Sigma J Eng & Nat Sci 38 (4), 2020, 2177-2191

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article

PARABOLIC FILTER MEL FREQUENCY CEPSTRAL COEFFICIENT AND FUSION OF FEATURES FOR SPEAKER AGE CLASSIFICATION

Mohammed Muntaz OSMAN*[1], Osman BÜYÜK[2]

[1]*Kocaeli University, Department of Electronics and Communication Engineering, KOCAELI;*
ORCID: 0000-0001-6932-4159
[2]*Kocaeli University, Department of Electronics and Communication Engineering, KOCAELI;*
ORCID: 0000-0003-1039-3234

Received: 07.08.2020 Revised: 15.09.2020 Accepted: 19.10.2020

ABSTRACT

Speech is an acoustic signal initiated at the inner end of the human vocal tract and radiated as an audio wave at the tip of the outer end. The structure and length of the vocal tract makes distinctions on features taken from speeches similar in content, but uttered by different speakers. As a person grows his/her vocal tract changes in length which in turn modifies speech characteristics gradually. The mel frequency cepstral coefficient (MFCC) which uses triangular band pass filter banks has been widely regarded as the most popular feature used in most speech processing applications. To improve the accuracy of speaker age classification a new spectral based feature set named as parabolic filter mel frequency cepstral coefficient (PFMFCC) is proposed in this study. PFMFCC uses parabolic band pass filter banks instead of the triangular ones. This feature extraction technique uses 30 parabolic band pass filter banks to extract 42 features from each speech frame of length 20 ms. These features are applied to three classical classifiers, namely the Gaussian mixture model (GMM), cosine score, and probabilistic linear discriminant analysis (PLDA). The aGender database consisting of 47 hours of German speech uttered by a total of 852 speakers is used in this study. The new PFMFCC feature achieved 51.01%, 56.01% and 58.14% accuracies with cosine score, GMM and PLDA classifiers respectively on the female dataset. Similarly it achieved 50.44%, 52.74% and 57.23% accuracies with cosine score, GMM and PLDA classifiers respectively on the male dataset. Using feature fusion of seven feature sets overall accuracies of 60.18%, 52.17% and 56.35% are obtained on cosine score, GMM and PLDA classifiers respectively for all the seven speaker age classes. The feature fusion has improved the overall accuracy by 2.55% using cosine score compared to a related speaker age classification study carried out on the same database previously

Keywords: Parabolic filter, feature fusion, speaker age, classification, accuracy.

1. INTRODUCTION

Speech is the most convenient and natural way to communicate ever since humans exist. Apart from the content of a message, speech signal also carries information about the identity of speakers. Speaker age information is one of the identities. Speaker age classification or recognition is getting more interest among the research community recently for the following reasons. First, online commercials through internet and phone ads are rapidly spreading and

* Corresponding Author: e-mail: mohammedmuntaz@yahoo.com, tel: (553) 158 54 12

compelling an urgent need for better performance these days. Some companies advertise their products or services using the telecom service provider operated in a certain country. We sometimes receive automatic and anonymous phone calls from companies. If these calls are not up to our interest, we immediately reject them. But instead if we design a smart system that can predict whether the intended call recipient is child, young, adult or old from his/her short "hello" speech, the call will not be wasted and it will be effective. Once the age class of a "hello" utterance is recognized, an appropriate commercial will be retrieved from a database.

Second, speaker age classification can be used to narrow down the range of suspects in criminal investigation operations. Speech is used in some forensic studies as evidence. Before being used as a piece of evidence though, it is evaluated in terms of its authenticity, any modifications it includes, and its relevance to the goals of the investigation. Audio evidences can be gathered from different resources, such as an acoustical recording system, a call center recording, a voice mail message, or a surveillance tape acquired during a criminal investigation. These evidences assisted by speaker age classification system can help to reduce the range of suspects which leads to criminals eventually. In addition to targeted marketing and criminal investigation speaker age classification can be used in medicine, sporting events, user-profiling, human machine interaction, online dating, and other areas of service.

The first attempt to address the problem of speaker age classification was made six decades ago in the early 1950's [1]. However this problem was supported by computer aided systems dealt based on information obtained from speech only recently [2-3]. Speakers of two databases, JNAS and S(senior).JNA.S, were divided into two groups by listening tests [2]. Speakers whose speech sounds so aged that one should take special care when he/she talks to them were put in one group. The other group has the remaining speakers of the two databases. After that, each speaker group was modelled with GMM. Experiments of automatic identification of elderly speakers showed the correct identification rate of 91 %. To improve the performance, two prosodic features are considered, i.e., speech rate and local perturbation of power. Using these features, the identification rate is improved to 95%. Using scores calculated by integrating GMMs with prosodic features, experiments have been carried out to automatically estimate speakers' age. Accordingly high correlation between speakers' age estimated subjectively by humans and automatically calculated score of 'agedness' was reported [1].

Acoustic feature set for estimating speaker's age was developed on cross-sectional data [4]. Using the mean absolute error (MAE) metrics these feature sets offered 10 years difference between the actual chronological age and estimated age for both genders independently, whereas the performance is degraded to 12 years for combined gender experiments. MFCCs [5] extended by a set of prosodic features, pitch f0, and first four formant frequencies are used as baseline feature set. 220 features were obtained when these features are combined; then the 220 features are reduced by selecting the best feature subset by maximizing the R2 variance with R as correlation by using multiple regression/correlation analysis. Eventually, a mechanism is designed in their study to select the best subset composed of one feature, two features, and continues until there is no better subset. Their approach has been tested on the University of Florida Vocal Aging Database (UF-VAD) which contains 5 hours of speech for 150 different speakers and 1350 utterances of read English speech. This database has 3 age classes equally divided between males and females for young, middle-aged, and old age groups. Adding prosodic, pitch, and formant features to the MFCCs feature set improved the results by reducing the mean absolute error between 4-20% [5].

