Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2020 SCPE. Volume 21, Issues 3, pp. 359–368, DOI 10.12694:/scpe.v21i3.1704

COMPARATIVE STUDY OF SPEAKER RECOGNITION TECHNIQUES IN IOT DEVICES FOR TEXT INDEPENDENT NEGATIVE RECOGNITION

NEHA R. KASTURE [*], POOJA JAIN [†], AND TAPAN KUMAR [‡]

Abstract. Speaker recognition (SR) or identification is the subset of broad area of Pattern recognition. Given the features of the voice print, the recognition system identifies the speaker from the knowledge of the speaker models stored in the database. In today's world when many of our works are done through voice, recognition of the speaker is necessary.Recently, SR has also gained importance in Internet of Things (IoT) like setting up of smart environments for home, industries or educational and commercial applications. The race for high accuracy needs making the devices used in these smart environments as close to human hearing capacity as possible. Speaker identification is mostly used to establish negative recognition [1].Negative recognition is when the system decides whether a person is who he disagrees to be thus preventing a person from exploiting multiple identities. Only biometrics will be suitable to establish such identification. The feature extraction of voice sample along with comparative analysis of its methods is of fundamental interest in this paper. We try to compare the performance of features which are used in state of art speaker recognition models and analyse variants of Mel frequency cepstrum coefficients(MFCC) predominantly used in feature extraction which can be further incorporated and used in various smart devices.

Key words: Speaker identification, Pattern recognition, MFCC, Feature extraction, Inverted MFCC, GMM

AMS subject classifications. 68M11

1. Introduction.

Speech or voice is a dominant mode of communication in everyday life in many of the Internet of Things (IoT) devices and comprises of unique features relevant to the user. So, although the primary function of a speech is to convey the message, the same speech is also used as bio metric feature to recognize the identity of the person.The last few decades have witnessed speaker recognition technology emerging in various commercial domains like bio-metric, banking applications, indexing or structuring of audio information and Diarization[2]. The emergence of smart devices and home assistants like Google Home or Alexa have brought in ample opportunities for authentication and use of speech samples[3] This biometric characteristic in the person's voice can be used to control the IoT devices. Automatic speaker recognition systems enables to recognize a speaker and hence authenticate it for making any transaction [4]. Speaker recognition systems can be broadly categorized as: speaker identification and speaker verification [5]. Speaker identification attempts to find "who is speaking" from a set of known speakers. This method is also called as Closed set identification because the unknown speaker belongs to the group of speakers in the database whose models are present in advance for matching. In Open set identification problem, the speaker can be an outsider not present in the finite pool of speakers known to the system. Speaker verification differs from recognition in the sense that it confirms if he/she is the authenticated person behind the speech sample.

Furthermore, speaker recognition can also be distinguished as text dependent or text independent. Text dependent speaker identification requires speaker to utter predefined word or phrase from a limited vocabulary, while text independent speaker identification is more flexible and does not restrict user to utter the predefined keyword. The research paper [6] focuses on the scope of text-independent speaker verification using short utterances.

The speaker recognition system operates in two phases. The first is the training phase or enrollment phase where model is created from the speech samples of the different speakers who will need identification. This step is usually completed before the system goes live for voice identification. The second phase is testing phase or

[*] Department of CSE, IIIT Nagpur, India (nehakasture86@gmail.com)
[†] Department of CSE, IIIT, Nagpur, India (pooja.jain@cse.iiitn.ac.in)
[‡] Department of ECE, IIIT Nagpur, India (tapan.jain@ece.iiitn.ac.in)

