Chapter-5

Improved filter bank on multitaper framework

It has been demonstrated already that the heterogeneous feature extraction with hybrid HMM classifiers performs better on large vocabulary isolated Punjabi speech corpora. The fusion of different feature extraction approaches in mismatched train and test conditions achieved improved results in front end approaches. However, they still face the challenge of processing the speech information in real acoustic environment. In this chapter, the issue of robustness of acoustic information is addressed for practical dataset of the proposed approaches. Earlier research practices resulted into the challenge of increase of variance with leakage of spectral information through front end. Also, the handling of large number of feature information creates confusion with large speech vocabulary. To overcome this, PCA based multi-windowing techniques are integrated in baseline GFCC and MFCC based feature approaches systems. The proposed integrated approaches result into better feature vectors, which are further processed using DE+HMM based modeling classifier. The integrated subsystems show better performance for word recognition as compared to the conventional or fused feature extraction systems.

5.1 Introduction

The previous work shows that the progress made so far are implemented over mismatched trained and test conditions. However, the performance is not acceptable over large speech database. Therefore, there is a need to improve for generation of unique and compact feature vectors in speech recognition system and this is a challenging task. To improve the proposed system performance, the dissimilarity of environment in actual acoustic signal and correlated variable of the extracted feature

information are managed using PCA optimized improved multi tapered GFCC and MFCC approaches. The conventional and fused feature extraction approaches face the issue of high variance and low bias. In this study, the speech signal is carefully processed at the front end. Multi tapered windowing is implemented to reduce the variance and balance the spectral bias. Semi-processed information is further optimized using PCA established filter bank. The operation of multi taper is integrated using three approaches - Multipeak (Hansson, M., & Salomonsson, 1997), Thomson (Thomson, 1982), and SWCE (Hansson-Sandsten, M., & Sandberg, 2009). These multi tapered approaches are presented by optimizing their taper value to obtain better results in different techniques when fused inside the conventional MFCC and GFCC. An optimal value of taper is generated to produce low value of variance spectrum estimator that outruns orthogonal window periodgram. Further, the filter bank value is also optimized using PCA approach thus producing approximate solution with better results. In literature, numbers of aspects (such as bandwidth, shape, and positions of filters) are reformed to generate better feature vectors (Hung, 2004; Lee et al. 2001; Minh, V. D., & Lee, 2004). In this chapter, an attempt is made to handle the mismatch between auditory environment by integrating multi taper approaches and PCA in first stage of the feature extraction phase. Later DE+HMM approach is implemented at modeling phase. Rest of the chapter is organized as follows: Section 5.2 describes the background of robust speech feature analysis approaches, multi taper spectrum estimation approaches, selection of taper parameters and Section 5.3 discusses PCA based filter bank optimization process. Section 5.4, proposes an integrated Punjabi-ASR system. Sections 5.5 and 5.6 provide the detailed corpus setup and experiment analysis. Finally, the summary is presented in Section 5.7.

5.2 Robust speech analysis approaches

5.2.1 Gammatone frequency Cepstral Coefficients

Speech signals contain different information. The extraction of relevant information from an uttered signal is a major challenge in acoustic varied environment. GFCC is one of the widely accepted technique that converts signal from its time domain to its frequency domain. It extracts the robustness of the spoken utterance through cepstral analysis process with the help of gammatone filter bank. This section, briefly discusses the conventional GFCC based feature extraction process through step-wise procedure described in Algorithm 5.1. It adopts gammatone filter which models human cochlear filtering. The process is initiated after various levels of operations - implication of windowing, Fast Fourier Transform (FFT) calculations, Gammatone Filter bank analysis (GF), log function and Discrete Cosine Transform (DCT) (Maganti, H. K., & Matassoni, 2010; Schluter et al. 2007). An input speech utterance is fixed after implication of a windowing operation on it. It adopts single windowing technique that results into a frame of 10 to 15 ms. It perceives an assumption of a stationary audio signal for a non-stationary signal. It also includes the Gammatone (GT) filter bank analysis on frequencies which are sensitive to human ear. The output of GT filter bank is log processed that are unassociated through the process of DCT. Eventually it helps in energy efficient resultant feature vectors.

Algorithm 5.1: Conventional GFCC method

Step 1: Initialize the sampling frequency (fs) $= 16000$, number of channels (num_chan) $= 64$.

Step 2: Generate Gammatone filter bank.

2.1 Initialize the order of filter.

2.2 Evaluate the upper and lower bound of equivalent rectangular bandwidth.

2.2.1 Convert frequency scale (hz) to ERB-rate scale.

$ERB = 21.4 * log_{10}(4.37e - 3 * hz + 1)$

2.2.2 Perform segmentation on ERB scale.

2.2.3 Convert ERB-rate scale to its frequency scale.

$hz = (10\text{^}(erb/21.4) - 1)/4.37e - 3$

2.2.4 Calculate the rate of decay (rd).

$rd = 1.019 * 27.7 * (4.37 * \frac{central\ frequency}{1000} + 1)$

2.3 Compute the gammatone filter bank to filter the input signal and to produce GF features.

Step 3: Apply DCT on GF features to extract GFCC features.