The modulation cepstrum coefficients, instead of using the cepstral coefficients are proposed for speaker age and gender classification [6]. Smooth information of the cepstral over a period of times is extracted from the speech utterance. Discrete cosine transform (DCT) is used over a fixed duration window. Which means, speech utterance in modulation cepstrum domain is filtered by decomposing the utterance cepstral trajectories into group of low and slow frequencies and the mel cepstral modulation spectrum (MCMS) features are extracted. It is reported that the low

modulation frequencies of MCMS (3-14 hertz) have the efficient information needed for age and gender classification. A comparison of these features with the conventional MFCC was made and an accuracy of 50.2% using the MCMS features was reported.

An iterative technique is used to define 30 parabolic band pass filter banks. These filter banks are applied the same way triangular filer banks are used in mel frequency cepstral coefficients (MFCC) which leads to the generation of new feature sets here in after named as parabolic filter mel frequency cepstral coefficients (PFMFCC). While the magnitude of the triangular filter banks starts falling sharply from the top (center), the magnitude falls gradually in parabolic filter banks. The latter is closer to practical filters than the former. It is more difficult to realize the triangular filter bank as it contains a sharp corner which needs a huge memory resource.

PFMFCC feature is applied to three known classical classifiers and gives a better performance than other features conducted on the same aGender database in PLDA classifier [7]. In line with our study, new transformed feature sets named as transformed MFCCs (T-MFCCs) generated by deep neural network (DNN) bottleneck extractor were proposed for speaker age and gender classification in a recent research study [8]. This scheme used a GMM universal background model (GMM-UBM) classifier as a backend and applied the experiment on aGender database. T-MFCCs achieved accuracies of 67.02, 62.16, 41.62, 72.97, 34.81, 52.97, and 71.89% for children, young female, young male, adult female, adult male, senior female, and senior male, respectively. The overall accuracy of T-MFCC feature was 57.63%. Fusion of PFMFCC with six other feature sets has improved this performance by 2.55%.

Although several frontend feature extraction techniques are used for speech recognition and speaker recognition including MFCC, it is a limitation that most of these feature sets have not been employed for speaker age classification. In this study we explore on performance evaluation of 8 feature sets on speaker age classification for the first time to the best of our knowledge. MFCC, inverted MFCC (IMFCC), rectangular filter cepstral coefficient (RFCC), linear frequency cepstral coefficient (LFCC) and relative spectral transform - perceptual linear prediction (RASTA-PLP) [9] are among the magnitude based features. Modified group delay (MODGD) and cosine phase feature sets are phase based features used in our experiment [10-11]. Some of these features are used for replay attack detection as well [12]. MFCC has been repeatedly used with variety of classifiers for speaker age classification as well as regression.

The maın contrıbutıon of thıs study can be clearly stated in two points:

1. Develop an algorithm which creates a set of parabolic band pass filter banks and use these filter banks to generate a unique set of features called PFMFCC.
2. Explore the performance of selected features and their fusion for speaker age classification and compare the results with the newly proposed feature set.

The remaining part of this article is organized as follows. In section 2 feature extraction techniques are discussed. The PFMFCC is explained in depth and comparative analysis is made with MFCC. In section 3 three classical classifiers are presented. Section 4 discusses how the experimental setup is made. In section 5 a brief discussion of results and performance of feature fusion of PFMFCC with other magnitude and phase based spectral features is presented. In section 6 a concluding remark is made.

2. FEATURE EXTRACTION

Feature extraction is a set of operations applied on sequence of numbers taken from a sample speech to compress its size without losing important information. For instance a randomly picked 3.148s utterance from our database, which is sampled at 8 KHz, contains a row vector with 25184 number of values. However framing this sequence using hamming windows of 20 ms and extracting 42 important features from each frame removes 118 less relevant sequence values from each frame. For the sake of simplicity the setup of any feature extraction process in our study is presented symbolically as shown in fig. 1 below.

Details of each block in the setup shown in fig. 1 below are discussed in section 5. However this section presents a highlight about every block and a brief discussion about the core block here. At block A our main routine accesses a text file containing all the paths to audio samples collected from four age categories; children, young, adults and elders both from male and female genders. The program reads the first line of this text at block B. Following this our experimental setup checks if the accessed line is not an empty line. The previous two blocks are a onetime operation whereas the subsequent blocks are iterative until reading each line of the text file is completed. The program fetches utterances from the path accessed in block B or F. The core of this process is at block D where feature extraction is executed. At block E the system writes features on to text files before the system continues the process all over again excluding block A and B.

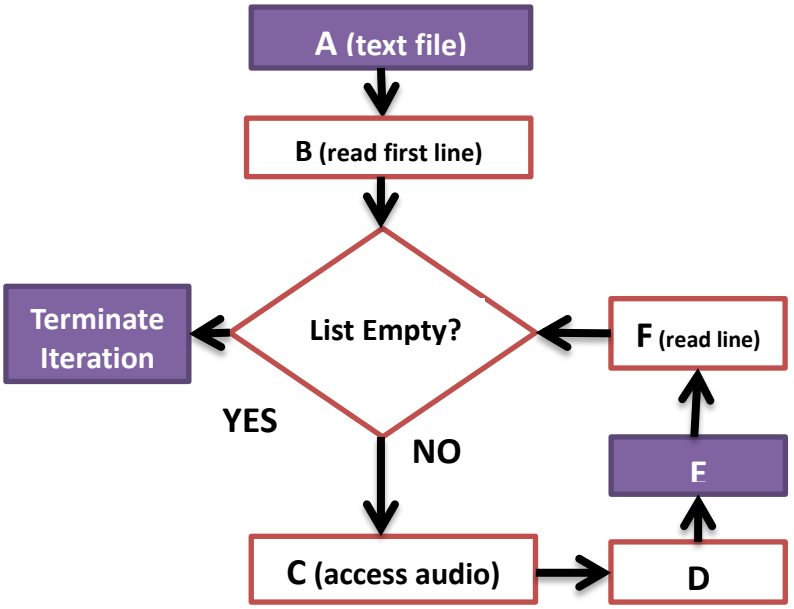


Figure 1. The setup of feature extraction operations.

