Spectro-Temporal Features Based Automatic Speech Recognition

by

Ankit Nagpal

201311022

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

INFORMATION AND COMMUNICATION TECHNOLOGY

to

Dhirubhai Ambani Institute of Information and Communication Technology

June, 2015

Declaration

I hereby declare that

(i) The thesis comprises of my original work towards the degree of Master of Technology in Information and Communication Technology at DA-IICT and has not been submitted elsewhere for a degree,

(ii) Due acknowledgement has been made in the text to all the reference material used.

Signature of Student

Certificate

This is to certify that the thesis work entitled "**Spectro-Temporal Features Based Automatic Speech Recognition**" has been carried out by Ankit Nagpal (201311022) for the degree of Master of Technology in Information and Communication Technology at Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT) under my supervision.

Prof. Hemant A. Patil

Thesis Supervisor

Acknowledgements

I express my deepest gratitude to Dhirubhai Ambani Institute of Information and Communication Technology for providing facilities and environment conducive for carrying out the research work.

I would like to express my sincere thanks to Dr. Hemant A. Patil for his constant support and guidance during my research period. I greatly acknowledge him the way he motivated me to carry out independent research.

I am extremely thankful to my subject experts, Dr. Suman Mitra and Dr. Pratik Shah for their valuable advices and suggestions.

I express my gratitude to Prof. Manjunath V Joshi for sparing some time from his schedule to review my thesis.

I express my profound thanks to all the members of speech lab, especially Mr. Mohammadi Zaki, Mr. Hardik Sailor, Ms. Tanvina Patel, Mr. Maulik Madhavi and Mr. Nirmesh Shah for sharing their expertise on various subjects and helping me get on with the tools and technologies used to carry out the research work.

I would like to thank Department of Electronics and Information Technology (DeitY), New Delhi, India, for its financial support, through the project, *viz.*, "Speech Based Access of Agricultural Commodity Prices and Weather Information in *12* Indian Languages", to carry out the research work.

Finally, I express my earnest thanks to my family for its moral support and patience.

Table of Contents

Abstract

ASR technology has found its application in almost every field in life. Today's world cannot be considered as noise-free and deploying ASR technology in such environments would incorporate the challenge to deal with various kinds of noises and channel effects. Thus, robustness of ASR is becoming increasingly important. State-of-the-art Mel Frequency Cepstral Coefficients (MFCC) features capture spectral information and some temporal dynamics in the speech signal. Spectro-temporal features, on the other hand, are more physiologically motivated, as they capture more perceptual information, and are able to perform better in the presence of noise. In this thesis, cepstral analysis, theory of cepstral coefficients (MFCC and Gammatone Frequency Cepstral Coefficients, i.e., GFCC) and motivation to use spectro-temporal features, are discussed. Furthermore, the work presents the theory behind Gabor filters and motivation to incorporate them for ASR task. Algorithm for the extraction of spectro-temporal features- Spectro-Temporal Gabor filterbank features (GBFB), is also presented in detail.

Experiments are carried out on TIMIT database, with various additive noises such as white, babble, volvo and high frequency (under various SNR levels) to compare spectro-temporal features, denoted by $GBFB_{mel}$+MFCC and the proposed $GBFB_{Gamm}$+GFCC (incorporating mel and Gammatone filters, respectively) and the state-of-the-art MFCC features. Experiments are carried out with HTK as back end, taking into account the effectiveness of acoustic and language model. It is concluded that with acoustic modeling only, spectro-temporal Gabor filterbank (GBFB) features (whether incorporating Gammatone filterbank or mel filterbank) when concatenated with cepstral coefficients perform better than the state-of-the-art MFCC features in clean conditions as well as in the presence of various additive noises or signal degradation conditions. This is because GBFB features are able to capture more local joint spectro-temporal information, than the MFCC features, from the speech signal.

List of Symbols

ω_n	Temporal modulation frequency
ω_k	Spectral modulation frequency
υ_n	Semi-cycles, in temporal-domain, under the envelope of Gabor filter
υ_k	Semi-cycles, in spectral-domain, under the envelope of Gabor filter
Q	Quality factor

Other minor symbols are defined at first occurrence; where necessary some symbols are redefined in the text.

List of Acronyms

ASR	Automatic Speech Recognition
HMM	Hidden Markov Model
LM	Language Model
TRAPs	TempoRAl Patterns
MFCC	Mel Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
STRF	Spectro-Temporal Receptive Field
WER	Word Error Rate
GFCC	Gammatone Frequency Cepstral Coefficients

HTK	Hidden Markov Model Toolkit
BM	Basilar Membrane
FFT	Fast Fourier Transform
STFT	Short-Time Fourier Transform
DCT	Discrete Cosine Transform
ERB	Equivalent Rectangular Bandwidth
GBFB	Gabor Filterbank
RASTA	Relative Spectra
GBFB_{mel}	GBFB features incorporating mel filterbank
$\text{GBFB}_{\text{Gamm}}$	GBFB features incorporating Gammatone filterbank
PRA	Phoneme Recognition Accuracy

List of Figures

Chapter 1

Introduction

1.1 Overview of ASR

Speech is the most basic and the simplest way of communicating your ideas and thoughts. Every human being cannot read and write, however, can communicate using the speech signal. Due to its high influence on the day-to-day life of human beings, speech processing has been one of the subjects of research from decades. Recognizing what has been spoken so as to act accordingly is known as speech recognition. Technological advancements have been able to make machines perform humans' tasks. Speech recognition performed by the machine is known as *Automatic Speech Recognition* (ASR), wherein speech is given as input to the machine and machine converts different speech sound units, e.g., phonemes, syllables and words, to text. Thus, ASR is also termed as "speech-to-text" conversion process.

1.1.1 Architecture of ASR system

Figure 1.1 shows the architecture of ASR system. ASR involves two stages; first stage being extraction of relevant features from the speech signal and the second being classification of the features to generate output text. Feature extraction should be such that they capture relevant information (like temporal information, spectral information or *joint* spectro-temporal information) from the speech signal, and are not redundant. Feature vector is an acoustical representation of the speech signal and is denoted by $\boldsymbol{o}$. Hidden Markov Models (HMMs) are most commonly used back-end feature classification systems for speech recognition task. HMMs are generative models and represent acoustic models for all the speech sound units. Acoustic models are statistical representation of sound units obtained from the speech signal and corresponding transcription. Language models (LMs) provide the

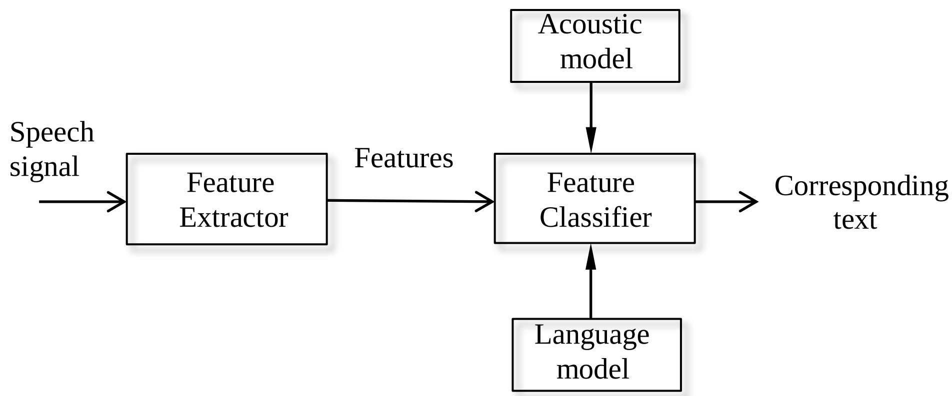


Figure 1.1: Basic architecture of an ASR system.

probability of occurrence of sequence of words in the language and are generated by considering word sequences in the training data. Speech is recognized with the help of acoustic models and language models. Let observation sequence (feature vector sequence) be represented by $\boldsymbol{O}$.

$$O = o_1 o_2 \dots \dots \dots \dots .. o_{T,} \tag{1.1}$$

where o_t is the observation at time t. The word recognition problem can then be regarded as computing

$$arg \max_i \{P(w_i|O)\}, \tag{1.2}$$

where w_i is the i^{th} word in the *vocabulary*. The above probability cannot be computed directly and is computed by using Bayes' rule which is [1]

$$P(w_i|O) = \frac{P(O|w_i)\, P(w_i)}{P(O)}, \tag{1.3}$$

where $P(O)$ remains same for all utterances and hence, given the probability of occurrence of word ($P(w_i)$), $P(w_i|O)$ depends on the probability of the observation given word w_i, ($P(O|w_i)$). Thus, the optimal word, w_{opt} for a given observation sequence is computed as

$$w_{opt} = arg \max_W \{P(O|W) P(W)\}, \tag{1.4}$$

where $P(O|w)$ represents the *acoustic* probability of the word and $P(w)$ represents the *language* probability.

1.1.2 Application of ASR

ASR incorporates machine in-place of humans. Machines recognize the speech signal (the command given to the machine) and then perform the task, which would otherwise be performed by the humans, accordingly. This reduces manpower and hence, finds application in various domains. Some of the domains, where ASR finds its use are [2]:

a) Healthcare

In healthcare area, speech recognition is utilized by the dictator where he/she dictates into the ASR system and gets out the corresponding text document. The dictator can then edit and sign off the document.

b) Military

In military aircrafts, speech recognition is used for setting radio frequencies, commanding autopilot system, controlling air traffic, setting weapon release parameters, etc.

c) In-car system

Speech recognition is able to eliminate manual control of the vehicle. Person, while relaxing, can control the vehicle navigation by his/her voice. Simple voice commands may be used to initiate phone calls, select radio stations or play music.

d) Multimedia

In mobile phones, instead of typing long messages, one can speak and the corresponding text will then be delivered to the recipient. Computers can be operated via speech signal, rather than the use of input devices. ASR technology is also being utilized in search engines, where speech is used to search over the Internet. Google search by voice is one such technology [3].

Other applications include gaming, Interactive Voice Response System (IVRS), robotics, transcription, automatic subtitling and many more.

1.1.3 ASR in real-life scenarios

ASR technology is now being deployed to the real-life scenarios where the environment is no more clean in terms of channel noises. Environment contains different noises and channel effects such as reverberations. Modern world, involving huge traffic and population, cannot be considered as noise free. If not reverberation, background noise would still be there. Thus, deploying ASR in those environments would have to tackle with various kinds of noises and different channel effects. Hence, *robustness* of ASR is becoming increasingly important. In the presence of channel noise (especially with highly noisy conditions), statistics of speech signal, like joint spectro-temporal intensity, change significantly. Thus, recognizing speech in the presence of noise is indeed a challenging task. Decades of research in the field of ASR brought numerous methods to improve the recognition performance by increasing robustness against variability of speech signals. History and theories of some of these methods are described in chapter 2.

1.2 Motivation and Objectives of the Thesis

ASR technology is being utilized in almost every field in life and thus, the efficacy of ASR has become a key factor. Real-life scenarios involve noisy channels and it is indeed a challenging task for ASR system to recognize speech signal in those signal degradation (or noisy) conditions, since characteristics of the speech signal would change significantly in the noisy environments. ASR systems do not perform robust enough in the presence of background noise, while there is a strong need to have accurately performing ASR systems in real-life scenarios that involve different background noises (such as babble, vehicle, music, etc). This work aims at improving the efficiency (the recognition performance) of ASR system in noisy environments. Several methods till date have been proposed to increase the robustness of ASR systems against background noises. Methods include capturing of temporal cues from the speech signal (TempoRAl Patterns (TRAPs) [4]), spectral information from the speech signal (Mel Frequency Cepstral Coefficients (MFCC) [5] and Perceptual Linear Prediction (PLP) [6]). MFCC features are concatenated with their first and second-order temporal derivatives (i.e., delta and double-delta features), to capture temporal dynamics in the speech signal. It resulted in improvement in ASR performance and hence, became a big motivation to use *joint* spectro-temporal features for ASR task. Another motivation to use spectro-temporal features in ASR is the fact that our brain responds to joint spectro-temporal patterns in the speech signal rather than temporal-only or spectral-only patterns [7]. Biological studies indicate that neurons in the primary auditory cortex (*A1*) of mammals are explicitly tuned to spectro-temporal patterns [8] and different neurons are excited by different spectro-temporal patterns depending upon their Spectro-Temporal Receptive Fields (STRFs). Hence, it would be worthwhile to explore and analyze spectro-temporal features of speech signal since these features are physiologically motivated and it is already known that human speech recognition system is better than any ASR system.

1.3 Contributions in the Thesis

- In this thesis, spectro-temporal Gabor filterbank features, proposed in [8], are experimented on TIMIT database [9] with various additive noises like white, babble, volvo and high frequency noises, at various SNR levels, and are compared with the state-of-the-art MFCC features.
- Gammatone filters are used in place of the standard mel filters for the time-frequency representation of the speech signal. Concatenating cepstral coefficients with these specto-temporal features results in increase in the recognition performance and hence, MFCC features and Gammatone Frequency Cepstral Coefficients (GFCC) features are concatenated with both the spectro-temporal features to form new feature set. Experiments are performed with HTK (HMM toolkit) [10] as back-end, taking into consideration the strength of acoustic and language model.

1.4 Organization of the Thesis

The organization of the thesis is as follows.

Chapter 2 details the literature survey that has been done for the research problem. It covers some theory and motivation behind using joint spectro-temporal features for speech recognition task. Method to extract spectro-temporal features from the speech signal is also briefly described

Chapter 3 describes the cepstral analysis and significance of cepstral coefficients for speech recognition. In this context, state-of-the-art MFCC features and similar, GFCC features are described.

Chapter 4 details the theory and implementation of *1-D* and *2-D* Gabor filters. Furthermore, motivation of using Gabor filters for speech recognition task is presented.

Chapter 5 presents the method to extract spectro-temporal features named “Spectro-Temporal Gabor filterbank features” from the speech signal. Algorithm and architecture for the method is presented.

Chapter 6 details all the experiments that have been carried out for the ASR task. Comparison between proposed features and the baseline features is shown.

Chapter 7 summarizes and concludes thesis work, along with the future research directions.

1.5 Chapter Summary

In this chapter, an overview of ASR, covering its architecture and applications in real-life scenarios, has been presented. Real-life applications involve different background noises and there is a strong need to improve ASR performance in those conditions. Motivation to use joint spectro-temporal features in ASR task is presented in great detail. Finally, contributions in the thesis is briefly described. Next chapter details the theories and some methods utilizing joint spectro-temporal features for ASR task.