5.2.2 Mel Frequency Cepstral Coefficient

The process of MFCC is accompanied by pre-processing step which helps in removing

DC along with its first order transformation using pre-emphasis. Further it helps in short term Fourier transformation analysis using Hamming window or by replacing it with different multi taper spectrum estimation function. The auditory spectral investigation is then conducted with Mel filter bank using a triangular filter (Davis, S. B., & Mermelstein, 1980). The result of filter bank is the weighted sum of its spectral magnitude after executing log on it. It calculates the sound frequency (f) that helps in generation of the mel frequency $Mel(f)$ using Eq. (3.13). The frequency here is overrun with a triangular filter that produces highly associated dataset. The quantity of filter is dependent on its sampling rate (such as 8 kHz, or 16 kHz) that originate the demand of varying number of filters such as it lies in the range of 15 to 20 filters. DCT is used to segregate the extracted information. Conclusively it results into the reduced dimension static feature vector. In actual practice, 13 static feature vectors are always kept to exhibit the important characteristics of the spoken utterances. Their second and third order derivatives are examined which carry various amount of features vector using Algorithm 5.2 (Davis, S. B., & Mermelstein, 1980).

Algorithm 5.2: Conventional MFCC method

Step 1: Initialize the lowest_frequency, linear_filters, linear_spacing, log_filters, log_spacing, fft_size (N) and cepstral coefficient.

Step 2: Framing: Segment the speech samples into a small frame with length of $20-40$ msec. The adjacent frames are being separated by an overlapping factor M and frame size $N = 256$.

Step 3: Compute the hamming window.

h(t) = 0.54 - 0.46*cos (2*pi*(0:window size-1)/window Size)

Step 4: Evaluate the fast fourier transform.

Y (w) = FFT [h(t)* X(t)] = H(w)* X(w)

Step 5: Perform mel filter bank processing.

Step 6: Calculate discrete cosine transform.

DCT = 1/sqrt(total Filters / 2)* cos ((0:(cepstralCoefficients-1))' * (2*(0:(totalFilters-1))+1) * pi/2/total Filters)

Step 7: Generate delta energy and delta spectrum.

Energy =$\Sigma X^2[t]$

13 delta features represent the change between frames in the corresponding cepstral or energy feature, while each of the 39 double delta features represents the change between frames in the corresponding delta features.

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Above methods shown in Algorithms 5.1 and 5.2 perform two level of feature

dimension reduction: by utilizing a filter bank and by transforming the information into a cepstral domain from its log spectral domain.

In this study, the mel scale and gammatone filters are used for feature extraction after fusion of PCA based approach. The process is implemented after integration of multiple window estimation methods enhancing the performance of proposed hybrid system.

5.2.3 Multi taper spectrum estimation

Reducing the spectral leakage helps in improving the variance by multiple windowing process. The task is accomplished using a power spectrum estimation approach. The issue of basing is effectively cured by selecting an optimal value of a taper. The value of m^{th} frame along with k^{th} frequency bin of its windowed periodogram is estimated through (Alam et al. 2013) Eq. (5.1).

$$S(m,k) = |\sum_{j=0}^{N-1} \mathrm{w(j)s(m,j)e^{-i2\pi jk/N}}\ |^2 \tag{5.1}$$

The equation belongs to a single taper or a modified Hamming window periodogram. Due to deviation between the calculated values of the spectrum corresponding to the actual value of required spectrum, it results into the decrease of taper value. Besides, it doesn't decrease the value of its variance. The high value of variance has drastic affect causing leakage through its spectral estimate (Kay, 1988). To reduce the variance, a simple and appropriate multi taper method has been to reduce variance. Numerous multi taper methods have been employed in the past (Hansson-Sandsten, M., & Sandberg, 2009; Hansson, M., & Salomonsson, 1997; Thomson, 1982) that shows good leakage characteristics for spectrum estimation. One of them is a sine taper which is pair wise orthogonal and one of the easiest method, calculated using (Alam et al. 2013) Eq. (5.2).

$$w_p(j) = \sqrt{\frac{2}{N+1}}\, sin\{\pi p(j+1)|N+1\}, j = 0,1,\dots\dots, N-1 \tag{5.2}$$

It constitutes a multiplicative constant orthonormal that retains a unit norm and an orthogonal shape. The optimal weighting of cepstrum estimation is performed on a sine taper resulting into a creation of SWCE and multipeak of a peaked spectra using step-

wise process of Algorithm 5.3.