The main purpose of this section is to present the feature extraction block represented by D shown in fig. 1 above. Every feature extraction scheme in our study follows similar procedures up to the stage of filter banks as shown in fig. 2. Filter bank based features follow the whole process in the diagram whereas other features such as MODGD, RASTA-PLP, etc. follow only until the discreet Fourier transform (DFT) stage. The new PFMFCC feature proposed in this study differs from other filter bank based features in the shape of the band pass filter banks it uses.

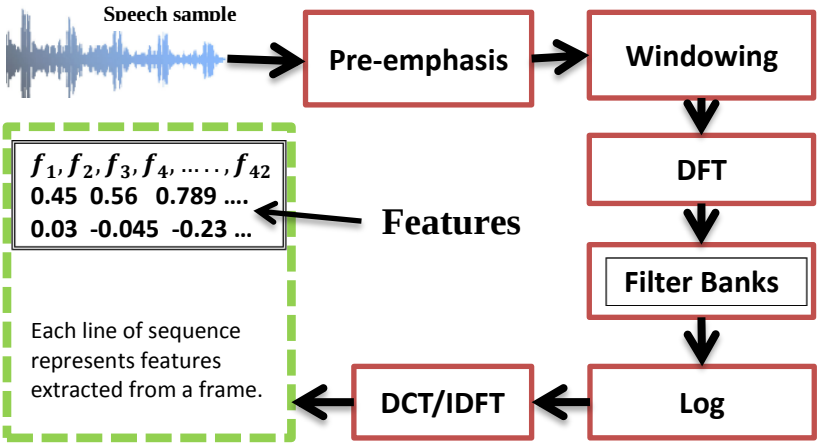


Figure 2. Complete feature extraction process

As shown in fig. 2 above, the process of feature extraction begins with applying a high pass finite impulse response (FIR) pre-emphasis filter on the entire speech signal to amplify the high frequency components, to balance the frequency spectrum, to avoid numerical problems during DFT operation and to improve the signal to noise ratio (SNR). This operation is described as follows:

$$y[n] = x[n] - \alpha x[n-1] \tag{1}$$

where $x[n]$ represents the entire speech signal before framing uttered by a particular speaker and α is pre-emphasis factor which can be determined using the expression given in (2) based on the values of the cut-off frequency F_c and the slope S in logarithmic scale. In this experiment the values of F_c and S are set to be 2122 Hz and 6dB/oct respectively. Hence the α value is calculated as 0.97.

$$\alpha = e^{\frac{-2\pi S}{F_c}} \tag{2}$$

Speech signal is highly dynamic in its entire scope. To avoid this splitting the signal as small as we can get a static short frame helps for further processing. In this study a hamming window of length 20 ms with a frame shift of 10 ms is chosen to split the entire speech signal in to smaller frames [13]. The 10 ms frame shift is used to create an overlap between adjacent frames which helps to avoid unexpected outcomes. Although the size of each frame is equal the number of frames from each speech sample varies depending on the duration of the utterance.

Calculating the DFT is the crucial step in many speech processing applications. This is mainly due to the nature of speech signal which becomes evident in its frequency domain description. It converts a finite sequence of uniformly-spaced samples in each frame in to a same length of uniformly-spaced samples of discrete time Fourier transform (DTFT) [13]. Since DFT is a complex valued function of frequency, it can be decomposed in to real and imaginary or magnitude and phase components eventually leading to many feature sets in further processing each of these components. Mathematically this process can be described as follows:

$$X[j,k] = \sum_{n=0}^{N-1} y[n] w[n-j]\, e^{-j\frac{2\pi kn}{N}}, 0 \le k \le L-1 \; and \;\; j = 1, 2, \ldots, M \tag{3}$$

where $X[j,k]$ is the DFT of the j^{th} frame, M is the number of frames in a certain utterance which varies according to the duration of the speech, N is the DFT point and $w[n-j]$ is the j^{th}

framing hamming window having equal length L for all j which is given by (4). The proceeding stages from this stage are applied on each frame rather than on the entire signal at one time.

$$\mathrm{w[n]} = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right), & 0 \le n \le L-1 \\ 0, & otherwise \end{cases} \quad (4)$$

Unique contribution is made at the filter bank stage in this study which resulted in a different set of features. The most commonly used triangular band pass filter banks in MFCC are replaced by parabolic filter banks inverted down and shifted to the right based on mel scaling of range of frequencies. The general description of the i^{th} filter bank function $H[i,k]$ is given in a compact form as:

$$H[i,k] = -A_i(k - f_i)^2 + B \quad (5)$$

In order to obtain the maximum and the vertical line of symmetry of this function a first derivative with respect to k must be applied and equated to zero. This leads to the following equation:

$$H'[i,k] = -2A_i(k - f_i) = 0 \quad (6)$$

which in turn leads to the point where the maximum of the function occurs. The vertical line $k = f_i$ is the line of symmetry at the same time the maximum value of the i^{th} parabolic function occurs here. And the maximum point is (f_i, B). The intercepts to the horizontal axis are determined based on the values of the partition made on the frequency range(f_{min}, f_{max}). The minimum frequency is set to be 0 and the maximum is half of the sampling frequency used in the speech database which can be written as (f_{min}, f_{max}) =$(0, 4000Hz)$.