Chapter 2

Literature Survey

2.1 Introduction

Decades of research in the field of ASR brought several theories and methods to improve recognition performance of the ASR. Several methods have been proposed till date, namely MFCC [5], TRAPs (TempoRAl Patterns) [4], PLP (Perceptual Linear Prediction) [6], spectro-temporal Gabor filterbank features [8], etc. Features, like MFCC and PLP have been proposed in the late *90*'s and are considered as the state-of-the-art, because of low computational complexity and low-dimensional representation. Computation of the state-of-the-art MFCC features involves framewise FFT operation, followed by logarithmic operation and DCT. Thus, these features are able to capture only spectral information in the speech signal. Usage of first and second-order derivatives of MFCC features leads to capturing of some temporal dynamics in the speech signal. Concatenation of MFCC features with their first and the second-order derivatives results in increase in the recognition performance and thus, became one of the biggest motivations to use joint spectro-temporal features of the speech signal. Spectro-temporal features are also physiologically motivated as our brain responds to joint spectro-temporal patterns in the speech signal. Biological studies show that neurons in the primary auditory cortex (*A1*) of mammals are excited by joint spectro-temporal patterns in the speech signal. Because of the above studies, exploring and analyzing spectro-temporal features for the ASR task would be worthwhile. Different neurons in the *A1* are excited by different spectro-temporal patterns in the speech signal depending upon their Spectro-Temporal Receptive Fields (STRFs). The shape of STRF of a neuron looks like a *2-D* Gabor filter as shown in Figure 2.2 (b) [11]. Thus, Gabor filters are physiologically motivated and are utilized in the extraction of spectro-temporal features of the speech signal. A detailed description of the Gabor filters is given in chapter 4.

This chapter presents some history and theories behind joint spectro-temporal features of the speech signal and their analysis in the human brain. Work like spectro-temporal analysis using *2-D* Gabor filter and spectro-temporal Gabor filterbank features, are briefly described. Some description about the state-of-the-art MFCC features and the Gammatone counterpart GFCC features is also given.

2.2 Speech Processing Stages in Humans

Figure 2.2 (a) shows the stages of speech processing in humans, from the signal entering the ear, till brain. The speech signal enters the pinna of the outer ear and reaches the auditory portion of the inner ear known as *cochlea*. Cochlea contains a membrane called Basilar Membrane (BM) which vibrates at different points depending upon the incoming frequency of the sound. Thus, BM acts as a bank of bandpass filters wherein different regions of the BM pass different ranges of sound frequencies. Speech signal strikes the BM and nerve fibers at the surface of the BM are fired converting the sound wave to electrical signal. The electrical signal characterizes the time-frequency representation of the speech signal which reaches the primary auditory cortex (*A1*) in the brain. Uptil this, it is the early stage of processing. Early stage is followed by the cortical stage where the time-frequency representation of the speech signal is analyzed in the brain [12]. Primary auditory cortex consists of millions of neurons which are excited by the spectro-temporal patterns in the speech signal. Different neurons in the *A1* are excited by different spectro-temporal patterns in the speech signal depending upon their Spectro-Temporal Receptive Fields (STRFs). Neuron excitation is considered as a linear process and thus, the neural response, which is called *cortical* representation, is the convolution of the input time-frequency representation of the speech signal with the *2-D* impulse response of the neuron which is nothing but the STRF of the neuron. STRF of neuron in the *A1* looks like *2-D* Gabor function as in Figure 2.2 (b) [11]. Arrow indicates highly varying *2-D* impulse response region, where red and blue colors indicate region of strongly excitatory and suppressed responses.

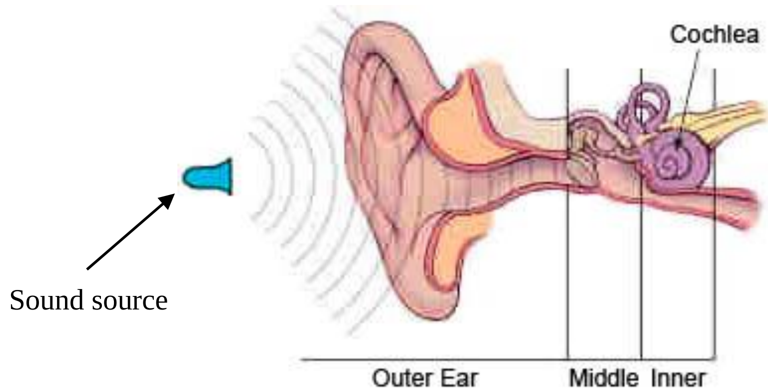


Figure 2.1: Human ear. Adapted from **[13]**.

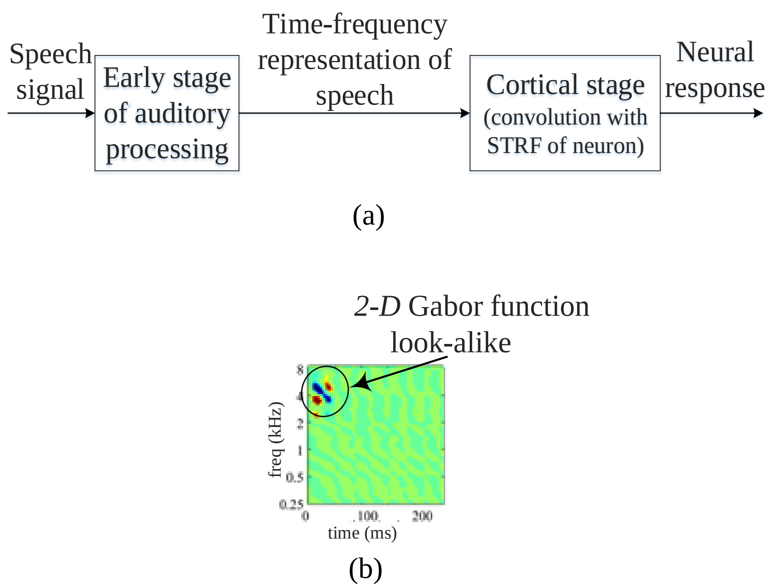


Figure 2.2: (a) Speech processing stages in humans, (b) STRF of a neuron in *A1*. Adapted from **[11]**.

2.3 Spectro-Temporal Analysis of Speech Using *2-D* Gabor Filter

A spectro-temporal analysis using *2-D* Gabor filter is done in [14]. Speech signal sampled at *16* kHz is taken and short-time Fourier transform (STFT), with Hamming window of *25 ms* duration, is calculated. Spectrogram is computed and patches from the spectrogram are extracted at every grid point (i, j). Size of the patch is taken as df and dt as height and width, respectively. The parameters df and dt must be large enough to be able to resolve the spectro-temporal components in the patch, however, small enough so that the underlying signal is stationary. Suitable parameter ranges are *5-15 ms* for *dt* and *600-800* Hz for *df*. Additionally, analysis parameters are the window hop sizes in time and frequency domains. Typically, these are taken as *3-5 ms* and *150- 350* Hz, which creates overlap between the patches. These patches are then multiplied with Gaussian window $W(f,t)$, as given in Eq. (2.2), located at the patch center (f_0,t_0) and are undergone *2-D* Fourier transform. The output we get is $R_{ij}(\Omega, \omega)$ given by:

$$R_{ij}(\Omega,\omega) = \sum_t \sum_f P_{ij}(f,t) W(f,t) e^{-j2\pi\frac{\Omega}{N_H}f}\; e^{-j2\pi\frac{\omega}{N_W}t}, \tag{2.1}$$

where $W(f,t)$ is defined as:

$$W(f,t) = \frac{1}{2\pi\sigma_t\sigma_f}\, e^{-\frac{1}{2}\left(\left(\frac{f-f_0}{\sigma_f}\right)^2+\left(\frac{t-t_0}{\sigma_t}\right)^2\right)}. \tag{2.2}$$

Here, $P_{ij}(f,t)$ is the patch at (i,j) point in the spectrogram and N_H x N_W is the Fourier transform size. The equation can be rewritten as

$$R_{ij}(\Omega,\omega) = \sum_t \sum_f P_{ij}(f,t) g_{\Omega,\omega}{}^*(f,t), \tag{2.3}$$

where $g_{\Omega,\omega}(f,t)$ is the Gabor filter function. This is nothing but the *2-D* Gabor transform of the patch and $R_{ij}(\Omega,\omega)$ can be viewed as the projection of a patch $P_{ij}(f,t)$ on the entire bank of spectro-temporal *2-D* Gabor filters $g_{\Omega,\omega}(f,t)$. In general, *2-D* spectro-temporal Gabor responses of spectrogram patches exhibit multiple Gaussian peaks that come in pairs. Each peak pair corresponds to a different spectro-temporal modulation contained in the patch. In effect, *2-D* Gabor filterbank decomposes a patch in the spectro-temporal components. Peaks with large amplitude in the spectro-temporal response $R_{ij}(\Omega,\omega)$ corresponds to dominant spectro-temporal modulations in the patch. The signal $R_{ij}(\Omega,\omega)$ which we get contains many useful properties, *viz.*, harmonicity, formants, plosive onset/offset and noise. Harmonic patch contains horizontal striations. Thus, the Gabor transform of the patch, which is nothing but the *2-D* Fourier transform of the windowed (Gaussian-windowed) patch, would be two dominating Gaussians on the spectral modulation Ω-axis. On the contrary, plosive patch contains vertical edges and thus, the Gabor transform would be two dominating Gaussians on the temporal modulation ω-axis. Figure 2.3 shows an exemplar harmonic patch and a plosive patch and corresponding spectro-temporal responses $R_{ij}(\Omega,\omega)$.

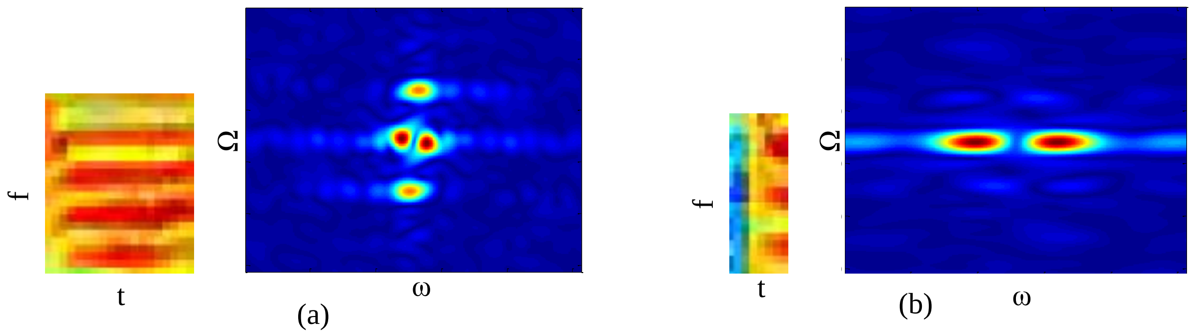


Figure 2.3: Pair of patch and corresponding spectro-temporal response $R_{ij}(\Omega,\omega)$ for harmonic and plosive segment, in (a) and (b), respectively.

2.4 Spectro-Temporal Gabor Filterbank Features for ASR

Features proposed by Marc. Schadler *et. al.* are purely physiologically motived [8]. Marc. Schadler *et. al.* tried to mimic speech recognition of mammals in ASR task. Speech recognition system of mammals is better than any ASR system and thus, the algorithm is intended to perform good enough in noisy environments [8]. Algorithm takes log-mel spectrogram (spectro-temporal patterns as input to neurons in *A1*) and passes it through a bank of *2-D* Gabor filters (real part of Gabor filters, as *2-D* impulse response of neurons known as STRFs) to generate corresponding time-frequency representations, known as the cortical representations. Outputs of all the Gabor subband filters are concatenated to form the features. These features are quite high-dimensional due to concatenation of the outputs of all the Gabor subband filters and thus, a channel selection scheme, described in [8], is applied to reduce the dimension. In [8], % word error rate (WER %) have been calculated for AURORA2 and NUMBERS95 databases with HTK and SRI Decipher recognizer, respectively, for various noisy conditions. These features were shown to perform better than the state-of-the-art MFCC features for AURORA2 database (with HTK as back-end) but were outperformed by MFCC features for NUMBERS95 database (with SRI Decipher recognizer), for clean as well as noisy conditions. A detailed description of spectro-temporal Gabor filterbank features is given in chapter 5.

2.5 Multi-Stream Spectro-Temporal Features for ASR

Another method to extract spectro-temporal features that focuses on dividing the features into multiple parallel streams based on the spectro-temporal modulation frequencies [15]. Modulation frequencies are evenly taken upto *16* Hz and *0.5* cyc/channel as humans are most sensitive uptil these frequencies. Each feature stream can be viewed as a patch of information in the spectro-temporal receptive field. Comparison (in terms of WER) of the features with the standard MFCC is done on the NUMBERS95 database. Experiments were conducted using TANDEM recognition system [16]. The spectro-temporal features were shown to perform better than the MFCC features in clean as well as signal degradation conditions. For better comprehension and the recognition results, [15] can be referred.

2.6 Cepstral Coefficients as Features for ASR

Cepstrum of a speech segment is defined as the spectrum of the log-magnitude spectrum of the speech segment. Cepstral analysis is a method to separate source (speaker) and system (vocal tract) information. System information, which is important for speech recognition task, is present at lower frequencies and can be extracted by using only lower FFT samples. Mel Frequency Cepstral Coefficients (MFCC) are widely used cepstral coefficients and are considered to be state-of-the-art features for the ASR task because of low computational complexity and low-dimensional representation. MFCC incorporate mel filterbank to mimic the frequency mapping in the BM. Human ear has better resolution for lower frequencies than the higher frequencies. Hence, mel filters are physiologically motivated as they represent the filterbank at the BM. These filters, though, are arranged tonotopically on the frequency-axis, the shape is kept triangular to reduce the complexity in designing them. Another type of filterbank which is used to mimic BM is Gammatone filterbank. Gammatone function is multiplication of Gamma function and a tone (sinusoid) function. Unike mel filters, Gammatone filters are not triangular and are being used in place of mel filters for the extraction of cepstral coefficients to give Gammatone Frequency Cepstral

Coefficients (GFCC) [17]. Lower cepstral coefficients capture mostly spectral information in the speech signal and capture some temporal dynamics in the speech signal when are concatenated with first and second-order derivatives, however, are unable to capture more local joint spectro-temporal information in the speech signal. Chapter 3 presents a detailed description of both the cepstral coefficients, *viz.*, MFCC and GFCC.