Algorithm 5.3: Multitaper method

Step 1: Initialize the frame shift, frame_size, window length (N), number of frames (num_frames) and number of taper (t).

Step 2: Arrange the speech features into the frames of dimension num_frames of x frame size.

Step 3: Choose any one method.

3.1 **Thomson method**

Compute the resolution of the spectrum

$$r\,(\text{Tapers}) = 2 * \sin\left(\frac{\pi*(t+2)*L}{N}\right)/(2 * \pi * L) \text{ Where } L=1, 2, 3, \ldots\ldots\ldots, N-1;$$

$$s\,(\text{weights}) = \text{ones}(t, 1)/t;$$

3.2 **SWCE method**

$$\text{Tapers} = \sqrt{\frac{2}{N+1}}\sin\left(\frac{\pi*t*(j+1)}{N+1}\right), \text{ where } j=1, 2, 3, \ldots\ldots\ldots,N-1. \text{ weights} = \frac{\cos\left(\frac{2*\pi*(t-1)}{\frac{M}{2}}\right)+1}{\sum_{t=1}^{M}\cos\left(\frac{2*\pi*(t-1)}{\frac{M}{2}}\right)+1},$$

where M= fix (N/t).

3.3 **Multipeak method**

The multipeak tapers are obtained as the solution of eigenvalue problem:

$$R_{B'}w_j = v_j R_Z w_j,$$

where j=1,2,3,…..N., $R_{B'}$ is (N*N) toeplitz covariance matrix of the spectrum model defined by:

$$S(f) = \begin{cases} \exp^{\frac{-2c|f|}{10\log_{10}(e)}} & |f| \le B'/2 \\ 0, & |f| > B'/2 \end{cases}$$

with C = 20 dB and a predetemined interval of width B' outside of which spectral leakage is to be prevented, R_Z is the toeplitz covariance matrix, chosen for decreasing the leakage from the sidelobes of the tapers, of the following frequency penalty function:

$$S(f) = \begin{cases} G, & |f| \le B'/2 \\ 0, & |f| > B'/2 \end{cases}, \text{ where G=30 dB}$$

A weight of the multipeak method is calculated for the highest eigenvalues M using:

$$\lambda_t = \frac{v_t}{\sum_{t=1}^{M} v_t}$$

Step 4: Evaluate the power spectra that estimates each frame.

```
for (i = 1:size(tapers,2))
spec = spec + weights(i)  * abs(fft((frames').*repmat(tapers(:,i),1,size(frames',2)), N)).^2;
end
```

where tapers is a matrix of size (frame_size x t) containing the tapers (window functions) as column vectors and weights are the vector of length 't' of taper weights obtained from step 3.

This chapter integrates three taper methods through Algorithm 5.3 in conventional GFCC and MFCC techniques resulting into better feature vectors. It embeds PCA based filter bank optimization through selection of an optimal value of its taper parameter.

5.2.4 Selection of the tapers

The taper variable in Hamming window plays a crucial role. The value of taper should be fixed and then decreases towards signal frame boundaries. In this study, three taper approaches are used as per requirement – a) white noise employed Thomson method, b) voiced speech adopts multipeak approach and c) cepstrum analysis performed using sine taper. Distinct types of tapers are designed for reduction of the variance of an uncorrelated sub spectrum in the estimation error using Eq. (5.3).

$$S(f) = \sum_{j=1}^{K} \lambda(\mathrm{j}) \left| \sum_{\mathrm{t}=0}^{\mathrm{N}-1} \mathrm{w_j(t)x(t)c}^{-\mathrm{i}2\pi\mathrm{tf/N}} \right|^2 \quad (5.3)$$

The selected three multitaper methods produce a smooth spectrum resulting into reduced variance in comparison to that produced by Hamming window. In the proposed work, the range of taper (t) is varied. For a small number of $t < 4$, it will save the harmonic and spectral envelope. For $t > 8$ the harmonic factors get smeared out. Thus, the selection of the optimal number of taper plays a crucial role in speech recognition process.

5.3 Principal component analysis based filter bank optimization

PCA was applied to reduce the dimensionality of the interrelated variable of a dataset. It results into an efficient presentation of the feature vectors. Consider x random variable $\{x_k(n), n = 1,2,3, \dots, M\}$ of length M, representing signal components within k^{th} filter bank. These components are controlled by k^{th} filter of the filter bank. It uses earlier work (Lee et al. 2001) to obtain the k^{th} filter bank coefficients. The k^{th} filter vector is the component of the column vector w_k calculated by (Hung, 2004) Eq. (5.4).

$$w_k = [w_k(1), w_k(2), \dots \dots . w_k(n)_k]^T \quad (5.4)$$

Filter bank weights the signal components within a frequency band. The weight of these signal components can be calculated to generate a single value with a high variation in Algorithm 5.4.

