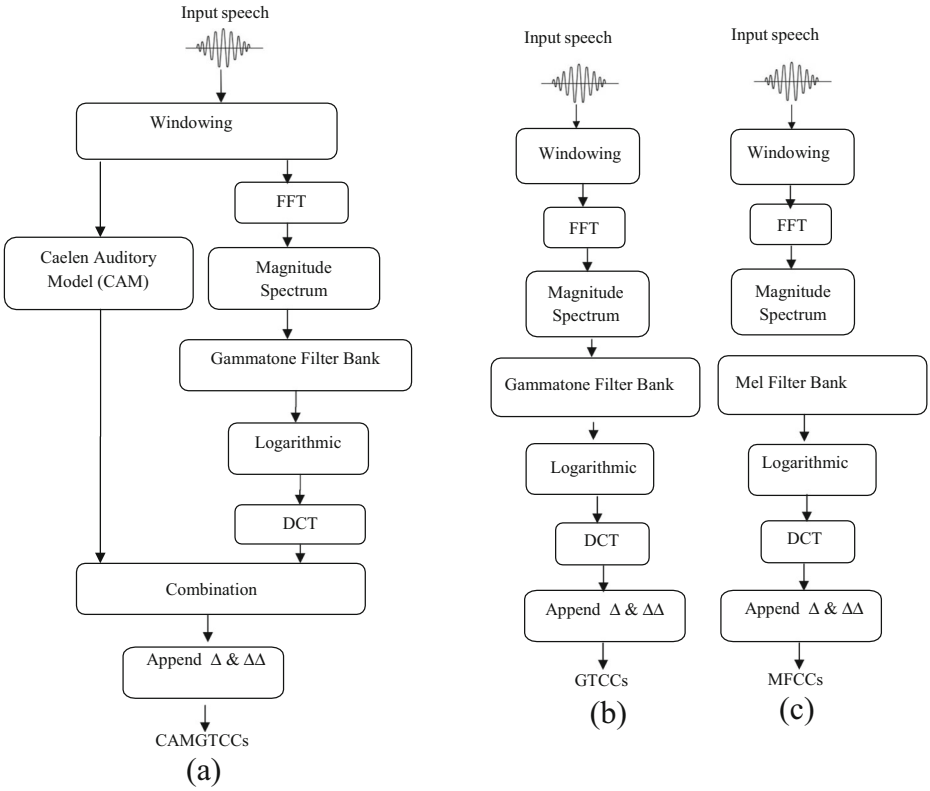

Fig. 4 Block diagram of, (a) CAMGTCCs, (b) GTCCs and (c) MFCCs features

generalized block diagrams of MFCC, GTCC and the CAMGTCC features extraction processes, respectively.

Thereafter, we presented an algorithm that summarized the different steps of the feature extraction proposed.

Algorithm, the steps involved in the CAMGTCCs feature extraction.

1. Read the speech signal s(t)
2. Calculate the Fourier Transform FFT
3. Applied the gammatone filter-bank
4. To compress the dynamic range of the estimated spectral parameters power-law nonlinearity is applied
5. Apply DCT on GTCC and retain first few coefficients to obtain GTCCs
6. Calculate the Caelen Auditory Model (CAM)
7. Combined Caelen Auditory Model (CAM) with GTCC
8. Append CAMGTCC with their first and second order derivative