2.7 Chapter Summary

In this chapter, a brief analysis of speech signal in the mammal's brain was presented. The analysis covered the importance of joint spectro-temporal information in the speech signal for speech recognition task. In addition, some theories behind and analysis of spectro-temporal information in the speech signal, like spectro-temporal analysis of speech signal using *2-D* Gabor filters, were briefly discussed. In addition, one of the methods to extract joint spectro-temporal features from the speech signal, like method for the extraction of spectro-temporal Gabor filterbank features, was briefed. Cepstral coefficients, used for the speech recognition task, like MFCC and GFCC, were also described in brief. Detailed descriptions of all of the methods/ features listed above are given in the subsequent chapters.

Chapter 3

Cepstral Coefficients – MFCC and GFCC

3.1 Introduction

Spectrum of a signal is the frequency-domain representation of the signal. *Cepstrum* is the spectrum of the log-magnitude spectrum of the signal. Signals that are added together and have disjoint spectral content can be separated by linear filtering. However, signals, often are not additively combined. In linear speech production model, source and system are convolutionally combined and thus, these components cannot be separated by linear filtering. For speech recognition task, only system information is important and source information is to be removed. This is because speech sound unit is perceived by the realization of several resonances in the vocal tract and thus, source information (pitch information) is of insignificant use. Hence, to reduce the redundancy of the ASR system, it is indeed imperative to separate source and system information in the speech signal and to use only the system (vocal tract) information.

Several methods, like linear prediction and homomorphic filtering [18], are there in the literature to separate nonlinearly combined signals. Linear prediction method is based on the prediction of current speech sample by having linear combination of some past samples. The error between the actual sample and the predicted sample, known as the LP residual, is minimized and the coefficients of the linear combination are obtained. These coefficients define nothing but the system characteristics and thus, obtaining the formant structures of the vocal tract. Hence, linear prediction analysis first extracts the system component through inverse filtering and then, it extracts the *excitation source* component. Further details regarding linear prediction of speech can be found in [18]. Another approach to separate signals that are non-linearly combined is homomorphic filtering. In this approach, convolutionally combined signals are first mapped to additively combined signals and then linear filtering is applied to separate them. Unlike linear prediction analysis, which is parametric approach, homomorphic filtering is nonparametric approach for deconvolution. In homomorphic filtering, specific model need not be imposed on the system transfer function in analysis. Cepstral analysis comes under homomorphic filtering where source and system components that are combined convolutionally (multiplied in frequency-domain), are separated by applying logarithmic operation or root operation in frequency-domain and then linear filtering, which would then be called as *liftering*, due to filtering operation in frequency-domain, is applied to separate the components. MFCC [5] and GFCC [17] are being used, by speech researchers, as the features for ASR task and they are the lower FFT coefficients that indicate nothing but the system information. These features also incorporate the function of the BM in the inner ear by utilizing mel filterbank and Gammatone filterbank [19]. These filterbank mimic the frequency mapping in the BM thus making the features capture more perceptual information from the speech signal. The features capture only spectral information in the speech signal (due to framewise FFT operation) and hence, are concatenated with the first and the second-order derivatives to capture some temporal dynamics in the speech signal. These features, however, unlike spectro-temporal features [8], are unable to capture local joint spectro-temporal information, by which a speech signal is characterized.

3.2 Cepstral Analysis

In linear speech production model, source signal (periodic, noisy or impulse) is convolved with the impulse response of the vocal tract to produce intelligible speech signal. Figure 3.1 shows the block diagram of the speech production mechanism.

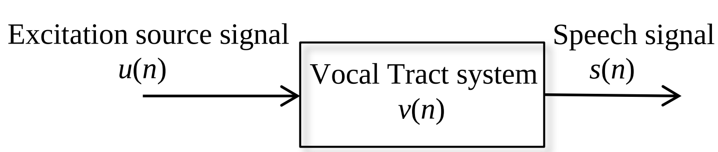


Figure 3.1: Speech production system.

In time-domain, output speech signal $s(n)$ is expressed as

$$s(n) = u(n) * v(n). \tag{3.1}$$

In frequency-domain, source signal is multiplied with the vocal tract function to give the speech signal (due to convolution theorem) as in Eq. (3.2).

$$S(\omega) = U(\omega)V(\omega), \tag{3.2}$$

where $S(\omega)$, $U(\omega)$ and $V(\omega)$ are the frequency-domain representation of $s(n)$, $u(n)$ and $v(n)$, respectively. For speech recognition, only system information, i.e., $U(\omega)$ is relevant and thus, there is a need to separate $U(\omega)$ and $V(\omega)$ which is done by applying log-operator as shown below.

$$log\{S(\omega)\} = log\{U(\omega)\} + log\{V(\omega)\}. \tag{3.3}$$

In this case, $log\{U(\omega)\}$ is present at lower side on the frequency-axis, i.e., $log\{U(\omega)\}$ contains lower frequencies as compared to $log\{V(\omega)\}$. Thus, $log\{U(\omega)\}$ is separated by applying lowpass filter. Fourier transform of $log\{S(\omega)\}$ is known as the *cepstrum* of $s(n)$.

3.3 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCC) [5] are the state-of-the-art features for speech as well as speaker recognition task because of their low-computational complexity and lower dimension representation. Architecture for the extraction of MFCC features, for speech recognition, is shown is Figure 3.2.

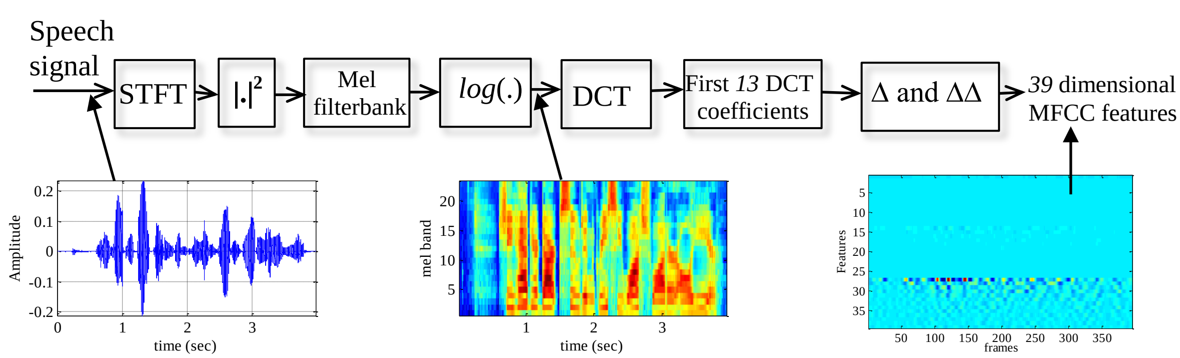


Figure 3.2: Architecture for MFCC feature extraction.

First, STFT, as defined in Eq. (3.4) of the speech signal is computed. $s[n]$ refers to the speech signal and $w[n]$ is the window. Squared absolute value of STFT, which represents the spectrogram, is taken and is passed through the mel filterbank. Mel scale (defined in Eq. (3.5)) [20] is perceptually motivated scale that mimics the mapping of frequencies to specific regions of Basilar Membrane (BM) in the inner ear. Human ear has lower resolution for higher frequencies and therefore, mel scale is logarithmic for frequencies above *1* kHz. Mel filters are taken as triangular in shape and a bank of *23* filters, aligned from *100* Hz to *8000* Hz, is shown in Figure 3.3. The output values of the spectrogram are compressed with the logarithm, roughly resembling the amplitude compression performed by the auditory system. Log-operation is also motivated from the cepstral analysis, i.e., to separate source and system information. DCT is applied to the output and lower coefficients (generally first *13*) are extracted, that represent a feature vector for speech recognition. Δ and $\Delta\Delta$ operations, to capture speech transition or the temporal dynamics in speech signal, are applied to the coefficients resulting in *39*-dimensional (*13*+*13*+*13*) feature vector. Δ and $\Delta\Delta$ capture transitions from a speech frame to the next frame and to the second next frame, respectively.

$$S(k,\tau) = \sum_{n=0}^{N-1} s[n]w[n,\tau]e^{-j2\pi kn/N}, \tag{3.4}$$

$$Mel(f) = 2595log_{10}\left(1 + \frac{f}{700}\right). \tag{3.5}$$

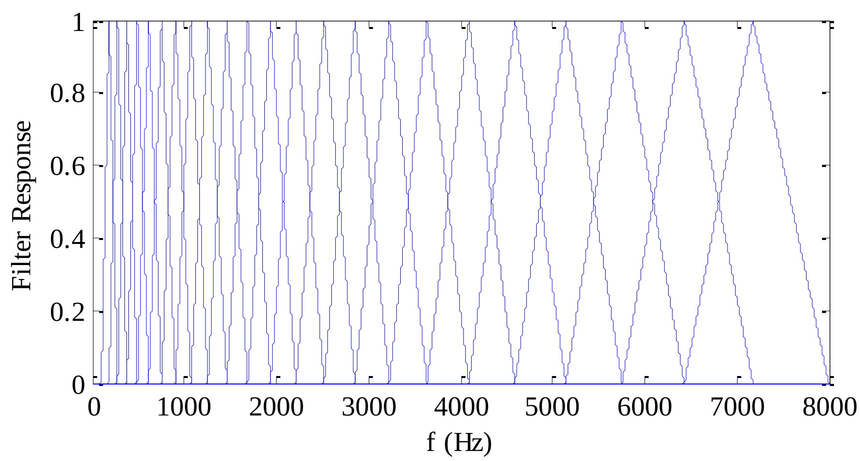


Figure 3.3: Mel filterbank consisting *23* filters.

An inherent characteristic that MFCC have is that the features are almost uncorrelated, which is due to the DCT operation. Uncorrelated features do not introduce redundancy and increase the efficiency.

3.4 Gammatone Frequency Cepstral Coefficients

Gammatone Frequency Cepstral Coefficients (GFCC), proposed by Shao *et.al.* [17], are same as MFCC, but incorporate Gammatone filterbank instead of mel filterbank. Architecture for the extraction of GFCC features is shown in Figure 3.3.

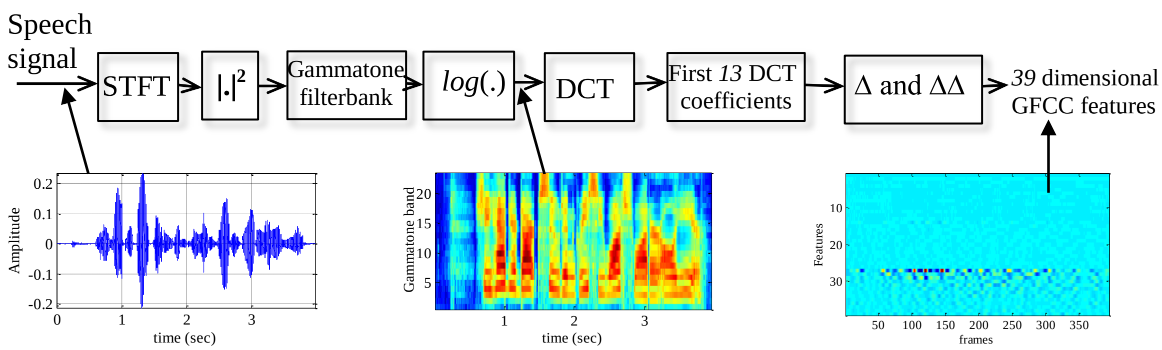


Figure 3.4: Architecture for GFCC feature extraction.

The impulse response of Gammatone filter centered at frequency f (Hz) is defined as:

$$\gamma(t) = \begin{cases} t^{a-1} e^{-2\pi bt} cos(2\pi f t), & t \geq 0 \\ 0 \quad , & else \end{cases}, \tag{3.6}$$

where a refers to the order of the filter and b is the rectangular bandwidth which increases with the center frequency f. Gammatone filters are aligned on ERB scale [21], which is another type of frequency scale used to mimic BM. Frequency on ERB scale, $ERBS$ is related to the linear frequency f (Hz) by:

$$ERBS(f) = 11.17 ln\left(1 + \frac{46.06538 f}{f + 14678.49}\right). \tag{3.7}$$

Gammatone filters of order *4* provide an excellent fit to the human auditory filters. Bandwidth of the filters increases with the center frequency and is given by:

$$b = 1.019 * 24.7(4.37 * f * (10^{-3})f + 1). \quad (3.8)$$

Figure 3.4 shows Gammatone filterbank of *23* filters aligned on the linear frequency scale, from *100* Hz to *8000* Hz.

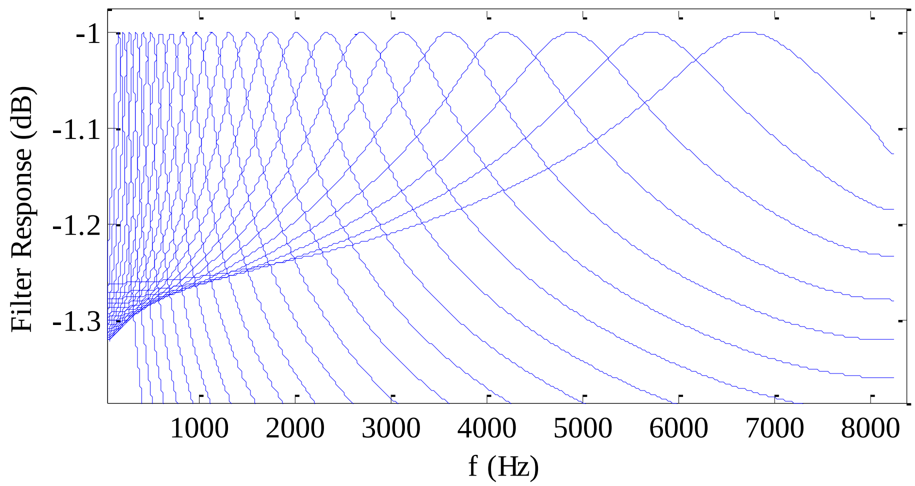


Figure 3.5: Gammatone filterbank consisting *23* filters.