Algorithm 5.4: PCA Optimization Method

Step1: Input a given set of features:
$X = \{x_1, x_2, x_3 \ldots\ldots\ldots, x_n\}$.
Step 2: Subtract the mean from each of the data dimensions.
$X = x_i - \bar{x}$, where $\bar{x}$ is the mean for the given data and i=1,2,3,……n.
Step 3: Evaluate the covariance matrix (cov).
Step 4: Calculate the unit eigenvectors and eigenvalues of the covariance matrix. $|cov - \lambda I| = 0$ where λ is the Eigen value and I is the identity matrix.
Step 5: Selection of components: Arrange eigenvectors by eigenvalues - This gives the components in order of significance.
Step 6: Formation of feature vector (matrix of vectors): Select the eigenvectors and frame a matrix with a eigenvector in the columns.
$Features = (eign_1, eign_2, eign_3 \ldots\ldots\ldots\ldots eign_n)$
Here $eign_1, eign_2, eign_3 \ldots\ldots\ldots\ldots eign_n$ are the eigenvectors for the corresponding eigenvalues $\lambda_1, \lambda_2, \lambda_3, \ldots\ldots\ldots, \lambda_n$.
Step 7: Final feature vectors = $(Features)^T$ * (mean adjusted features).

5.4 Proposed integrated Punjabi-ASR system with PCA optimized filter on multitaper framework

It is very important to extract discriminate and efficient parameters of an acoustic signal for training and testing of the Punjabi-ASR system. Input speech utterance both in training phase and testing phase are processed in front end of the system. Two parallel techniques - MFCC and GFCC are used for feature extraction. The integration of PCA and multitaper methods in baseline MFCC and GFCC technique helps in extraction of relevant feature vector before they are further classified in modeling phase. These feature vectors are used to discriminate their classes and must be insensitive to the test speech sample recognition. In this study, proposed feature vectors are processed through:

- Reduction in variance of an input speech signal and making a trade-off between their variance along with bias. It is possible through implementation of single taper (Hamming window) or multiple taper methods (such as Thomson, multi-peak and SWCE).
- Optimization of each filter bank (Gammatone and triangular filters) helps in data driven PCA approach to produce approximate solution.