4 Experimental setup

This section presents the results of the experiments to study the performance of the proposed features under noisy environment and speech coding distortion. We use the TIMIT corpus which contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers (438 male and 192 female) from 8 major dialect regions of the United States, both section from DR1 to DR8 and NIST-2008 Speaker Recognition Evaluation (SRE) corpora containing a single channel microphone recorded conversational segments of eight minutes or longer duration of the target speaker and an interviewer [24] [29]. The clean waveforms are transcoded by passing them through a coding and decoding AMR-WB codec [33]. For evaluating the robustness of these features in mobile noisy conditions, the speech test data are corrupted using (Babble, Car, Factory) at various noise levels (0, 5, 10,15 and 15 dB) which are taken from NOISEX-92 database [47]. In all experiments, we utilized three different acoustic features: (a) Mel frequency Cepstral coefficients (MFCC), (b) Gammatone Cepstral Coefficients (GTCC) and (c) Caelen Auditory Model Gammatone Cepstral Coefficients (CAMGTCC). The feature vectors contain 20 cepstral coefficients appended with the first and second order time derivatives, thus providing 60 dimensional feature vectors. After feature extraction, each speaker model is adapted from a 512-component in which the Universal Background Model (UBM) are trained using the entire database. For the total variability matrix training, the UBM is training dataset is used. The Expectation Maximization (EM) Algorithm training is performed throughout five iterations. We use 400 total factors (i.e., the i-vector size is 400) then Linear Discriminate Analysis (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 [45]. In practice, the MSR Identity Toolbox [41] is used for implementing the GMM-UBM and i-vector-GPLDA system. Speaker recognition systems are evaluated by speaker ID accuracy and equal error rate (EER). The speaker ID accuracy can be defined

$$Speaker\ ID\ accuracy\ (\%) = \frac{\text{Correctly identified recordings}}{\text{testing recordings}} \times 100 \tag{8}$$

EER (%) determines the threshold values for its false acceptance rate (FAR: probability (or rate) for an impostor to be accepted by the system) and its false rejection rate (FRR: probability (or rate) for a client (target speaker) to be rejected by the system). Its value indicates that the rate of false acceptances (FAR) is equal to the rate of false rejections (FRR). The lower the equal error rate value, the higher the accuracy of the verification system.

4.1 GMM-UBM and i-vector GPLDA based speaker recognition

The Gaussian mixture model–universal background Model (GMM-UBM) approach used for the first time in this work (Reynolds et al., 2000), represents speaker-independent distribution of feature vectors. The standard GMMs model which gives the distribution of feature vectors for speaker-dependent can be defined as:

$$p(x/\theta) = \sum_{k=1}^{M} \omega_k p_k(x) \tag{9}$$

where M represents the mixture weights, $p_k(x)$is the prior probability. The UBM modelthat is a large GMM trained on a diverse dataset to be speaker- and text-independent. TheGMM –UBM super-vector is obtained by concatenating the mean vector of all components ofthis aapted GMM.

$$M = \{\mu_1, \mu_2, \ldots\ldots\ldots, \mu_k\} \tag{10}$$

where μ_k is the mean super-vector of the k Gaussian component. However, The GMM – UBM super-vectors are very high dimensional vectors. In the total variability space, a speech utterance is represented by a vector in a low dimensional subspace in the GMM-UBM super-vector domain called *i-vector* where speaker and channel information is assumed dense. It is expressed as

$$M = m + Tw \tag{11}$$

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). In the GLPDA modeling approach, a speaker and channel dependent i-vector, can be defined

$$w_r = w + U_1 x_1 + U_2 x_2 + \varepsilon_r \tag{12}$$

where U_1 is the eigen-voice matrix and U_2 is the eigen-channel matrix, x_1 is the speaker factor and x_2 the channel factor; ε_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 [32]. Given two i-vectors, w_1 and w_2 for the target and test utterance, the batch likelihood ratio is as follows

$$Score(w_1, w_2) = \log \frac{P(w_1, w_2/\phi_s)}{P(w_1, w_2/\phi_d)} \tag{13}$$

where ϕ_s denotes the hypothesis that the i-vectors represent the same speakers and ϕ_d denotes the hypothesis that they do not. The block diagram of the proposed speaker recognition system is shown in Fig. 5.