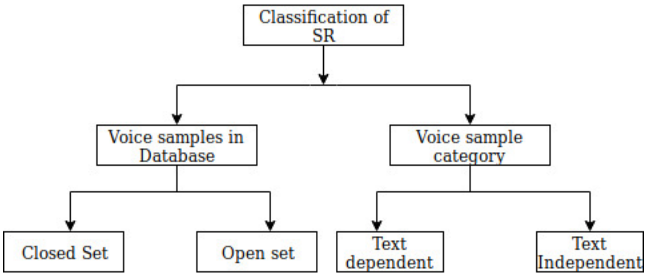

FIG. 1.1. *Broad Classification of Speaker Recognition Systems*

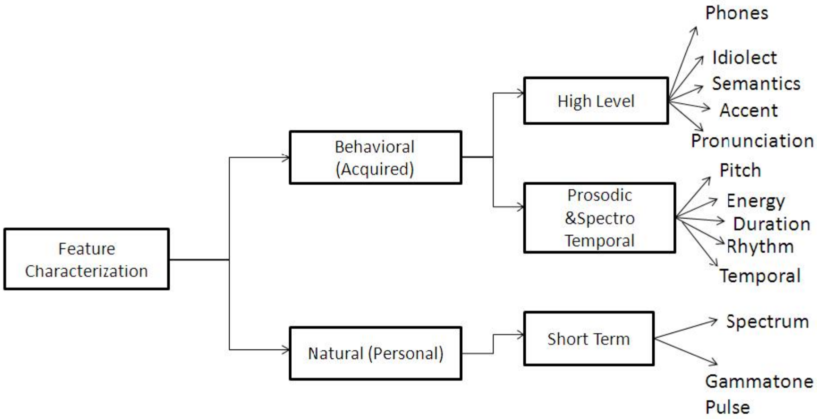

FIG. 1.2. *Broad Classification of Features*

verification phase where signal is matched with the models of speaker available in the database. The implicit assumption here is that each sample belongs to only one speaker. Signals are some quantifiable outputs. Once, the signal is received the fundamental interest lies in extracting a set of feature vectors from speech signals [7]. Feature extraction is one of the preliminary steps of acoustic modelling which quantifies the properties of input speech signal.

In this work we focus on studying and and analysing feature extraction techniques and its effectiveness when working with different Speaker recognition models like Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM). This study aims to choose better combination of feature and model to improve the accuracy of speaker recognition which can be then employed in various scenarios of Human Computer Interaction (HCI).

2. Acoustic Feature Extraction. Feature extraction is crucial to extract characteristics from spoken utterances received from the front end of the model. Features can be broadly classified as (1) short-term spectral (2) voice source (3) spectro-temporal (4) prosodic (5) high-level features based on their physical interpretations [8]. The authors [9] detail the language models for recognition of tamil language. The speech signals were segmented at phonetic levels on the basis of their acoustic characteristics. Spectral analysis determines the frequency content of an arbitrary signal. Spectral features can be obtained by by converting the time based signal into the frequency domain using the Fourier Transform, like: fundamental frequency, frequency components, spectral density, etc. These spectral features can be used to identify characteristics like notes, pitch, rhythm, and melody.

The most popular feature extraction technique used recently in tasks of speaker recognition in different applications is MFCC [10]. But pure MFCC approach modeled on human auditory system is observed to be efficient in non-noisy environments. With the increase in vocabulary or ambient noise the performance of MFCC features seems to decline. Here we analyse the variations of MFCC features which are more robust in nature for the task of SR. See Section 2.

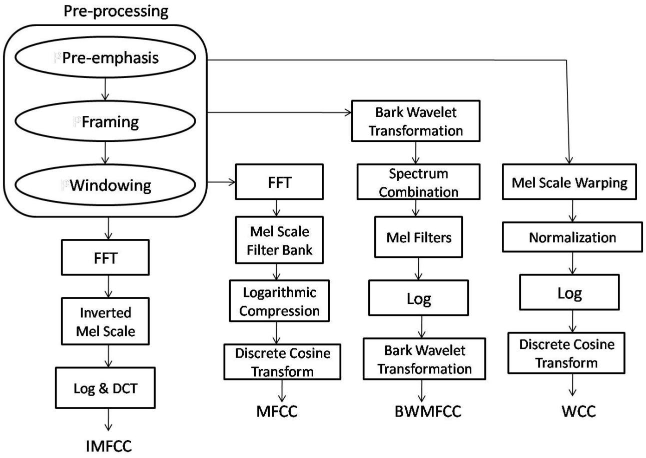

Fig. 2.1. *Pipeline of Feature Extraction*

2.1. Pre-processing. The speech waveform sampled at 8 Khz is used as input for feature extraction. Following steps are common for extracting MFCC, IMFCC and Fused MFCC [11].

$$\xi_{ij}(t) = P(x_t = i, x_{t+1} = j | y, v, w; \theta) = \frac{\alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i(t) a_{ij}^{w_t} \beta_j(t+1) b_j^{v_{t+1}}(y_{t+1})} \tag{2.1}$$

2.1.1. Pre-emphasis. This step is applied to improve Signal to noise ratio. Higher modulating frequencies are more susceptible to noise than lower ones. Hence, higher frequencies need to be boosted artificially. First order finite impulse response (FIR) filter is applied for spectral flattening as shown in equation 2.2 [12].

$$H(z) = 1 - \alpha z^{-1}, \quad 0.9 \leq \alpha \leq 1 \tag{2.2}$$

The value of αused for experimentation is usually 0.97.