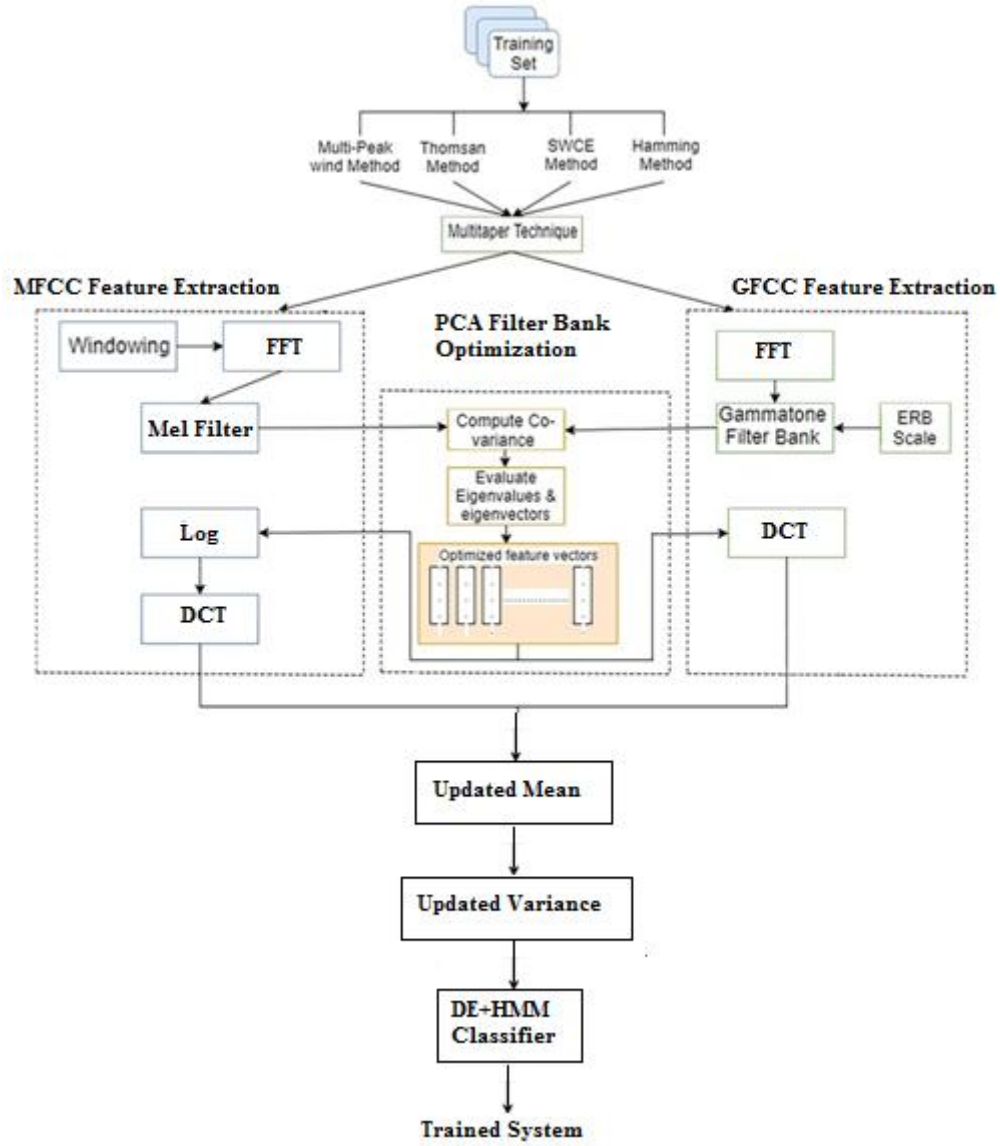

Figure 5.1: Block diagram of Parallel multitaper, PCA integrated GFCC and MFCC feature extraction approaches with DE+HMM employed acoustic modeling classifier

Figure 5.1 shows the integrated block diagram of the flow. To process the input speech signal, we initialize the input parameters (such as frame_shift, frame_size, window_length). These parameters help in arrangement of speech vectors $x_i = \{x_1, x_2, x_3, \ldots\ldots\ldots, x_N\}$ into a sequence of frames of matrix with a dimension (no of frame * frame_size) of N samples. According to earlier studies, conventional feature extraction approaches (like GFCC, MFCC and wavelets) calculate their power spectrum

with the help of windowed peridogram in Eq. (5.5).

$$\hat{P} = |\sum_{t=0}^{N-1} w(t)\, x(t) e^{-\frac{2j\pi f}{L}}|^2 \quad (5.5)$$

where frequency (f) lies in the range of $(0, L-1)$, $j = \sqrt{-1}$ is an imaginary unit and is used as window in speech processing application. A Hamming window is calculated using Eq. (3.9). From statistical perspective, the Hamming window is also known as a single taper approach which helps in reduction of biasing of an input speech spectrum. It is calculated by investigating the difference between the required estimates to its actual spectrum values. The output results into a higher value of its variance. To handle the problem oflarge variance, various types of taper approaches are used. A complex feature extraction model increases the complexity of its training phase but obtains better performance in testing phase of the system. These complex feature extraction approaches are employed in training and testing phase using multitaper spectrum estimation using Eq. (5.6).

$$\hat{S} = \sum_{i=1}^{K} \lambda(i) \, |\sum_{i=0}^{N-1} w_i\ (t) x(t) e^{-\frac{j\,2\pi}{N}}|^2 \quad (5.6)$$

where k denotes the number of tapers, w_i depicts the m^{th} data taper and λ(i) indicates the weight of i^{th} taper. For a single input speech frame, three different improved multitaper approaches in baseline MFCC and GFCC techniques e.g. Thomson, multipeak and SWCE are merged. All the three multitaper methods help in smoothening of its spectrum in comparison to a single taper approach (hamming window). The process is evaluated on isolated words speech corpora. To reduce the error in recognition phase, each speech utterance need to be discriminated and refined in training and testing phases. Unlike baseline feature and acoustic model based system which fails to combine information from a large speech corpus, the complex models with conceived value of taper are capable of doing so. Initially, taper value are randomly generated with values $2, 4, 6, 8, 10,$ and 12. To accomplish maximum word accuracy, a

steady value of taper is used such as 8 in case of multipeak and SWCE or 6 in case of Thomson approach. Furthermore, to escalate the feature vector, filter bank optimization is performed. It is implemented with the help of PCA approach because a default value of filter bank is not optimal in every situation for generation of required output. Then to fix the value of filter bank, the random variation takes place to obtain a static value for generation of better feature information. PCA optimized filter bank helps in retrieving a large value of dataset for filter bank processing and helps in calculating mean parameters. The relationship between each vector is determined after simulating its covariance for a given filter bank data. The appropriate selection of a component is examined through positioning of its eigen values and eigenvectors in steps 6 and 7 of algorithm 5.4. For every filter bank, its output is modeled on a random set of values of its filter bank ($17, 20, 24, 29,$ and 32). Too large or too small value for the number of filters doesn't provide any satisfactory output. The system obtained maximum word accuracy after achieving an optimized value of its filter bank that is 32. A better feature estimation of vector is obtained after feature extraction approach is required in modeling classifier. After processing of feature vector and before classification stage, the mean and variance are evaluated using Eqs. (3.15 and 3.16). The output of model parameters is refined to maximize the probability of an observation sequence. The question is how to generate an optimal value of model parameters? Sometimes the required size of the model parameter may go beyond its target value. The output of mean and variance parameters is refined with differential evolution approach in Algorithm 3.2 of Section 3.1.3. Population size is randomly changed ($10, 20, 25, 30, 40, 45, 50,$ and 60) for a trial and it is found to obtain success for a population size of 25. Numerous other parameter for DE based model parameters are varied like NP (the size of its population), CR (shows a crossover factor) and F (a scaling factor). DE is applied on mean and variance after splitting them to obtain a fixed value that result into a refined acoustic feature vectors. The key procedure to call and implement proposed technique is represented in