5 Results and discussion

5.1 Speaker ID performance under codec degradation

In this section, we analyzed the effect of speech coding over speaker identification performance by using different extraction features such as CAMGTCC, GTCC and MFCC. In speech coding, there are two reasons why a speech codec can degrade the recognition performance. The first and more important is the degradation by the compression itself. The second reason is using different speech codecs in the speaker identification system, then a mismatch between training and testing conditions. The original feature set and codec feature set are represented by a couple of points(C^{Clean}, C^{Codec}), where C^{Clean} is a particular feature computed from clean data, while C^{Codec} is the feature computed from coded-decoded speech using G722.2 at different bit-rates. Figure 6 shows an alignment of these GTCC, CAMGTCC feature from original speech and GTCC-codec, CAMGTCC-codec feature from coded-

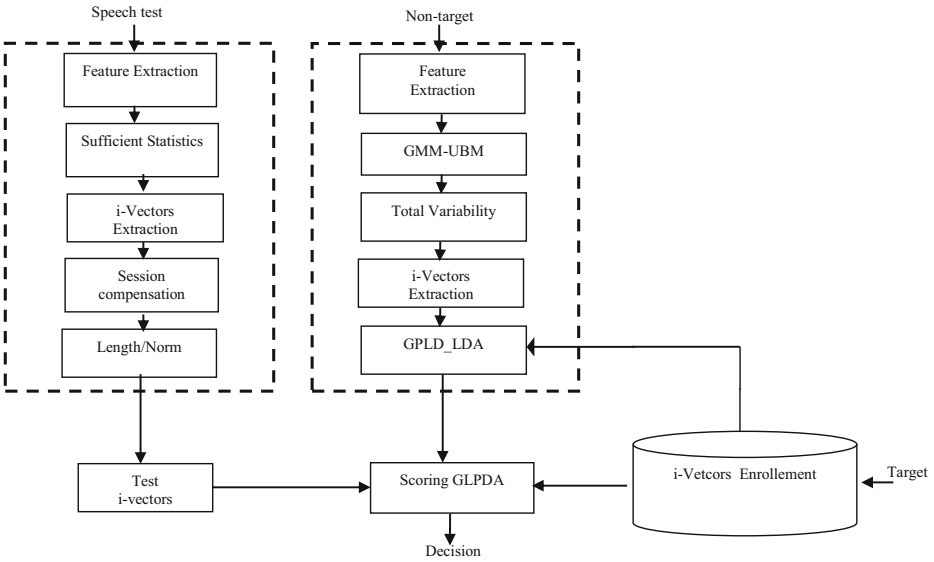

Fig. 5 A block diagram of GMM-UBM and G-PLDA based i-vectors for speaker recognition system

decoded speech. The coded–decoded distortion, defined as $D = C^{Clean} - C^{Codec}$ can be seen in Fig. 6, the coefficients of clean and speech codec with CAMGTCC feature give good linear distribution compared with GTCC feature.

Table 2 and Fig. 7 show the performance of the proposed CAMGTCC feature with the baseline system MFCC and GTCC for speaker identification in multi-condition training. It can be seen that the proposed CAMGTCC feature has a better performance compared with the GTCC and MFCC under the different bit-rate.

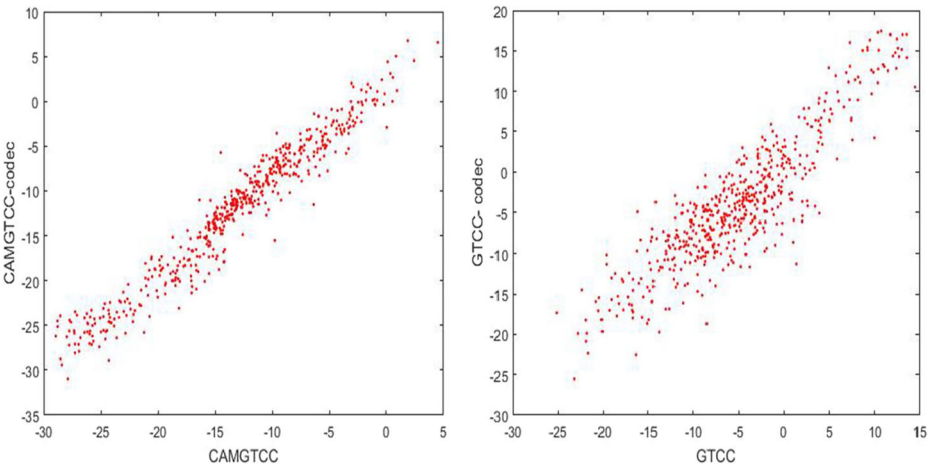

Fig. 6 Cluster plot of the two first coefficients of different features (CAMGTCC and GTCC) from original speech and coded-decoded speech