3.5 Chapter Summary

In this chapter, cepstral analysis and theory behind cepstral coefficients have been discussed. Mel and Gammatone filterbanks were described. Further, cepstral coefficients like MFCC and GFCC that are being used as features for ASR task, are described along with their architectures.

Chapter 4

Gabor filters

4.1 Introduction

Gabor filter, for the first time was used in image processing and its terminology was coined by Dennis Gabor. It is a linear filter and initially was used for edge detection in images. The invention of Gabor filters was from the fact that frequency and orientation representations of the human visual systems are similar to that of the Gabor filters. These filters have been found to be apt for texture representation and discrimination. In image processing, the usage of Gabor filters was physiologically motivated as the cells/neurons in the visual cortex of the brain of a mammal can be modeled by *2-D* Gabor functions. In other words, *spatio-temporal receptive fields* (STRFs) of neurons in the visual cortex of mammalian's brain look like *2-D* Gabor functions. For the last decade, research has been going on to incorporate Gabor filters in speech processing. Like, in visual cortex, neurons in the primary auditory cortex (*A1*) of mammalian's brain can also be closely modeled by *2-D* Gabor functions. Thus, speech analysis with Gabor filters is thought to be similar to perception in the human auditory system. Further sections present the detailed theory of Gabor filters, both *1-D* and *2-D* and the application of the filters in ASR.

4.2 *1-D* Gabor Filter

Gabor filter is defined as the product of complex sinusoid and Gaussian kernel. Gaussian kernel is defined by:

$$w(t) = e^{-t^2}. \tag{4.1}$$

Gabor filter is basically complex sinusoid modulated by Gaussian window and thus, the filter is a complex filter. Algorithm for the extraction of spectro-temporal Gabor filterbank features incorporates Hanning window instead of Gaussian window as Gaussian window has infinite support and cutting off the window would lead to unwanted higher harmonic frequencies in the modulation frequency domain. These distortions can, to a great extent, be eliminated using Hanning window [22]. Therefore, in this chapter, design of Gabor filters with Hanning window will be presented. Eq. (4.4) represents the *1-D* Gabor filter.

4.2.1 Time-domain Representation of *1-D* Gabor Filter

$$h_b(t) = \begin{cases} 0.5 - 0.5\, cos\left(\frac{2\pi t}{b}\right), & 0 < t < b \\ 0, & else \end{cases} \tag{4.2}$$

$$s_\omega(t) = e^{j\omega t}, \tag{4.3}$$

$$g(t) = s_\omega(t) h_{\frac{v}{2\omega}}(t - t_0). \tag{4.4}$$

Here, $h_b(t)$ represents the Hann window and $s_\omega(t)$ represents the complex sinusoid with fundamental frequency as ω. v denotes number of semi-cycles of the sinusoid under the Hann envelope. t_0 is the value by which Hann window is shifted so as to align the filter at the origin. The shifted Hann window is denoted by $h_{bs}(t)$. The value of x_0 would be -$\frac{b}{2}$. For the extraction of spectro-temporal features, as will be discussed in chapter 5, only real part of the filter will be used. Figure 4.1 shows the time-domain representation of the real part of the Gabor function, along with Hann envelope and the sinusoid carrier.

4.2.2 Frequency-domain Representation of *1-D* Gabor Filter

$h_b(t)$ can be represented as:

$$h_b(t) = \left\{0.5 - 0.5\cos\left(\frac{2\pi t}{b}\right)\right\} rect\left(\frac{t}{b} - \frac{1}{2}\right). \tag{4.5}$$

Hence, Fourier transform of $h_b(t)$ would simply be the convolution of the Fourier transform of the two terms in the product. *rect* function gives *sinc* in frequency-domain, while the first term in the product gives *3* impulses, at *0* and $\pm\frac{1}{b}$. Hence, Fourier transform would be *3 sinc* functions at *0* and $\pm\frac{1}{b}$. Since, this Hann window is to be placed at *0*, Fourier transform of the shifted window would be considered. Fourier transform of the shifted Hann window would then be:

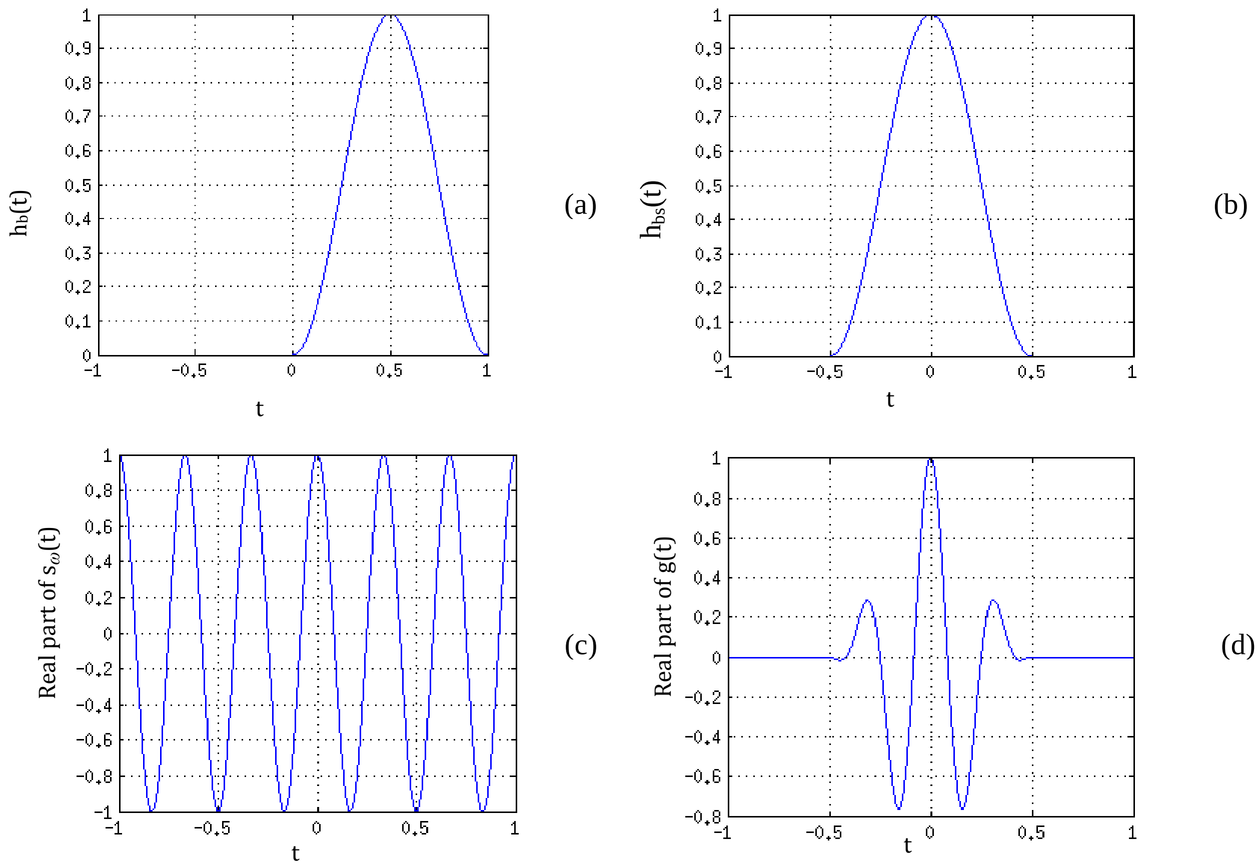
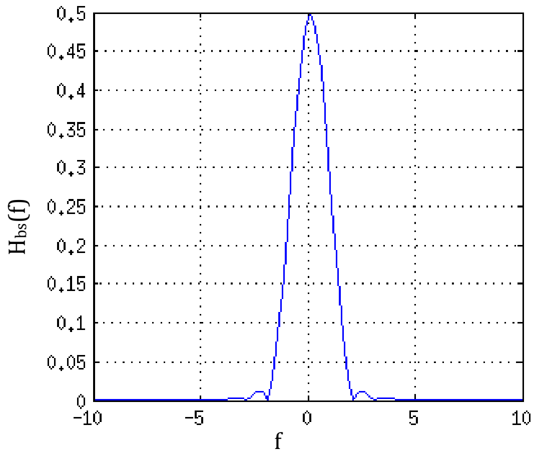


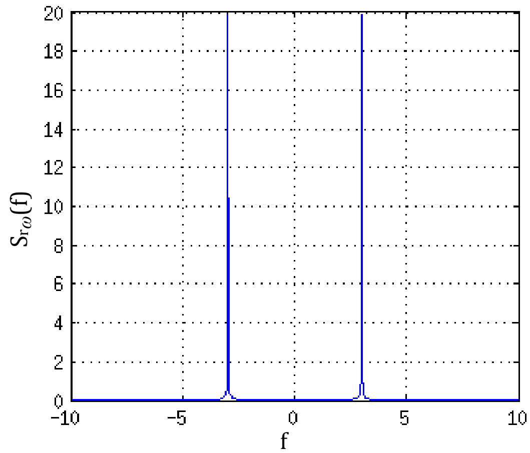
Figure 4.1: (a) Hann envelople $h_b(t)$, (b) shifted envelope $h_{bs}(t)$, (c) real part of complex sinusoid $s_\omega(t)$, (d) real part of Gabor function $g(t)$. Parameters taken are $\omega=6\pi$, $b=1$.

$$H_{bs}(f) = 0.5sinc(bf) + 0.25sinc(bf - 1) + 0.25sinc(bf + 1). \qquad (4.6)$$

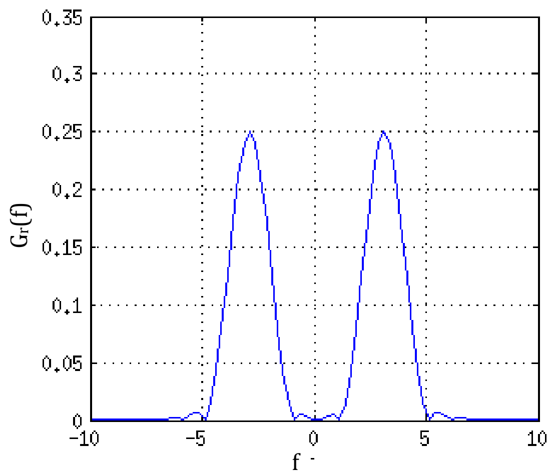
The shifted Hann window is then multiplied with complex sinusoid $s_\omega(t)$ to produce Gabor function. Thus, Fourier transform of the Gabor function would just be $H_{bs}(f)$ delayed by f_0, where $f_0 = \frac{2\pi}{\omega}$ (Hz). Figure 4.2 shows the spectrum of the real part of *1-D* Gabor function denoted by $G_r(f)$, along with that of the Hann envelope and the real part of the complex sinusoid denoted by $S_{r\omega}(f)$.



(a)



(b)



(c)

Figure 4.2: (a) Magnitude spectrum of shifted Hann envelope $H_{bs}(f)$, (b) magnitude spectrum of real part of carrier $S_{r\omega}(f)$, (c) magnitude spectrum of real part of the Gabor function $G_r(f)$. Parameters taken are $\omega=6\pi$, $b=1$. All frequencies are stated in Hz.

4.3 *2-D* Gabor Filter

2-D Gabor function is the function of *2* variables. In the context of speech signal, the independent variables are time and frequency, since speech signal is characterized by the variation in time and frequency. *2-D* Gabor filters are actually the filters which are being used in image and speech processing for that they are physiologically motivated and used to mimic human visual and speech perception.

4.3.1 Time-domain Representation of *2-D* Gabor Filter

$$h_{b_n b_k}(n,k) = h_{b_n}(n) h_{b_k}(k), \tag{4.7}$$

$$s_{\omega_n \omega_k}(n,k) = s_{\omega_n}(n) s_{\omega_k}(k), \tag{4.8}$$

$$g(n,k) = h_{b_n b_k}(n,k) s_{\omega_n \omega_k}(n,k). \tag{4.9}$$

The localized *2-D* Gabor filters are defined in Eq. (4.9), with the channel and time-frame variables, k and n, respectively; ω_k and ω_n the spectral and the temporal modulation frequencies, respectively.

$$b_x = \frac{v_x}{2\omega_x}, \tag{4.10}$$

v_k and v_n the number of semi-cycles under the envelope in spectral and temporal dimension, respectively. A *2-D* Gabor filter is the product of a *2-D* complex sinusoid carrier (4.8) with the corresponding modulation frequencies ω_k and ω_n, and an *2-D* envelope function defined in Eq. (4.7). p and q denote the shift in the envelope so as to align the Gabor filter at origin. Figure 4.3 shows the real part of the *2-D* Gabor function, along with *2-D* Hann envelope and *2-D* sinusoid carrier.

Change in the modulation frequencies ω_k and ω_n leads to rotation of the Gabor filter. Thus, *2-D* Gabor filters can also be modeled by rotation parameters. Further details about *2-D* Gabor filter are given in [23].

4.3.2 Frequency-domain Representation of *2-D* Gabor Filter

Concept of Fourier domain representation of *2-D* Gabor function is same as that described in Section 4.2.2, but now the representation would extend to two dimensions. For the spectrum of real part of *2-D* Gabor function, we would get *2* lobes centered at $(\pm\omega_n,\omega_k)$. Figure 4.2 shows the spectrum of the real part of *2-D* Gabor function denoted by $G_r(f_n, f_k)$, along with that of the Hann envelope and the real part of the complex sinusoid denoted by $S_r(f_n, f_k)$.