Algorithm 5.5. These model parameters will help in efficient training of a system using its statistical approach of HMM. The final output of training phase results into a compact and relevant feature vectors. The generation of a discriminated observation data helps in achievement of a highest likelihood probability which is further required to be presented as per surroundings. Therefore, it stores the value of its Real Time Factor (RTF) that increases with a same size and calculates its complexity in training stage of parallel integrated system. During decoding stage the best matched feature vector is produced using Viterbi Algorithm.

Algorithm 5.5: Proposed parallel hybrid technique

Step 1: Initialize the speech processing parameters i.e. speech segments $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$.

Step 2: Arrange the signal into its frames for a dimension (num_frames) that is divided into x frame size.

Step 3: Process the signal using three multi-taper techniques with any of the mentioned methods in step 3 of algorithm 5.3.

Step 4: Process the signal using any one feature extraction through (ch): 1 for GFCC method and 2 for MFCC method.

switch(ch)

case 1: GFCC feature extraction method:

- Follow step 1 to 2.2.4 of step 2 in algorithm 5.1.
- Call PCA filter bank optimization method of algorithm 5.4.
- Return to step 2.2.4 of step 2 of algorithm1

case 2: MFCC feature extraction method:

- Follow step 3 to step 5 of algorithm 5.2.
- Call PCA filter bank optimization method of algorithm 5.4.
- Return to step 2.2.4 of step 2 of algorithm 5.1.

Step 5: Normalize the extracted Feature vector using:

$$B = \sum \frac{(f-\mu)^2}{\sigma}$$

Step 6: Compute the mean and variance using Eqs. (3.15) and (3.16).

Step 7: Generate the observation sequence to maximize it with differential evolution process through:

Likelihood Probability $(likiProb) = \frac{1}{t}(\sum_{i=1}^{t} log(P(O_i|\lambda))$

5.5 Corpus setup

To evaluate the performance of the system, a total of 17,500 utterances are evaluated using Punjabi database. This dataset contain utterances from four dialect speakers. In this study, speech features are collected in two different environments (clean and real). The dataset is generated by 50 (men and women) speakers. Speech utterances are initially collected in four sets: SET1 is collected from Malwa dialect speakers, SET2 is

collected from Majha dialect speakers, SET3 is collected from Doaba dialect speakers and SET4 is collected from Powadh dialect speakers. These sets are further merged together to frame two training sets (train set1 and train set2) and two test sets (test set A and test set B). The train set1 and test set A contains the utterances of speakers in clean environment. On other hand, the train set2 and test set B is composed of utterances of real environment (multi-conditions). Real environment dataset is a collection of three different type of noise corpus (such as car noise, cafeteria noise and street noise). The experiments are conducted using 10 fold cross validation viz 100 out of 10 spoken utterances are kept in database for test sets and rest are used in training corpus.

5.6 Experimental analysis

The experiment is done using three different feature extraction approaches: First case involves conventional 39-dimensional MFCC and GFCC feature vectors, evaluated on 13 static features, their Ist order delta coefficients and double delta coefficients. Second case involves improved MFCC and GFCC feature vectors after integration of multitapered methods and in third case PCA based filter optimization is performed in improved GFCC and MFCC techniques. These approaches are used in four systems individually on designed dataset:

S1: This system employs varying number of tapers on improved multitapered (GFCC and MFCC) based feature extraction technique. Initially the baseline GFCC and MFCC approaches are tested which provide 78.23% and 81.43% results with single taper i.e. Hamming window. To achieve the best performance, varying value of taper (t) (such as 2, 4, 6, 8, 10, 12) is used to obtain a low bias and variance values on selection of an optimal value of its taper. The taper values are varied in all the three multitapered approaches - Thomson, multipeak and SWCE in baseline GFCC and MFCC techniques. S1 experiments are implemented on train set2 as listed in Table 5.1(a) and Table 5.1(b). System attains maximum word accuracy at a (t) value of 8 in all cases of improved

GFCC and MFCC but in GFCC-Thomson it attains better accuracy after tuning taper to a value of 6. It is noteworthy to specify that GFCC earns better word accuracy using SWCE in multi-environment dataset in comparison to baseline, Thomson or multipeak integrated MFCC and GFCC techniques. An improvement of 1% to 3% is obtained through improved GFCC approach in comparison to MFCC as a feature extraction technique in multi-condition corpora. Varying WA with different range of tapers recommend that Thomson procures better output at a small value of (t) than the SWCE and multipeak using improved GFCC technique.