Table 2 Comparison of speaker identification performance under different bit-rates of AMR-WB speech codec

Speaker ID Accuracy (%)									
Bit-rate	23.5	19.85	18.25	15.85	14.25	12.85	8.85	6.60	Clean
MFCC	55.15	54.55	53.94	53.94	53.48	52.58	48.71	44.09	93.33
GTCC	73.40	72.80	72.64	72.95	72.64	71.43	69.04	62.77	77.20
CAMGTCC	**76.29**	**76.44**	**77.05**	**76.44**	**75.53**	**75.38**	**72.53**	**67.02**	**83,89**

5.2 Speaker identification SID and speaker verification ASV performance under noise speech

In the next section, we evaluate the performance of speaker ID and speaker verification ASV system using MFCC, GTCC and CAMGTCC features, where the acoustic conditions of training and testing are mismatched; the training data and the testing data were set with different types of background noise (Factory, car and babble) at various SNR levels (0, 5, 10 and 15 dB). The speaker ID accuracy using the feature extraction techniques for different noise speech is shown in Table 3. It is evident to show that the proposed features perform better than GTCC and MFCC at almost all SNR levels and clean condition.

EER (%) performance using the feature extraction techniques for different noise speech is shown in Figs. 8, 9 and 10. Those figures presents the experimental results for this part of the study, ASV performance in terms of equal error rate (EER), where the acoustic conditions of training and testing are mismatched; the training data set was recorded under a clean condition and the testing data sets with different types of background noise (Factory, car and babble). The results were obtained by using a development set i-vector-GPLDA with different feature extraction methods (MFCC, GTCC and CAMGTCC) .

From the results of Figs. 8, 9 and 10, we can see that the CAMGTCC feature outperform the baseline MFCC and GTCC feature, more specifically, at low level noise (0db and 5db). Compared to the MFCC and GTCC baseline feature, the CAMGTCC feature achieves a reduction in average Equal Error Rate (EER) ranging from 1.05% to 0.25%, 1.05% to 0.25% and 10.88% to 6.8% at Babble, Car and Factory noises, respectively.

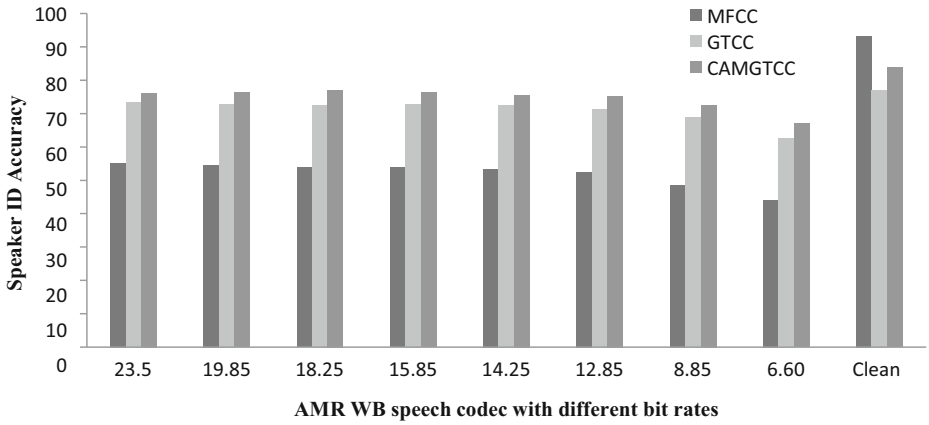

Fig. 7 Speaker ID Accuracy of clean and AMR WB speech codec with different bit rates using a proposed feature

Table 3 Speaker ID Accuracy (%) with different types of background noise (Babble, Car and Fctory)