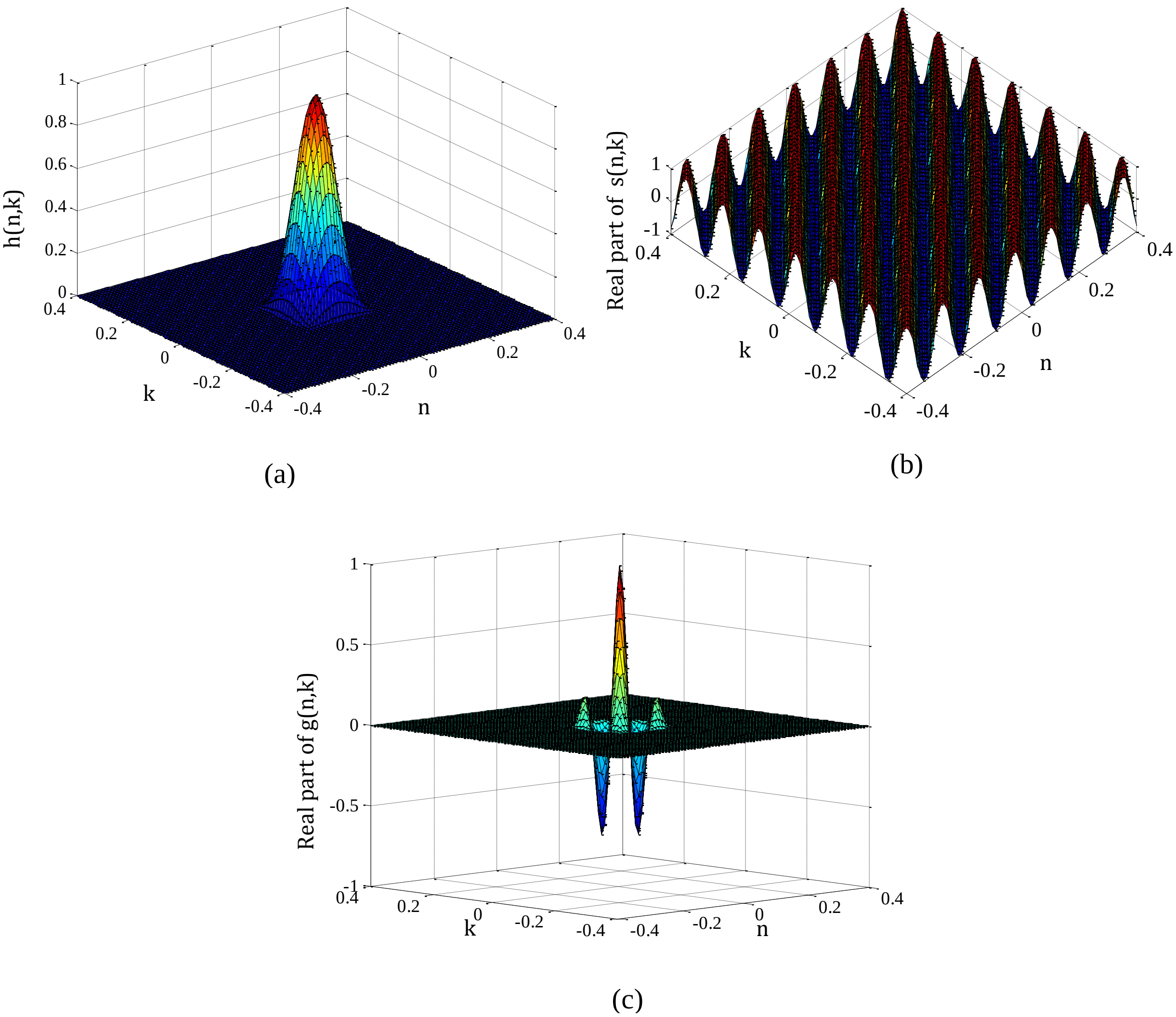


Figure 4.3: (a) Shifted *2-D* Hann envelople $h(n,k)$, (b) real part of *2-D* complex sinusoid $s(n,k)$, (c) real part of *2-D* Gabor function $g(n,k)$. Parameters taken are $\omega_n=16\pi$ rad-Hz, $\omega_k=16\pi$ rad-cyc/chan, $\nu_n= \nu_n=7\pi$.

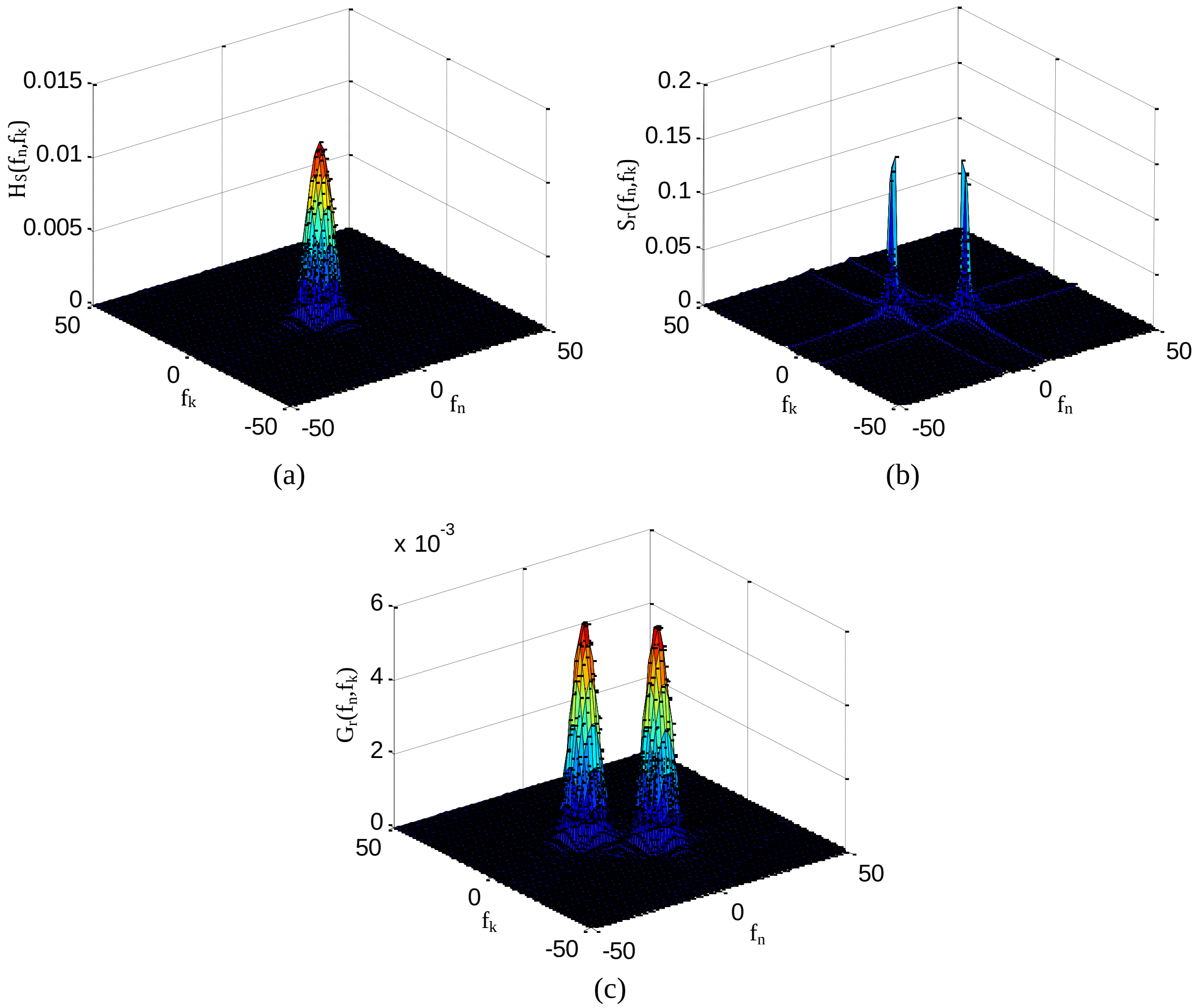


Figure 4.4: (a) Magnitude spectrum of shifted *2-D* Hann envelope $H_s(f_n,f_k)$, (b) magnitude spectrum of real part of *2-D* carrier $S_r(f_n,f_k)$, (c) magnitude spectrum of real part of the *2-D* Gabor function $G_r(f_n,f_k)$. Parameters taken are $\omega_n=16\pi$ rad-Hz, $\omega_k=16\pi$ rad-cyc/chan, $v_n= v_n=7\pi$. All frequencies are stated in Hz.

4.4 Gabor Filters in ASR

Gabor filters are now widely being used in speech recognition simply because they mimic human perception system and human's speech recognition system is better than any ASR system. Neurons in the primary auditory cortex of mammalian brain can be modeled by *2-D* Gabor functions since excitation patterns of the neurons closely match the patterns in the Gabor filters. In other words, *2-D* impulse response of neurons, known as STRF, in the primary auditory cortex, looks like *2-D* Gabor filter. Figure 2.2 shows the shape of STRF of a neuron of primary auditory cortex of a ferret. [12] and [24] give a detailed study and analysis of usage of Gabor filter for speech analysis. Gabor filters, thus, are being used by speech recognition researchers to extract physiologically motivated features from the speech signal and improve the efficiency of ASR. One such feature set is described in chapter 5.

4.5 Chapter Summary

This chapter describes the implementation of *1-D* and *2-D* Gabor filter. The implementation covers the time-domain as well as the frequency-domain implementation of the filters. The shape of the filters indicates that the Gabor filters are nothing but a type of bandpass filters and they are widely being used in the image and speech processing. Gabor filters are physiologically motivated and they mimic human perception system. In speech recognition, Gabor filters are being used for the extraction of physiologically motivated features so as to increase the recognition performance of ASR. The following chapter describes the method to extract joint spectro-temporal features, named spectro-temporal Gabor filterbank features, from the speech signal.

Chapter 5

Spectro-Temporal Gabor Filterbank features for Robust ASR

5.1 Introduction

Recently, ASR technology is being used in practical scenarios that involve various noises and channel effects and hence, robustness of ASR is becoming increasingly important. State-of-the-art MFCC features are known to be affected by acoustic noise. Hence, to cater to the needs of practical real-life applications that incorporate ASR, there is a need to come up with features that are able to perform good enough in the presence of background noise. Spectro-temporal Gabor filterbank (GBFB) features, proposed by Marc Schadler *et.al.* [8] intend to perform better in such signal degradation or noisy conditions. This is because these features are more physiologically motivated than MFCC features. Marc Rene Schadler *et.al.* [8] tried to mimic speech recognition of mammals, in ASR task. Algorithm takes log-mel spectrogram (i.e., spectro-temporal patterns as input to neurons in A*1*) and passes it through a bank of *2-D* Gabor filters (real part of Gabor filters, as *2-D* impulse response of neurons known as Spectro-Temporal Receptive Fields) to generate corresponding time-frequency representations (i.e., cortical representations). Speech processing in humans is described in Section 2.2. These cortical representations make up GBFB features. Following Section describes the feature extraction in detail. In contrast to mel filterbank, Gammatone filterbank [19] is also used to mimic frequency mapping in BM and thus, producing log-Gammatone spectrogram which acts as the input to Gabor filterbank.

5.2 Spectro-Temporal Feature Extraction

Figure 5.1 shows the architecture for spectro-temporal feature extraction from the speech signal. Log-mel or log-Gammatone spectrogram is passed through *2-D* Gabor filterbank to generate time-frequency representations corresponding to the Gabor filters. These time-

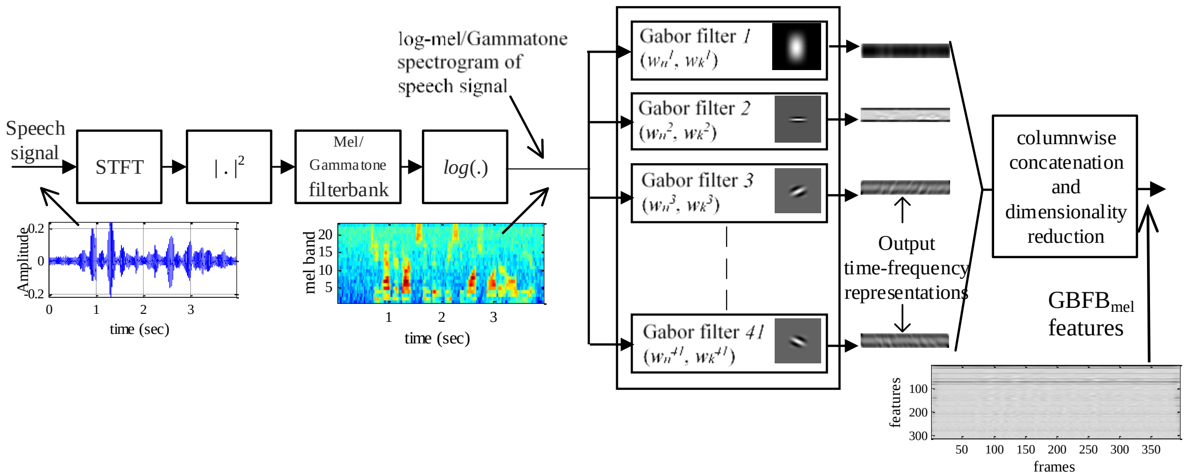


Figure 5.1: Architecture for spectro-temporal feature extraction. After **[25]**.

frequency representations are combined and dimensionality is reduced to form $GBFB_{mel}$ or $GBFB_{Gamm}$ features.

5.2.1 Log-mel Spectrogram as Input to Gabor Filterbank

Mel scale is logarithmic for frequencies above *1* kHz and therefore, mimics the mapping of frequencies to specific regions of BM in the inner ear. The output values of the spectrogram are compressed with the logarithm, roughly resembling the amplitude compression performed by auditory system. Spectrogram is expressed as:

$$S(k,\tau) = \sum_{n=0}^{N-1} x[n]w[n,\tau]e^{-j2\pi kn/N}, \tag{5.1}$$

$$X(k,\tau) = |S(k,\tau)|^2, \tag{5.2}$$

where $x[n]$ is the speech signal, $w[n]$ is the window function, τ is the time frame, N is the window length in samples and $S(k,\tau)$ *is* the short-time Fourier transform (STFT). Spectrogram is represented by $X(k,\tau)$. Log-mel spectrogram for the segment of clean speech signal, *dr1_faks0_sa1* from TIMIT database (with sampling frequency *16* kHz) and for additive white noise with *10* dB SNR are shown in Figure 5.2. Parameters used for calculating log-mel spectrogram are – window (Hanning) length = *25 ms*, window shift = *10* *ms*, number of channels/subband filters in mel filterbank = *23*, with center frequencies ranging from *100* Hz to *8000* Hz. It can easily be seen from Figure 5.2 (a region indicated by dotted circles), that joint spectro-temporal intensity pattern in the noisy signal has varied significantly from that of the clean version and thus, recognizing speech from a noisy speech signal is indeed a challenging task.