2.1.2. Framing. Frames should not be too long or too short. Longer frame result in rapid changes in signal properties across the window, thus negatively affecting the time resolution, while too short a frame comes at a cost of affecting the frequency resolution of the signal. So, there always exists a trade-off between time and frequency resolution [13]. So for such non-stationary signals as speech, usually 256 samples in each frame can be chosen with 128 overlapping samples in the adjacent frames with the intention of extracting any vital information occurring at the edges of the frames. In terms of time duration 25ms frame generated every 10 ms with a overlap of 15ms is a popular approach.

2.1.3. Windowing. Distortions in the frame boundaries can give rise to unwanted effects in the frequency response. A window function works with signal in such a way that it smooths the frame at the beginning and end at nearly zero to maintain the continuity [12]. Many window functions like rectangular window, flat top window and hamming window and hanning window are available to implement this step. Out of these possible options Hamming window is the most popular technique used in majority of feature extraction methods as it introduces minimum distortion. The equation 2.3 shows the hamming window function:

$$h[n] = \begin{cases} 0.54 - 0.46 \cos \dfrac{2\pi n}{N}, & 0 \leq n \leq N \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

2.2. MFCC Features. Most of the Speaker Recognition tasks today employ MFCC method for extraction of features [10]. Introduced in early 1980's, these features based on human auditory system [14]are still relevant. MFCC's are representative of vocal tract information. The significant feature of MFCC is its use of perceptually inspired Mel-spaced filter bank processing of the Fourier Transform. Another advantage is flexibility of use achieved through cepstral analysis. Following are the steps for the extraction of MFCC features.

2.2.1. Fourier Transform. Fast Fourier Transformation (FFT) is generally applied on each frame to calculate the components of frequency from the signal in time domain called as spectral values. FFT output is a set of complex numbers containing both real and imaginary part where, real values are dealt with and imaginary part is ignored. In a way, output of FFT and DFT transformation is same the only difference is in terms of computational complexity [15], FFT increases the processing rate of the signal. The following equation shows the DFT for input frame $x(n)$ of 256 samples. 256-point FFT can be used to convert frame of 256 samples into its equivalent DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\prod kn}{N}}, \quad 0 \leq k \leq N-1 \tag{2.4}$$

2.2.2. Mel Scaled Filterbank. The spectrum obtained from the above step contains lot of fluctuations and not the whole spectrum details are useful. Only the envelope of the spectrum is of use here. Hence the spectral envelope is obtained by multiplying the spectrum with Mel scaled filterbank. Each filter in the filterbank is a triangular filter which is uniformly spaced on the Mel frequency axis, having more filters in the low frequency region and less number of filters in the higher frequency region. Mel Frequency analysis is very close to how humans perceive sounds. Also, it is proved by experimentation that sensitivity of human ears is more towards low frequency than high frequency. The voice utterance does not follow linear scale frequency which is used in FFT hence, Mel scale is used which is linear upto 1kHz and logarithmic at higher frequencies. The equation sated below shows the relation between Linear scale and mel scale frequency.

$$Mel(f) = \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5}$$

2.2.3. Logarithmic Compression. This step aims to take log of spectral envelope obtained from the step above, since human ears cannot hear sounds in linear scale. Each co-efficient of envelope is multiplied by 20 to get the spectral envelope in dB.

2.2.4. Discrete Cosine Transform. This final step ensures conversion of log Mel spectrum into its spatial domain. This is achieved by taking Discrete Cosine Transform (DCT) which divides a finite sequence into discrete vector.Thus, DCT yields cepstral coefficients [16]as follows:

$$c_n = \sum_{k=1}^{K} S_k \cos\left[n(k - \frac{1}{2})\frac{\prod}{k}\right], n = 1, 2....L \tag{2.6}$$

where K is the number of log-spectral coefficients calculated in previous step, S_k are the log-spectral coefficients, and L is required number of cepstral coefficients that we want. The MFCC feature is finally achieved from lowest 12-15 DCT coefficients.