Noise	Features	0	5	10	15
Babble	MFCC	2.88	7.58	20.45	44.12
	GTCC	3.65	10.10	33.89	60.46
	CAMGTCC	8.58	13.26	37.66	60.06
Car	MFCC	19.70	28.48	49.47	66.72
	GTCC	34.19	57.71	66.41	70.23
	CAMGTCC	35.71	67.00	67.41	70.05
Factory	MFCC	4.89	17.10	35.73	56.64
	GTCC	8.24	25.50	46.72	59.08
	CAMGTCC	10.68	30.69	56.08	75.78

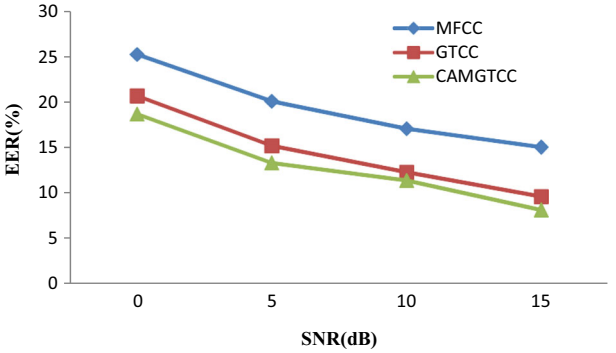

Fig. 8 ASV performance comparison (in terms of %EER) of the proposed feature with MFCC and GTCC baseline under Babbel noise at different SNR

6 Conclusion and future work

In this paper, we investigate the performance of speaker recognition system under noisy environment and speech coding distortion. We developed the new paradigm of feature based on Caelen auditory model and gammatone filterbank. This study of new feature was conducted by using the AMR-WB (Adaptive Multi-rate-Wideband) codec under mismatched testing

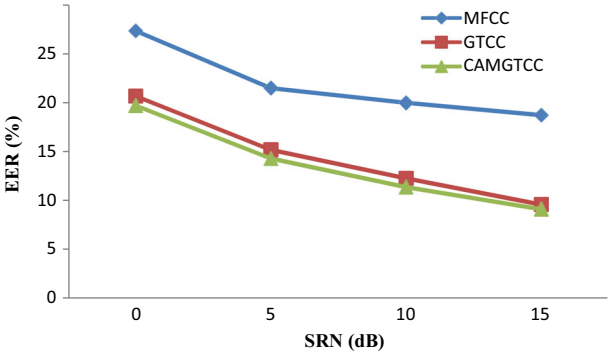

Fig. 9 ASV performance comparison (in terms of %EER) of the proposed feature with MFCC and GTCC baseline under Car noise at different SNR

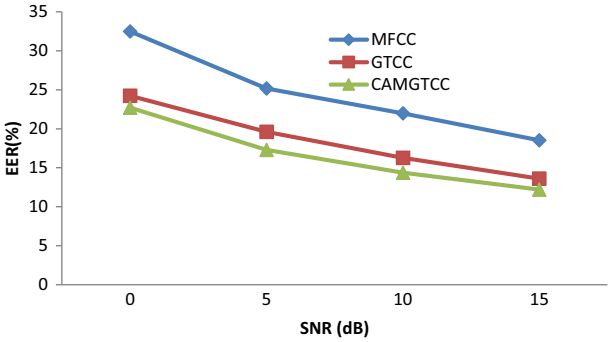

Fig. 10 **ASV** performance comparison (in terms of %EER) of the proposed feature with MFCC and GTCC baseline under Factory noise at different SNR

conditions. Furthermore, when we use the new feature is used with different acoustic noise which included Factory noise, car noise, and babble noise. Experimental results show that the CAMCTCC feature outperforms both the CTCC and MFCCs features in mismatched condition under speech coding distortion with different bit rates. In the noise environment, speaker recognition performance can be improved significantly even when the testing data are mismatched from the training data. Our system can achieve verification accuracy from 1.05% to 0.25%, 1.05% to 0.25% and 10.88% to 6.8% at Babble, Car and Factory noises, respectively.

We suggest two lines of research for further work. First of all, we would like to extend the experiments using the prosodic feature with acoustic distinctive to improve the performance of speaker recognition. Second, we plan to extend our study of auditory-based on acoustic distinctive features and gammatone filter to other speech application tasks, including speech synthesis and emotion recognition.