Table 5.1(a): Performance evaluation with various multitaper methods using MFCC on varying taper length

Type of Feature Set	**No of Tapers (Word Accuracy %)**					
	2	4	6	8	10	12
MFCC-Thomson	68.72	52.94	74.85	75.48	67.34	58.88
MFCC-Multipeak	64.70	58.79	58.79	70.58	70.20	52.46
MFCC-SWCE	76.40	52.94	76.47	78.82	58.82	68.82

Table 5.2(b): Performance evaluation of different multitaper methods using GFCC on varying taper length

Type of Feature Set	**No of Tapers (Word Accuracy %)**					
	2	4	6	8	10	12
GFCC-Thomson	70.48	57.45	78.35	76.18	67.34	56.47
GFCC – Multipeak	66.75	55.85	68.74	71.28	69.23	52.46
GFCC – SWCE	75.72	54.25	76.47	79.82	58.82	68.8

S2: After tuning of a better value of taper in three different tapered approaches, the efficiency of the S2 system is tested at different SNR (in db) levels varying from 5, 10, 15, 20 and 30 (clean). Here, the output of the system with different values of SNRs is taken into considerations. The conventional MFCC approach attains better output in clean environment only without integration of multiple tapered methods in it. The system performance is affected in multi-condition environment but baseline or improved GFCC produces better results. The performance of system (S2) with GFCC

and MFCC (baseline i.e. using single taper and multi-taper methods) techniques are as listed in Table 5.2. SWCE integrated system in baseline GFCC achieved better accuracy. The robustness of the system is tested and it is found that baseline GFCC technique gives better result than improved MFCC and GFCC techniques.

Table 5.2: Performance evaluations with different SNR level on baseline and multitaper estimation methods in different feature sets

Type of Feature Set	**Varying SNR level (Word Accuracy %)**					
	5dB	10dB	15dB	20dB	25dB	Clean
MFCC	54.18	64.65	70.15	71.48	73.82	75.63
GFCC	59.25	70.66	74.52	75.19	75.42	77.16
MFCC-Thomson	55.71	57.88	60.24	63.58	66.47	70.36
MFCC-Multipeak	52.42	54.74	57.58	61.33	67.59	62.78
MFCC-SWCE	53.61	57.49	60.78	64.34	67.78	73.67
GFCC-Thomson	59.83	61.97	66.54	69.93	74.84	74.88
GFCC-Multipeak	56.55	64.71	66.0	65.81	70.66	72.67
GFCC-SWCE	64.92	67.62	71.99	75.92	72.45	78.18

S3: For robust speech recognition, PCA is used to optimize the shape of the filters in the filter bank such as Mel filter bank in MFCC and Gammatone filter bank in GFCC. The PCA based filter bank is applied in two ways such as on baseline MFCC and GFCC, multitaper estimation method integrated GFCC and MFCC.

Table 5.3: Recognition rate using PCA optimized GFCC and MFCC filter with different multitaper methods

Type of Feature Sets/ Training Sets	**Train Set 2**							
	SET2		**SET1**		**SET3**		**SET4**	
Test set	**Clean**	**Real-Time Data**	**Clean**	**Real-Time Data**	**Clean**	**Real-Time Data**	**Clean**	**Real-Time Data**
MFCC	65.91	62.55	70.15	66.64	53.82	48.12	45.84	40.68
GFCC	70.57	64.32	74.47	71.46	57.42	54.92	50.23	43.52
Optimized MFCC	69.98	65.83	73.15	69.54	55.16	53.11	52.42	48.79

Optimized GFCC	74.85	71.46	79.06	76.18	59.26	56.19	56.82	54.74
MFCC-Thomson	66.27	63.48	70.46	66.57	52.55	49.20	44.47	42.27
Optimized MFCC-Thomson	68.31	64.92	72.55	68.34	55.03	51.86	47.12	44.62
GFCC Thomson	69.93	66.00	74.88	74.54	58.55	54.78	50.95	52.63
Optimized GFCC-Thomson	71.27	67.54	76.17	71.98	61.12	57.47	53.69	54.80
MFCC-Multi-peak	61.33	57.58	67.49	62.71	55.66	51.11	50.63	49.63
Optimized MFCC-Multi-peak	63.52	59.17	69.64	65.25	58.81	54.67	54.19	53.72
GFCC Multi-peak	65.77	63.51	70.85	68.02	50.63	48.86	41.66	49.64
Optimized GFCC-Multi-peak	69.04	65.41	72.671	70.56	54.34	52.48	45.22	52.32
MFCC-SWCE	64.47	64.54	73.67	67.78	60.18	54.52	47.82	4922
Optimized MFCC-SWCE	71.485	66.64	75.63	70.35	62.59	57.49	50.65	52.16
GFCC-SWCE	73.41	69.95	77.84	72.69	64.92	59.87	57.62	54.31
Optimized GFCC-SWCE	75.92	72.45	80.18	75.37	67.87	62.47	60.17	57.28