2.3. Bark Wavelet MFCC Feature. The DCT and FFT algorithms used in MFCC feature extraction do not prove to be a good option if the signal to be processed is non-stationary. The bark wavelet feature introduced by [17] proves as an anti-noisy feature that can substitute MFCC and overcome the disadvantages of fixed time-frequency resolution of DCT. Humans perception of speech is non linear if actual frequency is used but linear if Bark frequency is used. The relationship between linear frequency and Bark frequency can be represented as shown below:

$$b = 13 \cdot \arctan(0.76f) + 3.5 \cdot \arctan\left(\frac{f}{7.5}\right)^2 \tag{2.7}$$

where b represents bark frequency and f represents linear frequency. Following steps comprise of the general philosophy behind bark wavelet:

- Gaussian function is chosen as mother function of Bark wavelet to satisfy time and bandwidth product least.
- To maintain consistency with the frequency group,mother wavelet is chosen to have the equal bandwidth in the Bark domain.
- Unit bandwidth of 1 Bark keeps the consistency with the frequency group.

2.3.1. Pre-processing. The Pre-processing stage consisting of Pre-emphasis, Framing and Windowing is same as that of extracting MFCC features.

2.3.2. Bark Wavelet Transformation. Bark Wavelet Transformation can be performed by using the following equation on every frame:

$$s_k(n) = \sum_{l=0}^{N-1} S(l)W_k(l)e^{\frac{j2\Pi nl}{N}} \tag{2.8}$$

where N is the number of zeros in FFT, S(l) is the frequency spectrum of Speech signal, $s_k(n)$ is the speech spectrum of the k^{th} sub-band and $W_k(l)$ is a discrete form of $W_k(f)$ which is expressed as follows:

$$W_k(f) = c_2 2^{-4[13\arctan(0.76f+3.5\arctan(\frac{f}{7.5})^2-(b_1+k\Delta b)]^2} \tag{2.9}$$

where normalization factor c_2 can be calculated as

$$c_2 \sum_{k=0}^{K-1} W_k(b) = 1, \quad 0 < b_l \leq b \leq b_h \tag{2.10}$$

where $[b_l, b_h]$ is the Bark frequency bandwidth.

2.3.3. Spectrum Combination. Frequency Synthesis is obtained using the equation 2.11

$$s(n) = \sum_{k=0}^{K-1} s_k(n) \tag{2.11}$$

where $s(n)$ is the frequency synthesis spectrum.

2.3.4. Mel Filters. Signal $s(n)$ is passed through mel filters to reduce the effect of tone and pitch in the feature co-efficients and emphasize the original formant of speech.

2.3.5. Logarithm. Here, log of spectrum obtained through mel filters is taken as follows:

$$d(m) = \log(\sum_{n=0}^{N-1} |s(n)|^2 H_m(n), \quad 0 \leq m < L \tag{2.12}$$

where $H_m(n)$ is triangular mel frequency band pass filters, L is number of filters and N is sample number $s(n)$.

2.3.6. Bark Wavelet MFCC features. Finally the Bark Wavelet MFCC features (BWMFCC) is obtained by performing Bark Wavelet Transform on $d(m)$ as follows:

$$BWMFCC(n) = \sum_{m=0}^{L-1} |s(n)|^2 W_n(m) \cdot d(m), \quad 0 \leq m \leq L-1, \quad 0 \leq n < M \tag{2.13}$$

2.4. Wavelet Cepstral Coefficient. Another short term feature vector that is effective at keeping the effects of noise at bay is Wavelet Cepstral Coefficient (WCC) that uses Discrete wavelet transform (DWT). The detailed guidelines for DWT implementation is mentioned in [18]. A typical wavelet transform can be given as:

$$W_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \Psi \left(t - \frac{b}{a} \right) dt \tag{2.14}$$

where the function $\Psi(t)$ is a mother wavelet, a is scaling factor and b is translation parameter. DWT obtains the spectrum using multilabel resolution technique. In comparison with FFT used in MFCC, DWT distributes the signal into smaller frequency domains to obtain the local frequency spectrum. The advantage of such decomposition is, if the parts of signal is distorted by noise, the whole frequency spectrum won't be affected much. Thus making DWT more robust towards noise. Wavelet packet transform (WPT) offers a flexible multi-resolution approach which can vary the window length to suit better time or frequency resolution. This results in better time frequency characteristics of WPT but at the cost of computational overhead. The traditional WPT does not warp frequencies as per human auditory perception system. Therefore the work proposed by [19] has combined the advantages of multi-resolution WPT and Mel scale to give Wavelet Packet Based Mel Frequency Cepstral Features.