The intercept values determine the parameter A_i in each parabolic function. The value of the function below the minimum and beyond the maximum intercept should be set to zero. The maximum value B remains the same in all the band pass filter bank functions. The number of filter bank functions is set to be 30 in our experiment. Once the entire range of frequencies is converted in to mel scale using (7), this range is partitioned in to 30 smaller band of frequencies. It is known that the mel scale relates the perceived frequency to the actual measured frequency. The human ear is better at identifying small changes in speech at lower frequencies. The converted minimum and maximum frequency pair in mel scale is (0, 2146).

$$mel(f) = 1127\ln(1 + \frac{f}{700}) \quad (7)$$

The value of B is set to be 1 which is 0 dB. Substituting the values of B and f_i in the filter function given in (5) above, it is possible to determine the amplitude values A_i in each filter function equating it to zero and evaluating the equation at intercept points. The intercept points are the ages of the 30 filter banks. After partitioning (0, 2146) in to 30 smaller bands of frequencies a back conversion of these bands is needed as the final operation is carried out in terms of actual frequency instead of mel scale. The back conversion to actual frequency is done using (8). Where m represents individual mel scale values at the edge of each filter partition.

$$mel^{-1}(m) = 700(e^{\frac{m}{1127}} - 1) \quad (8)$$

In a mel scale each sub partition has a uniform spacing of 71.5. For instance the first five filter banks can be described by these edge parameters in a mel scale; (0, 143), (71.5, 214.5), (143, 286), (214.5, 357.5) and (286, 429). These edges have to be converted back to actual frequency measurement and their DFT indices should be calculated using (9).

$$f_i = \left\lfloor \frac{(N+1)mel^{-1}(m)}{F_s} \right\rfloor \quad (9)$$

The filter bank functions in (5) above need to be redefined considering the intercept points of the functions at the edges of each filter. The value of each filter bank function is made to vanish out of the ranges of the intercepts. Substituting the right edge of each filter f_{i+1} in the values of k

and equating it with zero results in $A_i = (f_{i+1} - f_i)^{-2}$, which eventually give the relation described in (10). Since the parabola in each function is symmetrical with the line $k = f_i$, the distance from the center to both eadges is equal.

$$H[i,k] = \begin{cases} -(\frac{k-f_i}{f_{i+1}-f_i})^2 + 1, & f_{i-1} \le k \le f_{i+1} \quad , \quad i = 1,2,3,\dots,30 \\ 0, & otherwise \end{cases} \tag{10}$$

Where the left most edge f_0 and the right most edge f_{31} of all the filter banks are 0 and 4000 respectively. In contrast to this $H[i,k]$ functions are given by equation (11) below for the conventional triangular filter banks. While the former is continuous throughout the sub band of a certain filter bank the later encounters a sharp corner at the center of the function as a result needs two mathematical functions to define the entire sub band. The graphs of both triangular and parabolic filter banks are shown in fig. 3 in the same order. In both cases the spacing is uniform until the thousandth frequency, and then it increases in each succeeding filter bank.

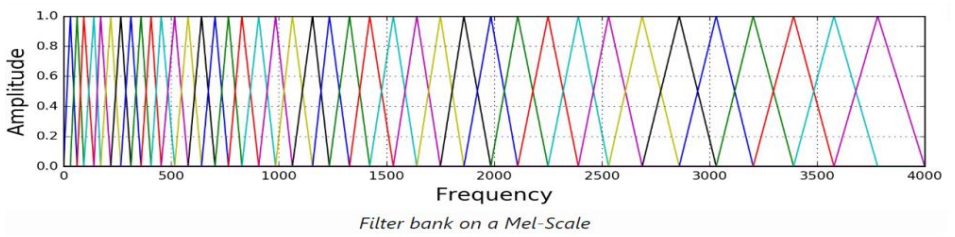
$$H[i,k] = \begin{cases} \frac{k-f_{i-1}}{f_i-f_{i-1}} & , \quad f_{i-1} \le k \le f_i \\ \frac{f_{i+1}-k}{f_{i+1}-f_i} & , \quad f_i \le k \le f_{i+1} \\ 0 & , \quad otherwise \end{cases} \tag{11}$$

Before finalizing this process with discrete cosine transform (DCT) and applying log on to it, the parabolic filter bank functions given in (10) are applied on the DFT described in (3). The filter bank effect on the power spectrum $\boldsymbol{X[j,k]}$ of each frame is given by (12). The effect is summed over the lower L_i and upper U_i frequencies of each particular filter bank i.

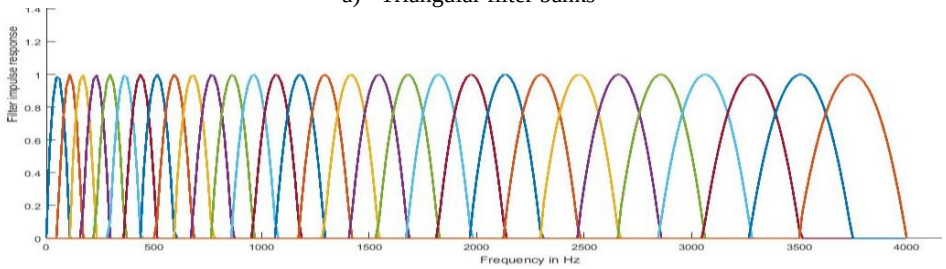
$$MF[i] = \frac{1}{A_i}\sum_{k=L_i}^{U_i} |H[i,k]\boldsymbol{X[j,k]}| \quad , \quad i = 1,2,3,\dots,30 \tag{12}$$

$$A_i = \sum_{k=L_i}^{U_i} |H[i,k]|^2 \tag{13}$$

$$\boldsymbol{X[j,k]} = \frac{1}{N}|X[j,k]|^2 \quad , \quad power\ spectrum\ of\ each\ frame \tag{14}$$



a) Triangular filter banks



b) Parabolic filter banks

Figure 3. Triangular and parabolic band pass filter banks

The last and most important stage of this feature extraction scheme is calculating the M static features using cepstral transform which in turn applies logarithm on $MF[i]$ values calculated in (12) above and uses either DCT or inverse DFT for eventual generation of these features. Equation (15 describes all this process. After collecting the M static values equation (16) is used to extract the M dynamic (delta) features from static features. The same approach is used to determine the M acceleration (double delta) features from dynamic features and shown in (17). In addition to the total 3M features, an energy component for each of static, dynamic and acceleration features are calculated which makes the total $3(\mathrm{M}+1)$. This study is made on a total of 42 features setting the M value to 13.