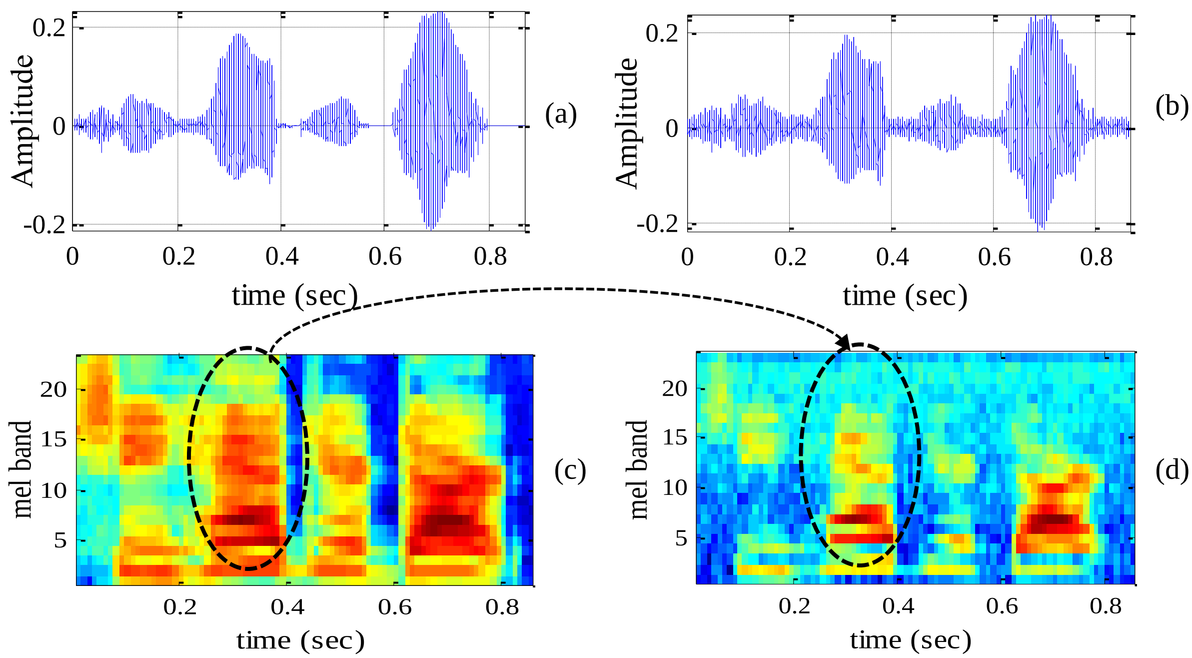


Figure 5.2: (a) Segment of the clean speech signal *dr1_faks0_sa1_te* from TIMIT database (Fs=*16* kHz), (b) signal in (a) with additive white noise at *10* dB SNR level, (c) log-mel spectrogram of clean signal, (d) log-mel spectrogram of the noisy speech signal.

5.2.2 Log-Gammatone Spectrogram as Input to Gabor Filterbank

Gammatone filterbank is commonly used filterbank to simulate the motion of the basilar membrane (BM) in the cochlea. Slaney's auditory toolbox is used to generate the Gammatone filterbank [19]. Figure 5.3 shows log-Gammatone spectrogram for the segment of clean speech signal, *dr1_faks0_sa1* from TIMIT database and for additive white noise with *10* dB SNR. Parameters used for calculatimg log-Gammatone spectrogram are the same as that for calculating log-mel spectrogram. Dotted circles in Figure 5.3 indicate the variation in the spectro-temporal intensity from clean to the noisy speech signal.

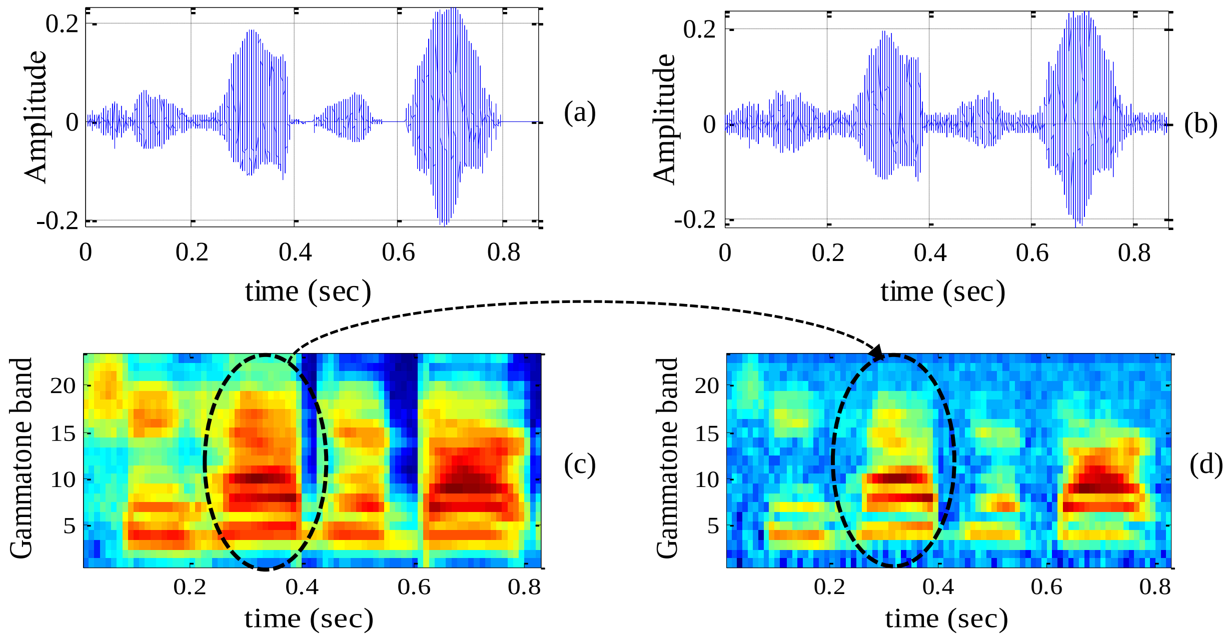


Figure 5.3: (a) Segment of the clean speech signal *dr1_faks0_sa1_te* from TIMIT database (Fs=*16* kHz), (b) signal in (a) with additive white noise at *10* dB SNR level, (c) log-Gammatone spectrogram of clean signal, (d) log-Gammatone spectrogram of the noisy speech signal.

5.3 Gabor Filterbank

2-D Gabor filterbank is already introduced in the previous chapter. The definition given via Eq. (4.9) would lead to infinite support for purely temporal or purely spectral modulation (ω_k=0 or ω_n=0) filters. Thus, filter size is limited to *69* frequency channels and *40* time frames. There is a linear relationship between the modulation frequency and the extension of the envelope (Eq. (4.7) - (4.9)) and hence, all the filters with same values for v_k and v_n are constant Q (i.e., the quality factor) filters. DC bias of each filter is removed since relative energy fluctuations are important for speech classification. Mean removal on a logarithmic scale is same as dividing on a linear scale and thus, this corresponds to normalization. While cepstral coefficients normalize spectrally, and RASTA (Relative Spectra) [26] processing and discrete derivatives normalize temporally, DC-free Gabor filters naturally normalize in both directions.

Temporal modulation frequencies up to *16* kHz and spectral modulation frequencies up to *0.5* cycle/channel are most sensitive to humans [7] and therefore, best performance is attained if maximum modulation frequencies of the filters are around these values. Empirically, we found that maximum modulation frequencies of *12.5* Hz and *0.25* cycle/channel produced the best performance. With the aim of evenly covering the modulation transfer space, modulation frequencies of the filterbank are decided as in (5.3), (5.4).

$$\omega_x^{i+1} = \omega_x^{i} \frac{1+c/2}{1-c/2}, \tag{5.3}$$

$$c = d_x \frac{8}{v_x}, \tag{5.4}$$

where d_x (in *x*-domain) is the distance factor between the two adjacent filters. Gabor filters with following frequencies are considered.

ω_k = -0.25, -0.12, -0.06, -0.03, 0, 0.03, 0.06, 0.12, 0.25,

ω_n = 0, 3.09, 4.92, 7.84, 12.5,

in cycles/channel and Hz, respectively. Hence, *41* unique *2-D* spectro-temporal Gabor filters are achieved whose real parts are used to process the log-mel/Gammatone spectrogram of the speech signal. The parameters for Gabor filterbank used here are given in Table 1. These parameters, empirically, found to perform the best and thus, are used for the speech recognition task considered in this work.

Table 5.1: Parameters used for Gabor filterbank

Parameters	ω_n (max)	ω_k (max)	v_n	v_k	d_n	d_k
Value	12.5	0.25	3.5	3.5	0.2	0.3

5.4 Output of the Gabor Filter

Log-mel/Gammatone spectrogram is convolved (i.e., *2-D* convolution) with the real part of the Gabor filter to get time-frequency representation that contains patterns matching the modulation frequencies associated with the filter (Figure 5.4). The dimension of the output time-frequency representation, of a single filter, is same as that of the log-mel spectrogram of the speech signal, i.e., *23* (number of mel channels) $\times$ number of frames of the speech signal. Results of all the *41* Gabor subband filters are concatenated columnwise to form the features. Some Gabor filters with different combinations of modulation frequencies (ω_n, ω_k), and corresponding outputs of noisy speech signal (generated by adding white noise at *10* dB SNR, to clean speech signal from TIMIT database), are shown in Figure 5.4. Arrows indicate the orientation of the Gabor filter as different combination of modulation frequencies (ω_n, ω_k) leads to different orientation of the Gabor filter.

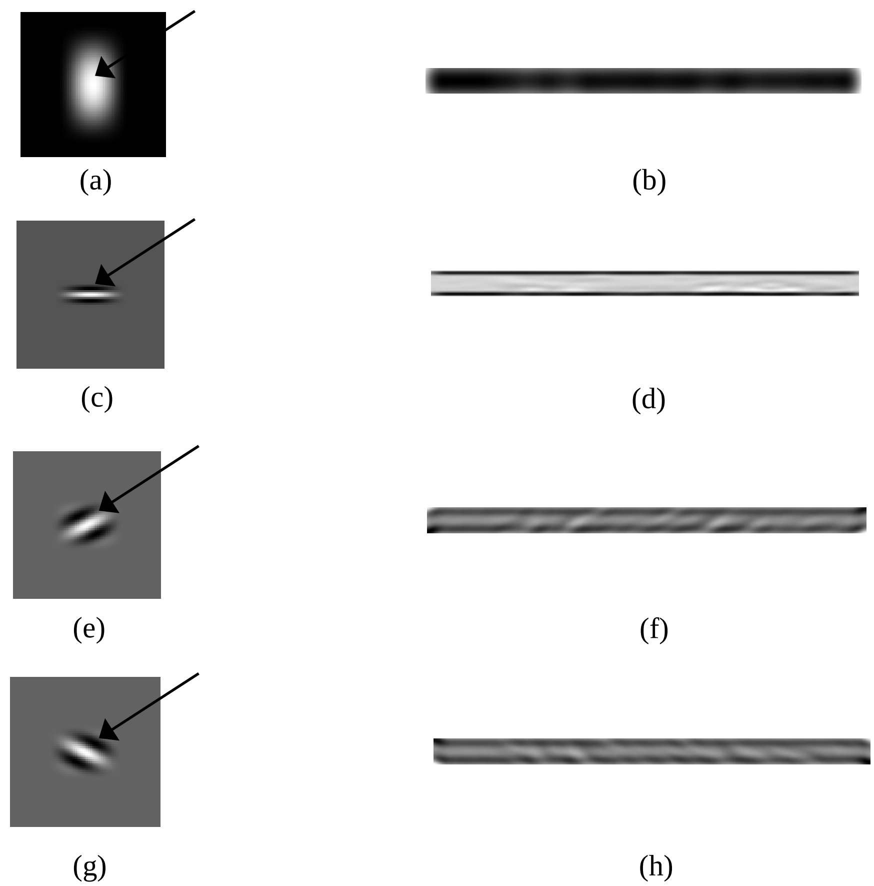


Figure 5.4: Four Gabor filters with (ω_n, ω_k) as (*0,0*), (*0,0.12*), (*3.09,-0.06*), (*3.09,0.06*) in (a), (c), (e) and (g), respectively. Corresponding output using log-mel spectrogram of speech signal *dr1_faks0_sa1_te* from TIMIT database, with white noise added at *10* dB SNR, in (b), (d), (f) and (h). Gabor filters parameters used are $v_n = v_k = 3.5$, $d_n = 0.2$ and $d_k = 0.3$.

The resultant concatenated output would be quite high-dimensional ($23 \times 41 = 943$). Such high-dimensional feature vectors would lead to high computational complexity and thus, there is a need to reduce the dimension. Dimensionality is reduced by exploiting the fact that the filter output between adjacent channels is highly correlated when the subband filter has a large spectral extent. Thus, channel selection scheme as discussed in [8] is applied to the complete feature matrix and dimensionality is reduced to *311*. Gabor filter size is limited to *40* time frames and hence, these features encode upto *400 ms* (*40* $\times$ *10 ms* window duration) context while MFCC features encode upto *45 ms* context. MFCC and GFCC features are concatenated with Gabor filterbank features, i.e., GBFB_{mel} and $\text{GBFB}_{\text{Gamm}}$, respectively (since it results in improvement in recognition performance [8]). Hence, dimension of our feature vectors is *350* (i.e., *311 +39 = 350*).

5.5 Chapter Summary

This chapter described an algorithm for the extraction of spectro-temporal Gabor filterbank features that intend to perform better than the state-of-the-art MFCC features for ASR task. Implementation with mel as well as Gammatone filterbank (for the time-frequency representation of speech), is shown. Following chapter covers the description of experiments that are performed for the ASR task using these features.

Chapter 6

Experimental Results

6.1 Introduction

Algorithm and the architecture for the Spectro-Temporal Gabor filterbank (GBFB) features have been presented in the last chapter. Algorithm incorporating both mel filterbank and Gammatone filterbank, producing $GBFB_{mel}$ and $GBFB_{Gamm}$ features, has been presented. To improve recognition performance, MFCC and GFCC features are concatenated with $GBFB_{mel}$ and $GBFB_{Gamm}$ features, respectively, to produce $GBFB_{mel}$+MFCC and $GBFB_{Gamm}$+GFCC features. Experiments are carried out on TIMIT database [9], with various additive noises for various SNR levels from *-10* dB to *20* dB, to compare the performance of the spectro-temporal features and the state-of-the-art MFCC features. Experiments are carried out with HTK [10] as back end, taking into account the effectiveness of acoustic model and language model (LM).