2.4.1. Pre-processing. The speech is initially sampled and subjected to common set of pre-processing steps as mentioned earlier consisting of Pre-emphasis, Framing and Windowing.

2.4.2. Mel scale warping. Mel scale warping consists of 3 sub-steps:

- Fast Fourier Transform (FFT) is used to transform the pe-processed signal from time domain to frequency domain.
- The frequency spectrum obtained through FFT is Mel-warped using triangular mel filter banks.
- The signal is again converted to time domain by Inverse FFT to carry out the further processing.

2.4.3. Wavelet packet decomposition. The speech signal is decomposed at depth 7 (level 7), with Daubechies type (db4) wavelet. The resultant wavelet consists of maximum frequency of 31.25 Hz producing 128 sub-bands.

2.4.4. Best basis formulation. 35 sub-bands out of 128 total frequency sub-bands are selected for further processing since higher frequency coefficient represents maximum amount of energy. The sub-band signal energies in each frame can be computed as

$$E_j = \frac{\sum_{j=1}^{N_j} [W_j^p f(i)]^2}{N_j}, \quad j = 1...35 \tag{2.15}$$

2.4.5. Log and DCT. Finally, The logarithmic compression is performed and DCT is taken to reduce the dimension of sub-band energies.

2.5. Inverted MFCC. MFCC effectively captures the low frequency region than high frequency region. Hence it is capable of extracting the formants[20] lying in the lower range of frequency. But this extraction of formants in lower frequency range neglects the formants if any lying in higher range of frequency. This essentially happens because of filter bank structure [21] where higher number of closely spaced overlapping triangular filters appear in lower frequency region of Mel filter bank and less number of overlapping triangular filters in higher frequency area. The approach of [22] is based on reversing the the normal MFCC filter bank structure to capture the characteristics in higher frequencies missed out by MFCC. This feature is called as Inverted MFCC (IMFCC). The initial steps of Pre-processing and FFT are common in this appraoch. The variation in implementation lies in the complementary way in which filter bank is used.

2.5.1. Inverted Mel Scale. The complementary filter bank structure is obtained by reversing the original from the general mid point of frequency range i. e., 0-4kHz in speaker recognition applications. Thus, reversing is done at 2kHz point of original filter bank. Mathematical expression for i^{th} filterbank of the same can be given as:

$$\widehat{\varPsi}_i(k) = \varPsi_{Q+1-i} \left(\frac{M_s}{2} + 1 - k \right) \tag{2.16}$$

where $\widehat{\Psi}(k)$ is the response of inverted Mel scale filter, $\Psi(k)$ is the original Mel scale filter response, Q is the number of filters, $(1 \leq i \leq Q)$ M_s is the number of points in DFT and $(1 \leq k \leq M_s)$ The relationship between inverted mel scale and original can be expessed as:

$$\widehat{f_{mel}}(f) = f_{mel}(f_{high}) + f_{mel}(f_{low}) - f_{mel}\left[\frac{F_s}{2} + \frac{F_s}{M_s} - f\right] \tag{2.17}$$

where $f_{mel}(f)$ is the relative pitch in the inverted scale corresponding to f, the actual frequency in Hz. Also to maintain uniformity in DFT calculation Inverted Mel Scale is made to have common boundary points as with the actual Mel Scale such that $\widehat{f_{mel}}(f_{low}) = f_{mel}(f_{low})$ and $\widehat{f_{mel}}(f_{high}) = f_{mel}(f_{high})$. This flipping in the Mel scale gives fine represenatation of high frequency regions not otherwise not justified by MFCC Mel Scale. Filter outputs $\{\hat{e}(i)\}_{i=1}^{Q}$ are computed from energy spectrum $|Y(k)|^2$ as

$$\hat{e}(i) = \sum_{k=1}^{M_s/2} |Y(k)|^2 \cdot \widehat{\Psi}_i(k) \tag{2.18}$$

2.5.2. Log and DCT. Logarithm of filter bank energies is taken as:

$$\{\log_{10}[\hat{e}(i)]\}_{i=1}^{Q} \tag{2.19}$$

The last step is to obtain the inverted MFCC coefficients by taking the DCT of log energies obtained in 2.19 as shown below:

$$\hat{C_m} = \sqrt{\frac{2}{Q}} \sum_{l=0}^{Q-1} \log[\hat{e}(i+1)] \cdot \cos\left[m \cdot \left(\frac{2l-1}{2}\right) \cdot \frac{\Pi}{Q}\right] \tag{2.20}$$

Usually with MFCC features we choose the first 19 coefficients as features to model the speaker but in case of IMFCC we choose the last 19 coefficients to model the speakers.