$$MFCC[m] = \frac{1}{R}\sum_{i=1}^{R} \log(MF[i])\, cos\left[\frac{2\pi}{R}\left(i+\frac{1}{2}\right)m\right] \ , m = 1, 2, 3, \ldots, M \tag{15}$$

$$delta[t] = \frac{\sum_{n=1}^{Q} n(MFCC_{t+n} - MFCC_{t-n})}{2\sum_{n=1}^{Q} n^2} \ , \quad Q = 2 \tag{16}$$

$$double\_delta[t] = \frac{\sum_{n=1}^{Q} n(delta_{t+n} - delta_{t-n})}{2\sum_{n=1}^{Q} n^2} \quad , \quad Q = 2 \tag{17}$$

where R is the total number of band bass filter banks used which is 30, t is an index used to identify adjacent frames whereas $t+Q$ and $t-Q$ are indexes of the farthest neighbour frames involved in calculation of the t^{th} frame dynamic and acceleration features.

3. CLASSIFICATION

A number of classical and modern classifiers have been being applied on speech features in order to make decisions over the years. The decisions can be summarized as regression or classification. Among the most widely used classifiers the Gaussian mixture model (GMM), support vector machine (SVM), k-nearest neighbor (KNN), cosine distance score are some of them.

The universal background model (UBM) which provides mean supervector (m) and helps us to determine the total variability matrix (TV) as an input to cosine distance score and probabilistic linear discriminant analysis (PLDA) is shown in fig. 4 below [14-16]. Both of these classifiers are used in this study. As shown in the figure, the method retrieves features from text files stored in a development set. The development set is allocated to train the supervector m and TV from our database. Each line of a certain text file in the database represents features extracted from a short frame of speech. UBM uses the expectation-maximization algorithm (EM) to determine TV and m which are used in determining model i-vectors for each class of speakers in the training set [17]. These two parameters are also involved in calculating i-vectors for each utterance of the test set. the mathematical expression relating l variability (TV) matrix T, super vector m independent of channel and speaker age, speech features x_i which depend on both channel and speaker age, and identity vectors (i-vector ω_i) is given by:

$$x_i = m + T\omega_i \tag{18}$$
$$\omega_i = (x_i - m)T^{-1}$$

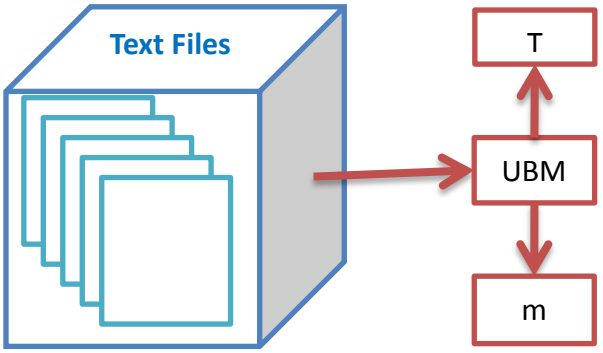


Figure 4. Universal background model (UBM)

Cosine score and PLDA classifiers use i-vectors extracted from each utterance of the training set to determine the model i-vector for each speaker age class. The later uses probabilistic LDA in addition to apply length normalization and data whitening. In addition to the model i-vector and test i-vector parameters used by cosine distance computation the PLADA classifier includes a structure variable containing the PLDA hyperparameters: plda.phi for Eigen voice matrix, plda.sigma for covariance matrix of residual noise, plda.mu for mean and plda.w for whitening transform. More over PLDA calculates i-vectors from development set which are needed to determine a low dimensional matrix in order to transform training model as well as test set i-vectors to a reduced dimension. Microsoft speaker recognition tool box is used for this purpose [18].

The Gaussian mixture model (GMM) calculates the scores for each test set sample using log likelihood ratio. The ratio is done comparing a single test sample with each of the speaker age classes modeled using the training set. Hence a score close to 1 shows high similarity and close to zero shows the rest sample does not belong to the model speaker age class under consideration.

Unlike GMM cosine score and PLDA classifiers used i-vectors calculated above in (18) which transform the variable length feature matrix in to uniform dimension in order to determine the score. The cosine score uses cosine distance computation between two vectors. The two vectors here are i-vector from each test utterance $\omega_{\text{test\_k}}$ and the model i-vector $\omega_{tarclass\_i}$ calculated as the mean of i-vectors for each speaker age class. This computation is shown in (19) and (20) below

$$\omega_{\text{tarclass\_i}} = \frac{1}{M_i}\sum_{k=1}^{M_i} \omega_{\text{train\_k}} \tag{19}$$

$$\cos\_\text{score}_{\text{k\_i}} = \frac{\omega_{\text{test\_k}}{}^{T} * \omega_{\text{tarclass\_i}}}{\|\omega_{\text{test\_k}}\|\|\omega_{\text{tarclass\_i}}\|} \tag{20}$$

where M_i represents the number of utterances in the i^{th} speaker age class of the training set.

4. EXPERIMENTAL SETUP

The aGender database excluding the child class is used for each gender separately for each standalone feature set in this study [7]. The database consists of a total of 6804 utterances from children age 7-14, 23779 utterances from female speakers of age 15-80 and 22493 utterances from male speakers of age 15-80. The audio was recorded over cell phones and landline connections in 8000 Hz, 8 bit alaw format. The male and female datasets are further classified in to three categories as young (ages: 15-24), adult (ages: 25-54) and old or senior (ages: 55-80). A total of 852 German speakers (at least 100 speakers in each class) have participated in the audio

recording which accounts to 47 hours of speech. All the seven classes including children class ranging from age 7 to 14 are considered to evaluate the overall performance of fusion of features.