The PCA based new drive features with GFCC-SWCE multitaper method yields performance improvement on other combination of feature extraction techniques such as optimized, improved or baseline GFCC and MFCC in mismatch environment. The experiments are performed on Train Set2 and with different speaker sets such as SET1, SET2, SET3 and SET4. Maximum accuracy is achieved with SET1 speakers followed by SET2 speakers then by SET3 and SET4 speakers with optimized GFCC-SWCE technique in feature extraction phase as shown in Table 5.3.

S4: It analyzes the performance of the integrated system (S4) with PCA based data driven approach; improved, baseline GFCC and improved, baseline MFCC approaches as given in Table 5.4. The word accuracy is calculated on a different train and test sets. The system is tested individually on two different trained datasets (Train Set1 and Train

Set2). System testing is performed using two type of test sets such as Test SetA and Test Set B. The SetB is further framed from three different type of multi-condition corpora. The improved GFCC-SWCE and multipeak methods perform better on medium vocabulary corpora during development test. As the size of corpora increases in training phase it result into issue of handling number of large feature vectors. The issue of handling large corpora feature is refined by employing PCA in log-spectral domain after optimizing the shape of its filter bank.

Table 5.4: Recognition performance obtained through single or multitaper methods based on GFCC and MFCC features with integration of PCA based filter bank

Type of Feature Set	**Training Condition**							
	Clean Training Condition				**Multi-condition Training Corpus**			
	Test setB (Word Accuracy)			**Test setA (Word Accuracy)**	**Test setB (Word Accuracy)**			**Test setA (Word Accuracy)**
	Car noise	**Street noise**	**Cafeteria noise**	**Clean**	**Car noise**	**Street noise**	**Cafeteria noise**	**Clean**
MFCC	66.66	45.54	57.52	68.44	62.48	30.68	40.32	64.74
GFCC	70.66	50.22	54.42	70.86	67.52	33.52	54.92	73.57
Optimized MFCC	69.26	52.42	55.16	69.54	65.85	38.69	52.33	69.98
Optimized GFCC	66.04	56.52	59.26	76.08	73.76	53.63	56.39	74.85
MFCC-Thomson	66.27	44.46	52.55	66.57	65.78	32.26	49.20	66.47
Optimized MFCC-Thomson	64.36	46.02	55.01	68.64	67.92	33.62	53.06	68.60
GFCC Thomson	76.26	50.95	55.55	74.54	66.00	52.63	54.50	69.96
Optimized GFCC-Thomson	70.94	52.69	61.12	70.98	67.57	53.80	55.45	70.47
MFCC-Multi-peak	62.70	50.62	55.66	64.70	57.58	39.63	53.33	60.66
Optimized MFCC-Multipeak	62.22	54.09	55.51	65.45	59.37	53.62	54.65	66.54
GFCC Multi-peak	64.02	40.66	50.61	68.34	65.53	39.63	40.06	65.77
Optimized GFCC-Multipeak	70.26	45.22	54.14	73.56	65.73	52.32	52.40	69.34
MFCC-SWCE	67.74	46.52	60.15	67.78	67.57	39.22	54.52	64.47

Optimized MFCC-SWCE	70.32	50.65	62.59	73.65	66.67	52.06	55.49	70.48
GFCC-SWCE	72.69	56.62	64.92	74.69	69.95	53.30	59.05	76.40
Optimized GFCC-SWCE	72.37	60.06	68.54	75.67	72.75	56.28	62.45	75.94

5.7 Summary

In this chapter, we present the hybrid feature extraction approaches on different sets of Punjabi mismatched train and test speech corpora. In proposed system, two techniques are hybridized in a complex model. Initially, the baseline GFCC and MFCC approaches are analyzed on DE+HMM based acoustic classifier. These feature extraction approaches are further hybridized after integration of multitapered methods and PCA technique on their respective filter banks after tuning of taper parameter. As a result, the proposed hybrid system in S1, S2, S3 and S4 shows a better performance than baseline MFCC, GFCC on HMM based ASR systems.

In the next chapter, we study the advantages of more feature reduction approaches at feature extraction level along with implementation of speaker adaptive model. We further employ hybrid acoustic modeling approaches such as DNN+HMM at modeling phase to improve its performance.