3. Speaker Modeling: Gaussian Mixture Model. The basic purpose of this step is building a model for any speaker 's' such that 'x' feature vector extracted from the utterance of speaker 's' can be represented by a unique model. Thus, matching an unknown voice sample with the speaker model can result in recognition of the correct speaker. One of the most universal modeling framework used for SR task is Gaussian Mixture Model (GMM)[23]. GMM's are very popular in text independent SR applications where the speaker is not restricted to use any pre-defined phrase as a voice sample. Gaussian distribution is particularly identified as a mean and a deviation about the mean. For a D-dimensional feature vector x, $\{\overrightarrow{x_t} \in \mathbb{R}^{\mathbb{D}} : 1 \leq t \leq T\}$, the mixture density used for the likelihood function is defined as [24]:

$$p(x \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x) \tag{3.1}$$

where GMM is denoted by λ, M is is the number of Gaussian components, w_i is the prior probability or mixing weight of the i^{th} Gaussian component constrained to $\sum_{i=1}^{M} w_i = 1$, and $p_i(x)$ is given by

$$p_i(x) = \frac{1}{(2\Pi)^{\frac{D}{2}} \mid \sum_i \mid^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu_i)'(\sum_i)^{-1}(x-\mu_i)\right\} \tag{3.2}$$

where $p_i(x)$ is is the D-variate Gaussian density function with mean vector μ_i and covariance matrix $\sum_i$. Collectively the GMM model is denoted as $\lambda = \{w_i, \mu_i, \sum_i\}$ where $i = 1 \ldots M$. The average log-likelihood of feature vector X with respect to model λ is defined as,

$$LL_{avg}(X \mid \lambda) = \frac{1}{T} \sum_{t=1}^{T} log \sum_{k=1}^{K} p(x_t \mid \lambda) \tag{3.3}$$

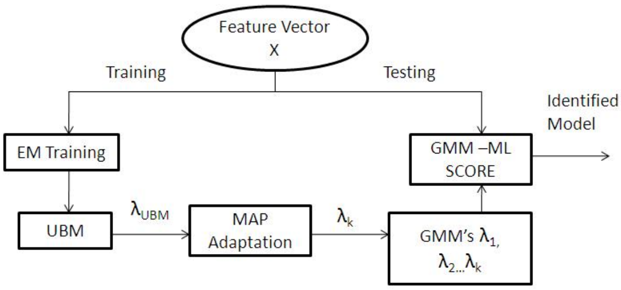

Fig. 3.1. *Training and Testing using GMM-UBM*

where $p(x_t \mid \lambda)$ is calculated as shown in equation 3.1. It has been empirically observed that diagonal matrix GMMs are better in performance and also computationally more efficient than full matrix GMMs. Estimating the parameters of a full-covariance GMM is very expensive [25]. Hence, diagonal covariance matrices are usually used. Maximum Likelihood (ML) estimation is used in training the GMM to estimate the parameter λ $= \{w_i, \mu_i, \sum_i\}$ where $i = 1 \dots M$ for a feature vector X.

Once the training vectors are ready, the iterative expectation–maximization (EM) algorithm [26] is used to maximize the likelihood with respect to the training data [27]. GMM parameters are refined with each iteration of EM algorithm to increase the likelihood of the estimated model for the observed feature data.K-Means can provide for the initialization of EM algorithm [28]. In general, five iterations are considered to be enough for parameter convergence. In applications pertaining to SR a model should adapt well with different types of speakers, their environments, speaking styles etc. Hence in GMM based SR a speaker-independent universal background model (UBM) is created. This UBM is trained with EM algorithm from hundreds of hours of speech data gathered from a large number of speakers. When a new speaker is enrolled into the system, the UBM adapts its parameters to the feature distribution of newly enrolled speaker. This adapted model is used as a model representing that speaker. Thus, prior knowledge is utilized for estimating model parameters. The maximum a posteriori (MAP) method [29] is used to extract speaker-specific GMM from the UBM. In the Testing phase, the MAP-adapted model and the UBM are combined, and the recognizer is called as Gaussian mixture model - universal background model, or "GMM-UBM". The test features received from the voice sample are compared with the speaker models available in the database and the model with highest log likelihood ratio (LLR) is chosen:

$$LLR_{avg}(X, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(x_t \mid \lambda_{target}) - \log p(x_t \mid \lambda_{UBM})\} \tag{3.4}$$