The distribution of utterances in each class for development, training and test sets is presented in table 1 below. Due to the lack of labeling on the test set of the original dataset we received the training dataset is split in to training and test sets. A total of 131, 186 and 18 speakers were used in the development, training and test sets respectively for female gender. Likewise 130, 199 and 15 speakers were used in development, training and test sets for male gender. Once the path to each speech utterance in the three datasets is established matlab commands are exploited to trace and pick for processing. A total of 9644 audio samples are used to train the total variability subspace matrix (TV) in the female speaker age classification experiment. Similarly 8505 utterances are used in the male gender.

A hamming window of length 20 ms with 10 ms overlap is used for framing utterances [19]. 512 DFT point and a total variability dimension of 200 are applied. 13 static, 13 dynamic and 13 acceleration features are extracted from each frame. This makes up a total of 42 features including an energy component for each of the three feature sets.

Table 1. Distribution of utterances along development, training and test sets in each class.

Age Classes	Development set	Training set	Test set	total
Child 7-14	2397	4000	407	6804
Female 15-24	2722	4254	384	7360
Female 25-54	3361	4187	386	7934
Female 55-80	3561	4544	380	8485
Male 15-24	2170	3631	388	6577
Male 25-54	2512	4051	366	7295
Male 55-80	3826	5224	325	9700
Female Total	9644	12985	1150	23779
Male Total	8508	12906	1079	22493
Grand Total	20549	29891	2636	53076

The utterances vary in length from 0.28 to 11.3 second. The average duration is 2.55 second.

After computing scores for each test set using the three classifiers performance of each classifier is evaluated using accuracy. Confusion matrix is generated for each gender and feature-classifier pair. The mean of the diagonal values of these matrices gives the accuracy. A shell script in combination with a code written using Perl is used for this purpose.

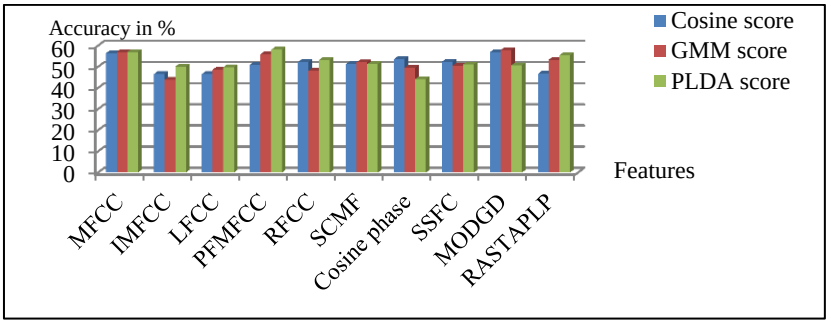
5. RESULTS AND DISCUSSION

5.1. Results

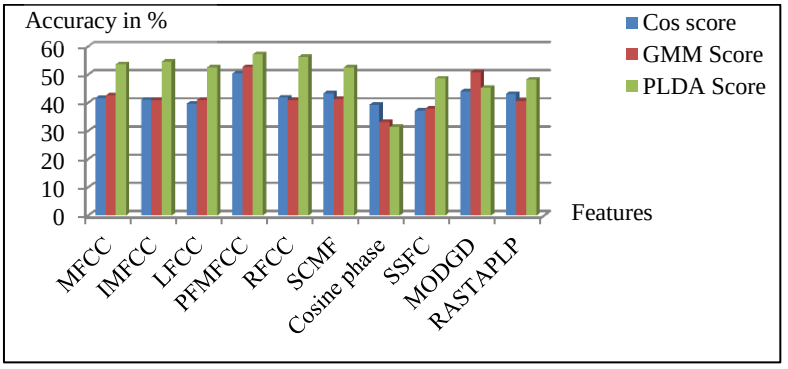
Fusion and standalone Performance evaluation results of 10 features and 3 classifiers are presented in this section. Except the popular MFCC all the other features have never been used for speaker age classification before to the best of our knowledge. We developed an algorithm that creates a set of band pass filter banks which closely resemble very much like practical filters. We used these filter banks to generate a unique set of features using all the steps in MFCC feature extraction. These newly generated feature sets are named after the filter bank shape used. Hence they are called as parabolic filter mel frequency cepstral coefficients (PFMFCC).

In the female dataset the choice of the backend process does not affect the performance significantly. Notable frontend feature extraction techniques which performed relatively better than others include MFCC, PFMFCC and MODGD. The best performance from all feature-

classifier pairs in female dataset is obtained using PFMFCC-PLDA which resulted in an accuracy value of 58.18%. Unlike the female dataset the classification process on the male dataset has shown significant variation across the three backend classifiers. The PLDA classifier performed better than the other two in all feature sets except in the phase based features; cosine phase and MODGD. The best performance is obtained again using PFMFCC-PLDA feature-classifier pair which resulted in an accuracy of 57.23%. The performance evaluation results are presented in fig. 5 a) and b) for female and male datasets respectively. These figures include performance evaluation of two additional feature sets which characterize subband energy named as subband spectral flux coefficient (SSFC) and subband centroid magnitude and frequency (SCMF) [20-22].



a) Performance evaluation on female test set from simulation results



b) Performance evaluation on male test set from simulation results

Figure 5. Female and male performance evaluation collected from confusion matrices of simulation results

Applying voice or speech activity detection (VAD or SAD) removes non-speech frames and reduces the number of frames in each original utterance. Although its effect was supposed to improve the performance, the PLDA classifier has shown significant degradation due to VAD in all feature sets except SSFC for male dataset as shown in fig. 6 below. The variation of performance in the other two classifiers due to VAD is insignificant both in male and female datasets. Despite smaller energy compared to other frames, non-speech frames have contributed for the PLDA classifier constructively.