Declarations

Conflict of interest The Authors declare no conflicts of interest.

References

1. Al-Ali AKH, Dean D, Senadji B, Chandran V, Naik GR (2017) Enhanced forensic speaker verification using a combination of DWT and MFCC feature warping in the presence of noise and reverberation conditions. IEEE Access 5(15400–15):413–15413
2. Bellot O, Matrouf D, Merlin T, Bonastre JF (2000) Additive and convolutional noises compensation for speaker recognition In Sixth International Conference on Spoken Language Processing
3. Caelen J (1985) Space/Time data-information in the ARIAL project ear model. Speech Commun 4(1)
4. Chandra, M., Nandi, P., & Mishra, S. (2015). Spectral-subtraction based features for speaker identification. In proceedings of the 3rd international conference on Frontiers of intelligent computing: theory and applications (FICTA) 2014 (pp. 529–536). Springer, Cham
5. Chaouch H, Merazka F, Marthon P (2019) Multiple description coding technique to improve the robustness of ACELP based coders AMR-WB. Speech Commun 108:33–40
6. Chu W (2003) Speech coding algorithms: Foundation and Evolution of Standardized Coders A. John Wiley & Sons
7. Dunn RB, Quatieri TF, Reynolds DA, Campbell JP (2001). Speaker recognition from coded speech and the efects of score normalization. In proceedings of conference record of ThirtyFifth Asilomar conference on signals, systems and computers (Vol. 2, pp. 1562–1567)

8. 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.
9. 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.
10. Gallardo LF (2016) Human and automatic speaker recognition over telecommunication channels. Springer Science + Business Media, Singapore
11. Ghitza O (1994) Auditory models and human performance in tasks related to speech coding and speech recognition. IEEE Trans. Speech AudioProcess. 2:115–132
12. Glasberg M (1990) Derivation of auditory filter shapes from notched-noise data. Journal of Hering Elsevier 47(1–2):103–138
13. Grassi S, Besacier L, Dufaux A, Ansorge M, and Pellandini F (2000) Influence of GSM speech coding on the performance of textindependent speaker recognition," in Proceedings of EUSIPCO pp. 437–440
14. 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
15. 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
16. Kadi KL, Selouani SA, Boudraa B, Boudraa M (2016) Fully automated speaker identification and intelligibility assessment in dysarthria disease using auditory knowledge. Biocybernetics Biomed Eng *36*(1):233–247
17. Kinnunen T, Alam MJ, Matejka P, Kenny P, Cernocky J, OShaughnessy D (2013) Frequency warping and robust speaker verification: a comparison of alternative mel-scale representations. In: Proc. INTERSPEECH. Lyon, France, pp. 3122–3126
18. Kinnunen T, Li H (2010) An overview of text-independent speaker recognition: from features to supervectors. Speech Commun 52:12–40
19. Krobba A, Debyeche M, Amrouche A (2010) Evaluation of speaker identifcation system using GSM-EFR speech data. In proceedings of international conference on design and Technology of Integrated Systems (nanoscale era Hammamet) (pp. 1–5).
20. Krobba A, Debyeche M, Selouani SA (2019) Maximum entropy PLDA for robust speaker recognition under speech coding distortion. Int J Speech Technol 22(4):1115–1122
21. Krobba A, Debyeche M, Selouani SA (2019) Multitaper chirp group delay Hilbert envelope coefficients for robust speaker verification. Multimedia Tools Appl:1–18.
22. Li Z, Gao Y (2016) Acoustic feature extraction method for robust speaker identification. Multimed Tools Appl 75(12):7391–7406
23. Li Q, Huang Y (2011) An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Trans Audio Speech Lang Process 19(6):1791–1801
24. Linguistic Data Consortium (1990) The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc:CD1–1.1
25. Lyon RF, 1982. A computational model of filtering, detection, and compression in the cochlea. In: IEEE International Conference on Acoustics,Speech, and Signal Processing, pp. 1282–1285
26. McCree A (2006). Reducing speech coding distortion for speaker identifcation. In Annual Conference (Interspeech) (pp. 941–944)
27. Mclaren M, Abrash V, Graciarena M, Lei Y, Pesan J (2013) Improving robustness to compressed speech in speaker recognition. Interspeech, In, pp 3698–3702
28. Ming J, Hazen TJ, Glass JR, Reynolds DA (2007) Robust speaker recognition in noisy conditions. IEEE Trans Audio Speech Lang Process 15(5):1711–1723
29. NIST Year (2008) Speaker recognition evaluation plan, Technical report, NIST. http:www.itl.nist.gov/iad/mig/yest/ser/2008
30. Peinado A, Segura J (2006) Speech recognition over digital channels: robustness and standards. isbn:978-0-470-02400-3
31. Pohjalainen J, Hanilçi C, Kinnunen T, Alku P (2014) Mixture linear prediction in speaker verification under vocal effort mismatch. IEEE Signal Processing Letters 21(12):1516–1520
32. Rahman MH, Kanagasundaram A, Himawan I, Dean D, Sridharan S (2018) Improving PLDA speaker verification performance using domain mismatch compensation techniques. Comp Speech Lang 47:240–258
33. Recommendation G (2003) 722.2:Wideband Coding of Speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB).
34. Reynolds DA (2002) An overview of automatic speaker recognition technology. In IEEE international conference on acoustics, speech, and signal processing (Vol. 4, pp. IV-4072)