4. Discussion and conclusions. In this work we reviewed two architectures of SR namely GMM and HMM when working with MFCC features and its variants. Its comparative performance is listed in the above table. This study focuses on the performance analysis of MFCC and its variants discussed in the literature. Table 4.1 shows the recognition rate using the discussed feature extraction techniques like MFCC, IMFCC, BWMFCC and WCC as experimented by various researchers. The database used by the authors include TIMIT [31], YOHO [32], VoxForge [33] and also author created manual database. Comparison of MFCC and IMFCC features can be seen on YOHO database where MFCC outperforms IMFCC. IMFCC supports the extraction of information lying in the higher frequency range which is not considered by MFCC. The experimentation done by [34] shows that the results improve when fusion of MFCC and IMFCC is taken. WCC is tested on comparatively smaller corpus of 30 speakers but shows to provide better time and frequency resolution for limited data than MFCC. Bark wavelet based MFCC can be used as a good anti-noise substitute since it overcomes the disadvantage of fixed time-frequency resolution. The robustness of this feature is also demonstrated in detail by introducing the noise component as shown in the work of [30].The advantages and disadvantages of using any of the features mentioned above can be tabulated as shown in Table 4.2. Variations over pure MFCC can improve the system performance if used in applications where robustness is required where humans and smart assistants are involved.

Table 4.1
Comparison between MFCC feature variants

Referred System	Database used	Feature Used	Dimension	Modeling Technique	Accuracy achieved (%)
[22]	YOHO	IMFCC	138 speakers GMM Mixing Coefficient=32	GMM	95.23
[22]	YOHO	MFCC	138 speakers GMM Mixing Coefficient=32	GMM	96.82
[19]	VoxForge	WCC	30 speakers GMM Mixing Coefficient=15	GMM	100
[19]	VoxForge	MFCC	30 speakers GMM Mixing Coefficient=15	GMM	93.33
[30]	Author created Word Pronunciation	BWMFCC	16 speakers SNR=Clean Words=30	HMM	95.69
[30]	Author created Word Pronunciation	MFCC	16 speakers SNR=Clean Words=30	HMM	93.74

Table 4.2
Advantages and Disadvantages of Features

Feature	Advantages	Disadvantages
MFCC	Good choice for clean speech, Represents human auditory system, Easy and relatively fast to compute	Unsuitable in noisy conditions, performance degrades with larger vocabulary, Only low frequencies are considered and high frequencies are ignored
BWMFCC	Robust to noise, Suitable for larger vocabulary	Works for low signal to noise ratios, Complex due to additional Bark wavelet transformation
IMFCC	Capable of representing information in high frequency region, less computation burden as compared to other variants	Gives better results when fused with MFCC than individual IMFCC
WCC	Frequency spectrum obtained through wavelet is noise-resistant, good time and frequency resolution	To find the optimum mother wavelet, time consuming

REFERENCES

[1] James Wayman. Fundamentals of biometric authentication technologies. *Int. J. Image Graphics*, 1:93–113, 01 2001.

[2] Homayoon Beigi. *Fundamentals of Speaker Recognition.* Springer Publishing Company, Incorporated, 2011.

[3] Zhanibek Kozhirbayev, Berat Erol, Altynbek Sharipbay, and Mo Jamshidi. Speaker recognition for robotic control via an iot device. pages 1–5, 06 2018.

[4] Jyoti Singhai and Rakesh Singhai. Automatic speaker recognition: An approach using dwt based featureextraction and vector quantization. *IETE Technical Review*, 24(5):395–402, 2007.

[5] R. Togneri and D. Pullella. An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, 11(2):23–61, Secondquarter 2011.

[6] Rohan Kumar Das and SR Mahadeva Prasanna. Speaker verification from short utterance perspective: a review. *IETE Technical Review*, 35(6):599–617, 2018.