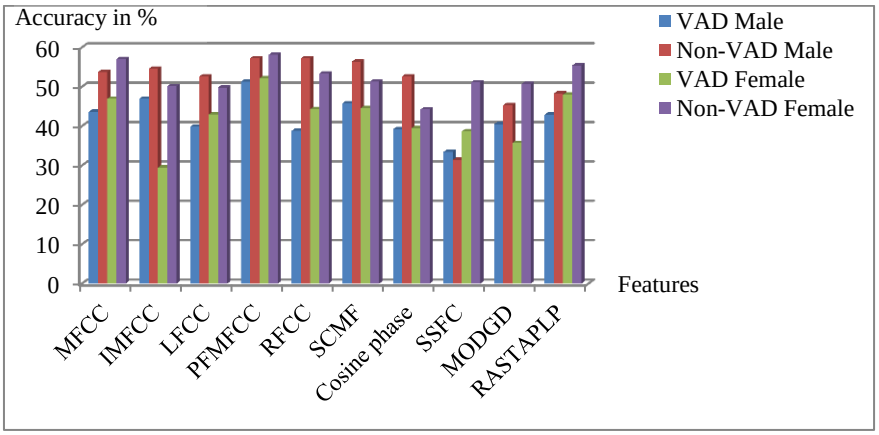


Figure 6. Effect of VAD on the PLDA classifier for male and female datasets from simulation results

In addition to stand alone performance a random selection of features were made and feature fusion was carried out on different kinds of features. At first a fusion of all the ten features was done using concatenation which resulted in a performance below the best performance of a single feature set. In the next steps few feature sets that showed the worst performances were removed one feature set at a time and arrived at a fusion of seven feature sets that consists of MFCC, MODGD, RFCC, SCMF, SSFC, RASTA-PLP and PFMFCC. VAD is not applied on these features. The cosine score classifier gave the best performance in both genders. The Matlab simulation result carried out on all the seven classes consisting children aged 7-14, young female aged 15-24, adult female aged 25-54, old female aged 55-80, young male aged 15-24, adult male aged 25-54 and old male aged 55-80 resulted in overall accuracies of 60.18%, 52.17% and 56.35% using cosine score, GMM and PLDA respectively. The age classes are made based on the aGender database [7]. According to this result the cosine score classifier has made an overall improvement of the accuracy by 2.55% compared to speaker age classification study carried out on the same database in [8].

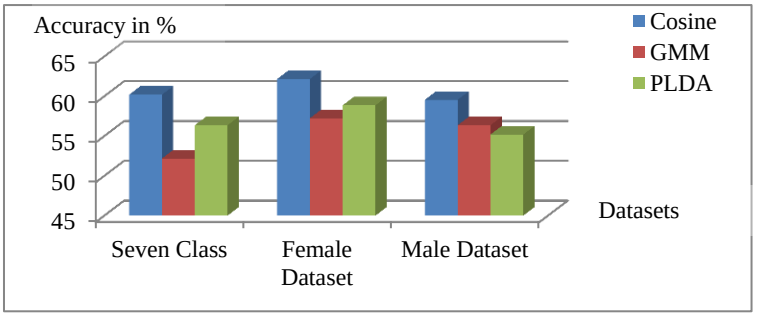


Figure 7. Performance evaluation results of feature fusion of seven feature sets on three classifiers

5.2. Discussion

After critically examining most research studies carried out on speaker age classification we found out that they heavily depend on MFCC features. Their focus is on the backend (classifiers). Having observed encouraging results from a couple of other features for speech recognition, speaker recognition and replay attack detection; we inquired if these features could perform well on speaker age classification and decided to apply them in our study. After observing their performances we got motivated to develop techniques to generate a new set of features using parabolic band pass filter banks. Table 2 below presents the summary of comparisons between the popular MFCC with our proposed feature set PFMFCC. Based on this summary we can say that PFMFCC contains more age information than MFCC.

Table 2. MFCC vs PFMFCC

Criteria	**MFCC**	**PFMFCC**
Filter bank shape	Triangular	Parabolic
No of function used per filter bank	2 linear functions	A single polynomial function of degree 2
Number of features in a frame	13 static + 13 dynamic + 13 acceleration + 3 Energy components = 42	13 static + 13 dynamic + 13 acceleration + 3 Energy components = 42
Performance for female dataset	Cosine score 56.44%, GMM 57.03%, PLDA 57.03%	Cosine score 51.06%, GMM 56.01%, PLDA 58.14%
Performance for male dataset	Cosine score 41.8%, GMM 42.63%, PLDA 53.75%	Cosine score 50.44%, GMM 52.74%, PLDA 57.23%
Realizability	Not easier to implement as it consists a sharp corner.	Can be approximated with practical filters.

The performance variations across different feature sets imply that further processing of either the phase or magnitude components of power spectrum of utterances could lead to a much better performance. This clearly indicates that much effort in frontend analysis is needed more than the choice of backend classifiers.

An apparent limitation of our study lays on the difficulty of finding convenient boundaries especially between adult and old speaker age classes where one speaker age class ends and the next one begins. For instance putting age 55 together with age 80 and putting age 54 in a different speaker age class is a challenging decision. This problem could be solved by using regression instead of classification if we had a large number of utterances which can represent each and every age very well in the training sets. Another limitation of our study is the short duration of utterances (2.55 second on average) which also includes non-speech frames within it. VAD further reduces the number of frames in an utterance by discarding the non-speech frames that have energy below a certain threshold and PLDA also reduces dimensıon due to its linear discriminant analysis (LDA) function within it. The performances of features with and without VAD on PLDA classifier as shown in fig. 6 indicates that PFMFCC and fusion of features could perform better on a database with longer duration of utterances.

6. CONCLUSION

In this study we have proposed a new filter bank based feature extraction technique named as parabolic filter MFCC (PFMFCC) for speaker age classification and performance evaluation of 10 feature sets including PFMFCC is carried out. PFMFCC gave better performances both in male and female datasets using the PLDA classifier compared to other feature sets in our study.