35. Saeidi R, Pohjalainen J, Kinnunen T, Alku P (2010) Temporally weighted linear prediction features for tackling additive noise in speaker verification. IEEE Signal Processing Letters 17(6):599–602
36. Samia AE, Nassar MA, Dessouky MI, Nabil A, Adel I, El-Fishawy S, El-Samie FEA. Text-independent speaker recognition using LSTM-RNN and speech enhancement. Multimedia Tools and Applications, 1–16
37. Selouani SA (2011) Speech Processing and Soft Computing. Springer, New York
38. Selouani SA, O'Shaughnessy D, Caelen J (2007) Incorporating phonetic knowledge into an evolutionary subspace approach for robust speechrecognition. Int. J. Comput. Appl. 29:143–154
39. Selouani SA, Alotaibi Y, Cichocki W, Gharsellaoui S, Kadi K (2015) Native and non-native class discrimination using speech rhythm-and auditory-based cues. Comp Speech Lang 31(1):28–48
40. Seneff S (1988) A joint synchrony/mean-rate model of auditory speech processing. J. Phon. 16:55–76
41. 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
42. Shabtai NR, Rafaely B, Zigel Y (2011) The effect of reverberation on the performance of cepstral mean subtraction in speaker verification. Appl Acoust 72(2–3):124–126
43. Sreenivasa RK, Vuppala AK (2014) Speech processing in mobile environments. Springer, ISBN: 978–319–03116-3.
44. Tan ZH, Lindberg B (2008) Automatic speech recognition on mobile devices and over communication networks. Springer Science & Business Media
45. Tan Z, Mak MW, Mak BKW, Zhu Y (2018) Denoised senone i-vectors for robust speaker verification. IEEE/ACM Trans Audio, Speech, Language Process 26(4):820–830
46. Valero X, Alias F (2012) Gammatone cepstral coefficients: Biologically inspired features for non- speech audio classification. IEEE Transactions on Multimedia *14*(6):1684–1689
47. Varga A, Steeneken HJ (1993) Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 12(3):247–251 http://spib.rice.edu/spib/selectnoise.html
48. Vuppala AK, Rao KS, Chakrabarti S (2013) Improved speaker identifcation in wireless environment. Int J Signal Imaging Syst Eng 6(3):130–137
49. Yu D, Deng L, Droppo J, Wu J, Gong Y, Acero A (2008) A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition. In 2008 IEEE international conference on acoustics, speech and signal processing (pp. 4041-4044). IEEE
50. Zhao X, Shao Y, Wang DL (2012) CASA-based robust speaker identification. IEEE Trans Audio, Speech, Lang Process 20(5):1608–1161