[7] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz, and Douglas A. Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4):101962, Apr 2004.

[8] Tomi Kinnunen and Haizhou Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52:12–40, 01 2010.

[9] SS Saraswathi and TVG Geetha. Language models for tamil speech recognition system. *IETE Technical Review*, 24(5):375–383, 2007.

[10] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing mel-frequency cepstral coefficients on the power spectrum. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 73–76 vol.1, May 2001.

[11] Siddhant C. Joshi and Dr. A. N. Cheeran. Matlab based feature extraction using mel frequency cepstrum coefficients for automatic speech recognition.

[12] Yuan Meng. Speech recognition on dsp: Algorithm optimization and performance analysis. 06 2019.

[13] Lindasalwa Muda, Mumtaj Begam, and I. Elamvazuthi. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *CoRR*, abs/1003.4083, 2010.

[14] Ben Gold and Nelson Morgan. *Speech and Audio Signal Processing.* 01 1999.

[15] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah, and Arpit Bansal. Feature extraction using mfcc. 2013.

[16] Bruce P. Bogert. The quefrency analysis of time series for echoes : cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. 1963.

[17] X. Zhang, J. Bai, and W. Liang. The speech recognition system based on bark wavelet mfcc. In *2006 8th international Conference on Signal Processing*, volume 1, Nov 2006.

[18] Y. Zhao, L. Zhang, J. Hu, and T. Liao. Mallat wavelet filter coefficient calculation. In *2013 International Conference on Computational and Information Sciences*, pages 963–965, June 2013.

[19] Smriti Srivastava, Saurabh Bhardwaj, Abhishek Bhandari, Krit Gupta, Hitesh Bahl, and J R. P. Gupta. *Wavelet Packet Based Mel Frequency Cepstral Features for Text Independent Speaker Identification*, pages 237–247. 01 2013.

[20] Ursula G. Goldstein. Speaker-identifying features based on formant tracks. *The Journal of the Acoustical Society of America*, 59:176–82, 02 1976.

[21] Fang Zheng, Guoliang Zhang, and Zhanjiang Song. Comparison of different implementations of mfcc. *J. Comput. Sci. Technol.*, 16:582–589, 11 2001.

[22] S Chakroborty, A Roy, and Goutam Saha. Improved closed set text-independent speaker identification by combining mfcc with evidence from flipped filter banks. *International Journal of Signal Processing*, 4:114–122, 01 2007.

[23] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, Jan 1995.

[24] Douglas Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41, 01 2000.

[25] Kuo-Hwei Yuo and Hsiao-Chuan Wang. Joint estimation of feature transformation parameters and gaussian mixture model for speaker identification. *Speech Communication*, 28:227–241, 07 1999.

[26] Arthur Dempster, Natalie Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 01 1977.

[27] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *International Computer Science Institute*, 4:126, 1998.

[28] Yoseph Linde, Andrés Buzo, and Robert Gray. An algorithm for vector quantizer design. *Communications, IEEE Transactions on*, 28:84–95, 01 1980.

[29] Jean-Luc Gauvain and Chin-Hui Lee. Lee, c.: Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. ieee trans. speech audio process. 2, 291-298. *Speech and Audio Processing, IEEE Transactions on*, 2:291 – 298, 05 1994.

[30] Zhang Jie, Li Guo-liang, Zheng Yu-zheng, and Liu Xiao-ying. A novel noise-robust speech recognition system based on adaptively enhanced bark wavelet mfcc. volume 4, pages 443 – 447, 09 2009.

[31] J S Garofolo, Lori Lamel, W M Fisher, Jonathan Fiscus, D S. Pallett, N L. Dahlgren, and V Zue. Timit acoustic-phonetic continuous speech corpus. *Linguistic Data Consortium*, 11 1992.

[32] Johan Koolwaaij. Speaker identification and assessment on the yoho database. 01 1997.

[33] Voxforge.org. Free speech... recognition (linux, windows and mac) - voxforge.org. `http://www.voxforge.org/`. accessed 06/25/2014.

[34] Sandipan Chakroborty and Goutam Saha. Improved text-independent speaker identification using fused mfcc & imfcc feature sets based on gaussian filter. *Signal Processing*, 35, 11 2009.

Edited by: Anand Nayyar
Received: Mar 3, 2020
Accepted: May 25, 2020