Accuracies of 58.18% and 57.23% are obtained using PFMFCC-PLDA feature-classifier pair for female and male datasets respectively. On the female dataset choice of feature sets rather than classifiers has shown significant variations on performance whereas selection of classifiers significantly affected the performance rather than feature sets on male datasets as the PLDA classifier gave the better performance on all magnitude based feature sets. Based on our simulation results we can conclude that PLDA classifier with magnitude based feature sets is preferable for male speaker age classification.

Applying VAD has played significant role in degrading the performance of the PLDA classifier in all feature sets except the SSFC for female dataset. VAD effect is insignificant on other classifiers and inconsistent with selection of feature sets and gender. Therefore, ignoring VAD is advised while simulating with PLDA.

Merging features together and creating feature fusion has significantly improved the performance of the cosine score classifier. Whereas, it is inconsistent to make a conclusion for the other two classifiers. An overall accuracy of 60.18% is achieved with feature fusion of 7 features using cosine score applying on all the seven speaker age classes. This has shown an overall improvement of the accuracy by 2.55% compared to speaker age classification study carried out on the same database previously.

Further studies should investigate on speaker age classification performance of time-delay neural network (TDNN) which is used to compute speaker embedding from variable length utterances. This fixed-length speaker embedding is called x-vector. X-vector is state-of-the-art method which can be used instead of i-vector. In addition regression instead of classification is advised with large data size in order to avoid boundary problems in future studies. Further studies could also explore the effects of applying dimensionality reduction techniques such as principal component analysis (PCA) for feature selection. Furthermore DNN classifiers of different models can be applied on the feature sets independently or fused.

REFERENCES

[1] Mysak, Edward D., (1959) Pitch and duration characteristics of older males, *Journal of Speech& Hearing Research*, 2(1), pp.46-54.

[2] Minematsu, Nobuaki, M. Sekiguchi, and K. Hirose, (2002) Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers, *in IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, vol. 1, pp. I-137-I-140.

[3] Muller, Christian, F. Wittig, and J. Baus, (2003) Exploiting speech for recognizing elderly users to respond to their special needs, *in Eighth European Conference on Speech Communication and Technology*, pp. 1305-1308.

[4] Spiegl, Werner, G. Stemmer, E. Lasarcyk, V. Kolhatkar, A. Cassidy, B. Potard, et al., (2009) Analyzing features for automatic age estimation on cross-sectional data, *In INTERSPEECH* 2009, vol. 10, pp. 2923-2926.

[5] Li M, Jung C-S, Han KJ , (2010) Combining five acoustic level modeling methods for automatic speaker age and gender recognition, *In: INTERSPEECH2010*, pp. 2826–2829.

[6] Ajmera, J., Burkhardt, F., (2008) Age and gender classification using modulation cepstrum, *In: Proc. Odyssey*, pp. 025.

[7] F. Burkhardt, Eckert, M., Johannsen, W. and J. Stegmann, (2010) A database of age and gender annotated telephone speech, *Proceedings of the Language and Resources Conference* (*LREC*).

[8] Mallouh, Arafat Abu, Zakariya Qawaqneh, and Buket D. Barkana, (2018) New transformed features generated by deep bottleneck extractor and a GMM–UBM classifier for speaker age and gender classification. *Neural Computing and Applications* 30(8): pp. 2581-2593.

[9] H. Hermansky and N. Morgan, (1994) RASTA processing of speech, *IEEE Transactions on Speech and Audio Processing*, 2(4): pp. 578–589.

[10] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, (2006) Significance of the modified group delay feature in speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1): pp. 190–202.

[11] R. Schluter and H. Ney, (2001) Using phase spectrum information for improved speech recognition performance, *in 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings* vol. 1, pp. 133–136.

[12] C. Hanilçi, (2017) Features and classifiers for replay spoofing attack detection, *in 2017 10th International Conference on Electrical and Electronics Engineering (ELECO)*, pp. 1187–1191.

[13] Harris, Fredric J. (1978) On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1): pp. 51-83.

[14] Douglas A. Reynolds, T. F. Quatieri, and R. B. Dunn, (2000) Speaker verification using adapted Gaussian mixture models, *in Digital Signal Processing*, Vol. 10, pp.19–41.

[15] K. W. Gamage, V. Sethu, P. N. Le, and E. Ambikairajah, (2015) An i-vector GPLDA system for speech based emotion recognition, *in 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 289–292.

[16] GRZYBOWSKA, Joanna; KACPRZAK, Stanislaw, (2016) Speaker Age Classification and Regression Using i-Vectors. *In: INTERSPEECH*. pp. 1402-1406.

[17] Moon, T. K. (1996) The expectation-maximization algorithm, *IEEE Signal processing magazine*, *13*(6), 47-60.

[18] Sadjadi, Seyed Omid, Malcolm Slaney, and Larry Heck. (2013) MSR identity toolbox v1. 0: A MATLAB toolbox for speaker-recognition research. *Speech and Language Processing Technical Committee Newsletter* 1(4): pp. 1-32.

[19] Saini, J., & Mehra, R., (2015) Power spectral density analysis of speech signal using window techniques. *International Journal of Computer Applications*, *131*(14), 33-36.

[20] Lie Lu, Hong-Jiang Zhang, and Hao Jiang, (2002) Content analysis for audio classification and segmentation, *IEEE Transactions on Speech and Audio Processing*, 10(7): pp. 504–516.

[21] K. K. Paliwal, (1997) Spectral subband centroids as features for speech recognition, *in IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 124–131.

[22] Kua JM, Thiruvaran T, Nosratighods M, Ambikairajah E, Epps J., (2010) Investigation of spectral centroid magnitude and frequency for speaker recognition, *In Odyssey*-2010, paper 007.