6.2 Experimental Setup I

Recognition experiments are conducted on TIMIT database with additive white, babble, volvo and high frequency noises at various levels of SNR ranging from *20* dB to *-10* dB. The database contains *6300* sentences, of which *4620* are used for training and *1680* are used for testing. For our experiments, training and testing environments are kept same. Hidden Markov Model (HMM) is used as the back end and phoneme-level accuracy, as given in Eq. (6.1), is used as the performance measure with one phoneme modeled by *5* states and each state modeled by mixture of *8* Gaussians. HTK is used to carry out the experiments. The % phoneme recognition accuracy (PRA) is defined as [10]:

$$\% \, PRA = \frac{N - D - S - I}{N} \times 100, \tag{6.1}$$

where N is the total number of labels (phonemes) in the reference transcriptions, S is the substitution errors, D is the deletion errors and I is the insertion errors.

6.3 Experimental Results I

A comparison between Gabor filterbank (GBFB) features concatenated with MFCC features ($GBFB_{mel}$+MFCC, dimension = *350*) and MFCC features (dimension = *39*) is shown in Figure 6.1, for additive white, babble, volvo and high frequency noises, respectively. Experiments are conducted with (i) no LM (ii) LM with weight *5.0* (iii) LM with weight *1.0*. When experimented with *5.0* and *1.0* weighted LMs, we found that MFCC features perform better than $GBFB_{mel}$+MFCC features for clean and noisy environments with SNR ranging from *20* dB to *-10* dB. For SNR = ∞ (clean conditions), *20* dB, *15* dB, *10* dB, *5* dB, *0* dB, *-5* dB, *-10* dB, MFCC features perform better (in terms of % *PRA*) by an average (computed over various SNR levels from *-10* dB to *20* dB) of *5* %, *3.5* %, *2.2* % and *4* % for LM *5.0*, for white, babble, volvo and high frequency noises, respectively. For LM *1.0*, MFCC features perform better by an average of *1.5* % for both white and babble noises and by an average of *0.7* % and *3.6* % for volvo and high frequency noises, respectively. When experimented without incorporating LM, we see that $GBFB_{mel}$+MFCC features perform better than MFCC alone by *3.6* % in clean conditions and perform better by an average of *3.5* %, *2* %, *3* % and *3.6* % in the noisy conditions, respectively.

For white noise, without LM, performance gap between $GBFB_{mel}$+MFCC features and MFCC features keeps on increasing as we go from *20* dB to *-5* dB SNR indicating higher robustness of the $GBFB_{mel}$+MFCC features for extremely low SNR conditions (*-5* dB i.e., severe signal degradation conditions). For high frequency noise, without LM, we get similar trend as white noise, with slightly larger performance gap. While for babble and volvo noises, without LM, there is a constant performance gap of about *5* % and *3* % between $GBFB_{mel}$+MFCC features and MFCC features, uptil SNR of *5* dB and *-10* dB, respectively. These results show that with acoustic modeling only, $GBFB_{mel}$+MFCC features perform better than MFCC features in clean conditions as well as in the presence of various additive noises. This is because GBFB features are able to capture more local joint *spectro-temporal* information in the speech signal (by passing log-mel spectrogram of the speech signal through overlapping subband *2-D* Gabor filters centered at several spectro-temporal modulation frequencies important for speech intelligibility), than the MFCC features (which capture more global spectral information and some temporal dynamics in the speech signal by adding delta and double-delta features). Thus, $GBFB_{mel}$+MFCC features produce more accurate acoustic models and hence, are more robust than the state-of-the-art MFCC features, even in severe signal degradation or noisy conditions (SNR -*5* dB or -*10* dB).

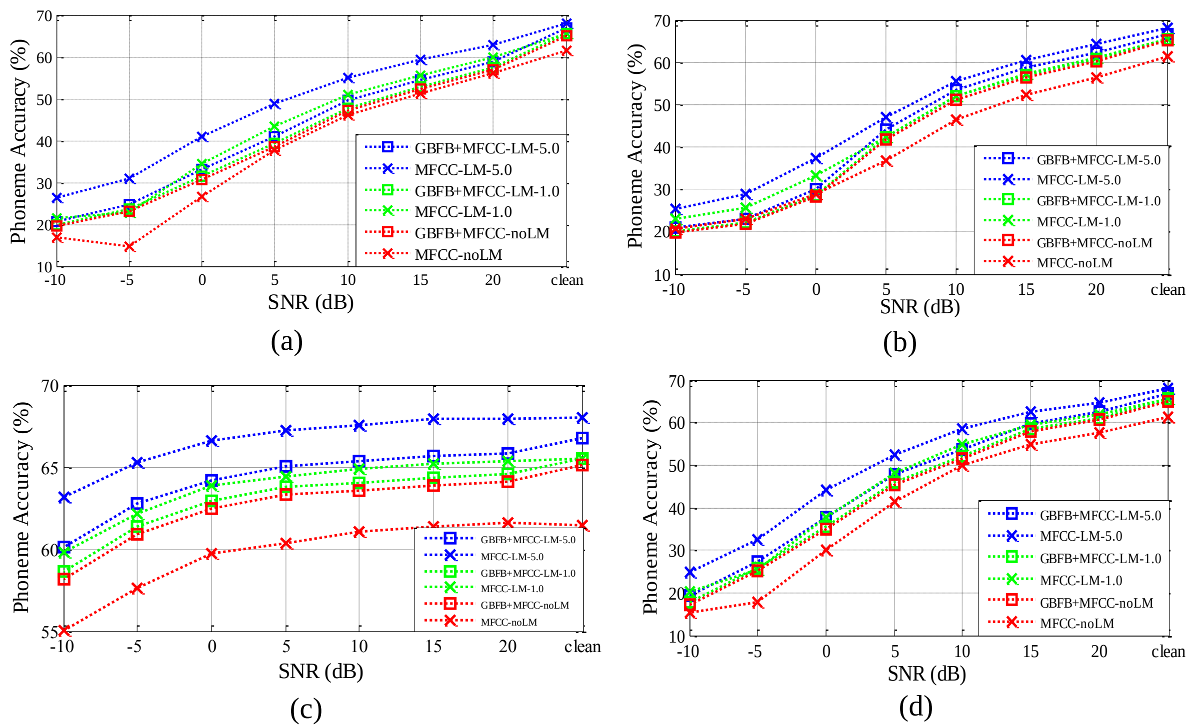


Figure 6.1: Comparison of phoneme-level accuracy (in %) between MFCC features and $GBFB_{mel}$+MFCC features for LM *5.0*, *1.0* and with no LM, for white, babble, volvo and high frequency noises in (a), (b), (c) and (d), respectively.

6.4 Experimental Setup II

Recognition experiments are conducted on TIMIT database with additive white, volvo and high frequency noises at various SNR levels ranging from *20* dB to -*5* dB. Core training sentences (*3696*) and core testing sentences (*192*) of TIMIT database are used in these experiments. % *PRA* as defined in Eq. (6.1) is used as the performance measure. Rest all the setup is same as that in Setup I.

6.5 Experimental Results II

A comparison between proposed features, i.e., $GBFB_{Gamm}$ concatenated GFCC ($GBFB_{Gamm}$ +GFCC, dimension = *350*) features, $GBFB_{mel}$ concatenated MFCC ($GBFB_{mel}$ +MFCC, dimension = *350*) features and MFCC features (dimension = *39*), for additive white, babble, volvo and high frequency noise, for various SNR levels is done and results are shown in Figure 6.2. Experiments are conducted with *5.0* weighted LM and for without LM. When experimented with *5.0* weighted LM, it is found that MFCC features perform better than the other two features for clean and noisy environments with SNR ranging from *20* dB to -*5* dB. For SNR = ∞ (clean conditions), *20* dB, *15* dB, *10* dB, *5* dB, *0* dB, -*5* dB, MFCC features perform better (in terms of % *PRA*) than the proposed features by an average (computed over various SNR levels from -*5* dB to *20* dB) of *2.6* %, *2.7* %, *1* % and *2.3* % and perform better than $GBFB_{mel}$+MFCC by an average of *3.5* %, *2.3* %, *1.2* % and *2.5* % , for additive white, volvo and high frequency noise, respectively. Thus, with *5.0* LM, the proposed features perform better than $GBFB_{mel}$+MFCC by an average of *0.9* % for white noise and *0.2* % for volvo and high frequency noises. When experimented without incorporating LM, it is seen that the proposed features outperform both MFCC and $GBFB_{mel}$+MFCC under signal degradation or noisy conditions, for additive white, volvo and high frequency (HF) noises. For signal degradation conditions, the proposed features perform better than MFCC by an average of *4.6* %, *3.5* % and *5.4* % and perform better than $GBFB_{mel}$+MFCC by an average of *1* %, *0.2* % and *0.8* % for white, volvo and high frequency noise, respectively. Under clean conditions, without LM, the proposed features perform almost similar to $GBFB_{mel}$+MFCC features but perform better than MFCC features by *3.7* %. For babble noise, the proposed features perform better than MFCC by an average of *2.9* % but are outperformed by $GBFB_{mel}$+MFCC by an average of *0.6* %. It can be observed that, with acoustic modeling only, spectro-temporal Gabor filterbank (GBFB) features (whether incorporating Gammatone filterbank or mel filterbank) when concatenated with cepstral coefficients perform better than the state-of-the-art MFCC features in clean conditions as well as in the presence of various additive noises. This is because GBFB features are able to capture more local joint spectro-temporal information in the speech signal. In addition, when Gammatone filterbank is used instead of mel filterbank, to extract GBFB features, the recognition performance under signal degradation or noisy conditions (SNR ranging from *20* dB to -*5* dB), is improved.

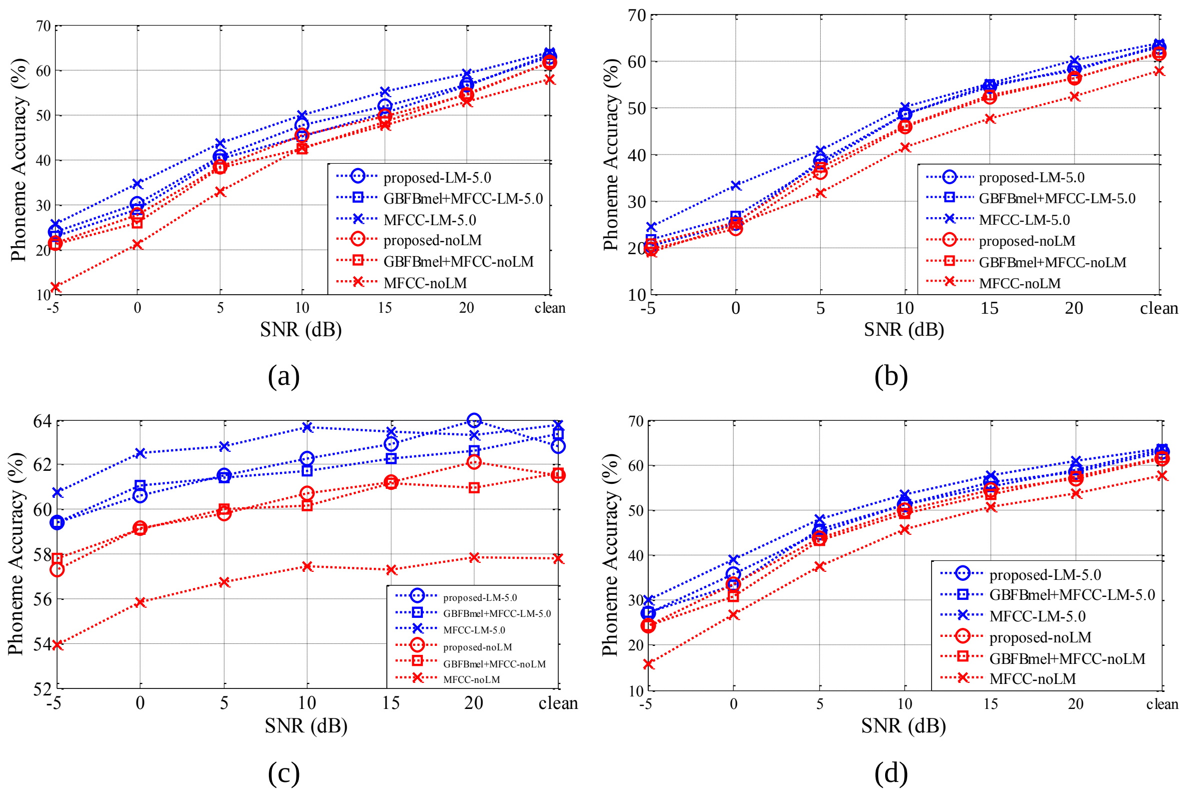


Figure 6.2: Comparison of phoneme-level accuracy (in %) between the proposed features, $GBFB_{mel}$+MFCC features and MFCC features for LM *5.0* and with no LM, for additive white, babble, volvo and high frequency noises at various SNR levels in (a), (b) and (c), respectively.

6.6 Discussions on Results

The state-of-the-art MFCC features capture spectral information from the speech signal, and capture some temporal dynamics when concatenated with their temporal derivatives (i.e., delta and double-delta features). The spectro-temporal Gabor filterbank features are inspired by the process of human speech recognition and are thus, able to capture more local joint spectro-temporal information in the speech signal. Experimental results show that, with acoustic modeling only, the spectro-temporal GBFB features when concatenated with cepstral coefficients perform better than the state-of-the-art MFCC features because of the fact that GBFB features are able to capture more local joint spectro-temporal information in the speech signal (by passing spectrogram of speech through various *2-D* Gabor subband filters aligned at modulation frequencies important for speech *intelligibility*). Thus, spectro-temporal features are preferred for the languages/databases which do not have enough accurate language models (due to scarcity of training data). When Gammatone filterbank is used instead of the standard mel filterbank, the recognition performance of the spectro-temporal features is improved.

6.7 Chapter Summary

This chapter presented all the experiments that have been carried out for the ASR task. The performance of spectro-temporal features ($GBFB_{mel}$+MFCC and $GBFB_{Gamm}$+GFCC features) is compared with that of the state-of-the-art MFCC features, for various additive noises, under various SNR levels. Furthermore, an analysis on the results is done. In the next chapter, results presented in this thesis are summarized along with the conclusions and future research directions.

Chapter 7

Conclusions and Future Works

7.1 Summary and Conclusions

ASR technology, is now, being used in almost every field in life, from academics to healthcare to military. Deploying ASR technology in today's world incorporates the challenge to deal with various background noises and channel effects. Efficiency of the ASR cannot be compromised and thus, there is a strong need to have features that are able to perform good enough in the those signal degradation or noisy conditions. The state-of-the-art MFCC features are known to be affected by acoustic noise. Physiologically motivated spectro-temporal features intend to perform better in noisy conditions as they capture additional information, local joint spectro-temporal information, from the speech signal. In this thesis, the state-of-the-art MFCC features and its Gammatone counterpart GFCC features, along with their limitations have been discussed. This is followed by the motivation to use spectro-temporal features in ASR and incorporation of Gabor filters in feature extraction for ASR. STRF of neurons in *A1* of mammalians' brains look like *2-D* Gabor function and hence, neurons can be modeled by *2-D* Gabor filters. Since Gabor filters are physiologically motivated, they are used in the extraction of features, which would then capture more relevant information from the speech signal resulting in higher recognition accuracy. A method, described in [8], [25] and [27], to extract spectro-temporal features, named "Spectro-Temporal Gabor filterbank (GBFB) features", is discussed. The algorithm takes the log-mel/Gammatone spectrogram of the speech signal and passes it through a bank of *2-D* Gabor filters (that act as impulse response of the neuron in *A1*) to generate time-frequency representations which inturn form the feature vectors. Cepstral coefficients (MFCC and GFCC) are concatenated with the GBFB features to increase recognition performance. Experiments are carried out on TIMIT database, with various additive noises, for various SNR levels, to compare the proposed spectro-temporal features and the state-of-the-art MFCC features. Results show that with acoustic modeling only, spectro-temporal features perform better than the MFCC features, for clean as well as noisy conditions. In addition, when Gammatone filterbank is incorporated instead of the standard mel filterbank to extract spectro-temporal features, recognition performance increases. Spectro-temporal features are thus able to build more accurate acoustic models and hence, are preferred for the languages/databases which do not have enough accurate language models (due to scarcity of training data).

7.2 Limitations of Present Work

- Algorithm for the extraction of spectro-temporal Gabor filterbank features is computationally complex as it involves computation of log-mel/Gammatone spectrogram followed by several convolutions with the Gabor filters followed by dimensionality reduction using channel selection scheme. This results in higher time complexity, than the state-of-the-art MFCC features, in the extraction of the features.
- Spectro-Temporal Gabor filterbank features are quite high dimensional (dimension =*350*). Hence, it takes quite higher time than MFCC features, to train HTK and further, to decode.

7.3 Future Research Directions

- GBFB features are quite high-dimensional as compared to the standard MFCC features. Hence, future research work will be to reduce the dimensionality of GBFB features without diminishing the actual strength of the features, i.e., to capture local joint spectro-temporal information in the speech signal.
- In reality, neurons in the *A1* are data adaptive. Thus, future work could also be to learn the Gabor filters using any machine learning technique (Restricted Boltzmann Machine (RBM) or Deep Neural Networks (DNN)), instead of using static filters, to extract spectro-temporal features from speech.

Appendix

HTK System Building

A.1 Introduction

This supplementary material describes the main steps for building ASR system in Hidden Markov Model toolkit (HTK). Installation of the toolkit and then using it to build the complete ASR system will be discussed.

A.2 Software Installation and Setup

To install HTK, HTK.tar.gz and HTK_samples.tar.gz (versions may change) are to be downloaded for which voxforge website can be referred to [28]. Box A.1 shows the stepwise installation of the toolkit in Ubuntu. For windows, [28] can be referred.

Box A.1: Stepwise installation of HTK in Ubuntu

From the downloaded folder, run:

- tar -xvzf HTK.tar.gz
- tar -xvzf HTK_samples.tar.gz

From the extracted HTK folder, run:

- ./configure -prefix=/home/username/bin/htk
- make all
- make install

Create another folder and copy the following files to the created folder

1. mkclscript.prl (from /htk/samples/RMHTK/perl_scripts)
2. prompts2mlf (from /htk/samples/HTKTutorial)

3. prompts2wlist (from /htk/samples/HTKTutorial)

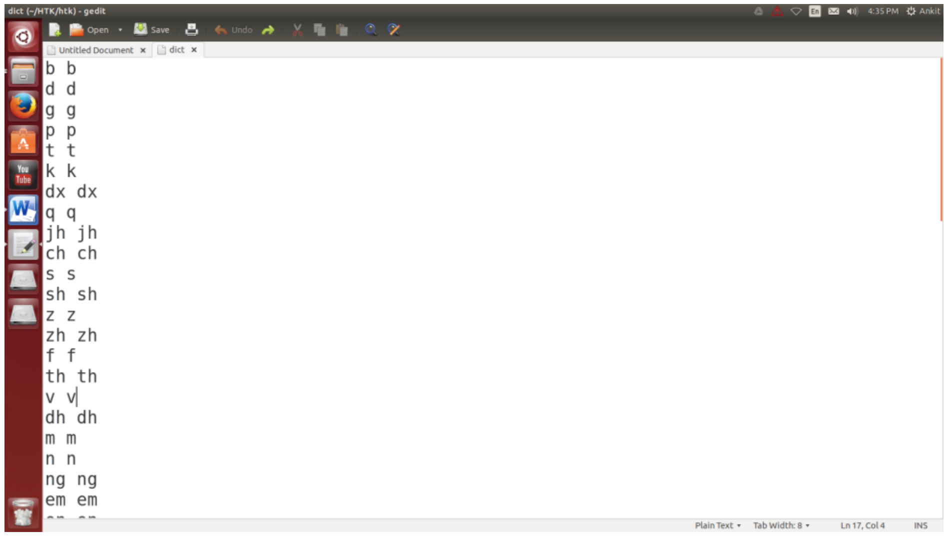
A.3 Files required to Build ASR System

1. **prompts**: It contains text representation (phoneme-level) of all the utterances in the database. Left column of the file contains utterance id and right column contains the text representation of the utterance. Example is shown below. Any scripting language can be used to construct the file.

Box A.2: Exemplar prompts file

*/dr1_fcjf0_sa1_tr.lab h# sh ix hv eh dcl jh ih dcl d ah kcl k s ux q en gcl g r ix s ix w ao sh epi w ao dx axr ao l y ih axr h#
*/dr1_fcjf0_sa2_tr.lab h# d uh n ae s kcl k m ix dx ix kcl k eh r ix n oy l ax r ae gcl g oy kcl dh ae tcl h#

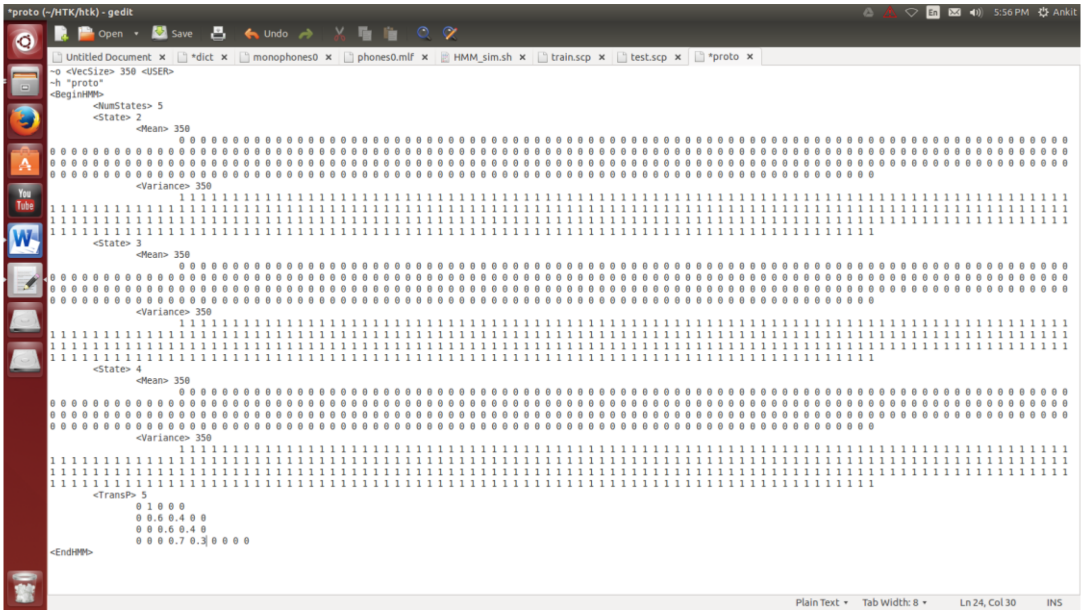
2. **monophones0:** It contains all the words, i.e., phomemes in the database. HTK script prompts2wlist is used to construct the same. Words would be listed in the sorted order.
   - **perl prompts2wlist prompts monophones0**
3. **dict:** It is the phoneme-level pronunciation dictionary also known as the lexicon. It maps phonemes to phonemes. !ENTER [] and !EXIT [] are to be entered at the end of the file. These are HTK internal entries used for creating Acoustic Models.



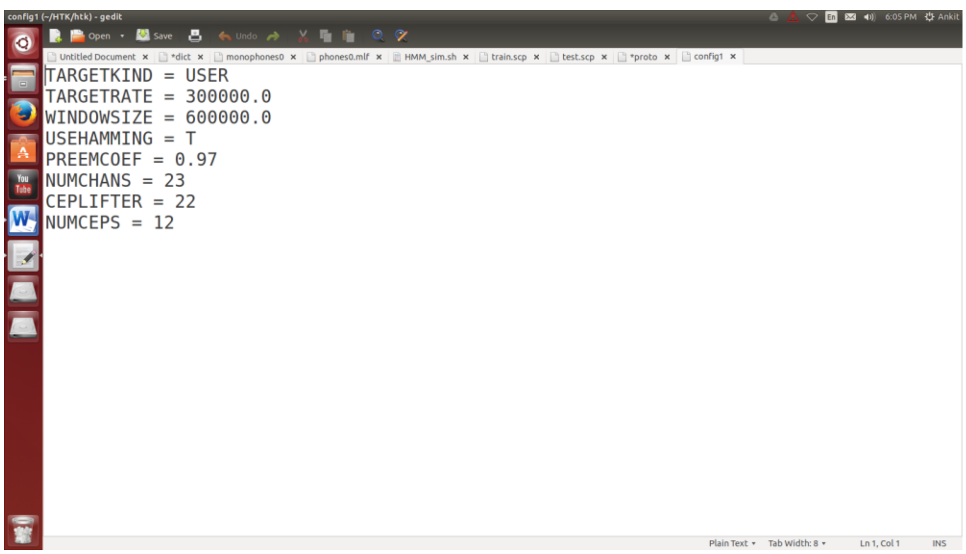
4. **phones0.mlf:** It is the Master Label File (MLF) that contains file name followed by one phoneme in each row. prompts2mlf script is used to construct it.
    - **perl prompts2mlf phones0.mlf prompts**
5. **train.scp:** File containing paths of all the files for training the ASR system.
6. **test.scp:** File containing paths of all the testing files.
7. **testref.mlf:** MLF for testing files only. It is used for testing the results.

A.4 Training of System

1. Create a prototype model called “**proto**” to create a structure. The file defines the number of states (in this work, 5 states are defined) to model a phoneme along with the state transition probabilities. It creates a structure for mean and variance of Gaussians in a state. Any scripting language such as perl can be used to create it.



2. Create a configuration file "**config**". This file is used to configure the features used for ASR.



TARGETKIND field contains the type of features. If it is USER, it means user is providing his features for building the system. HTK can also compute features within itself. Providing MFCC_0_D_N to the TARGETKIND field will make HTK to compute MFCC features.

3. Now, construct a proto file in another folder name **hmm0** by using HCompV command.

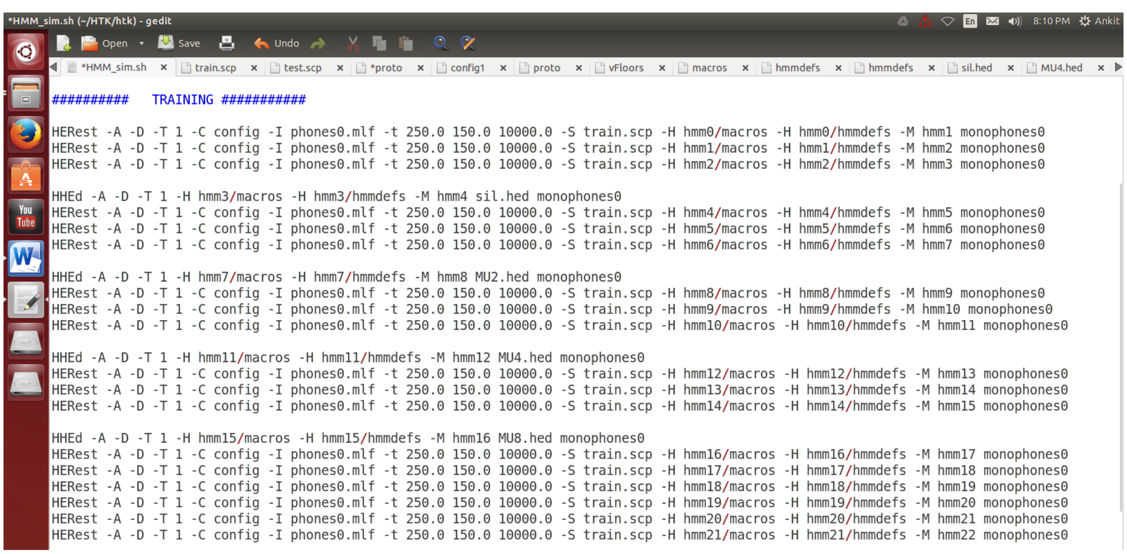
- HCompV -A -D -T 1 -C config -f 0.01 -m -S train.scp -M hmm0 proto

This will create two files, **proto** that contains global mean and variance (by scanning all the training files) and sets them for all the Gaussians, and **vFloors** containing floor of the variance which has the value *0.01* times the global variance.

4. Create **hmmdefs** using monophones0 file.
   i. Put the phonemes in double quotes.
   ii. Add '~h' before each phoneme.
   iii. Copy from line 5 onwards from the proto file in hmm0 and paste it after every phoneme.

   Thus, hmmdefs contains flat-start monophones.

5. Create macros using vFloors.
   i. Create a new file called macros in hmm0.
   ii. Copy vFloors to macros.
   iii. Copy first 3 lines of proto and add them to the top of the macros file.
6. Create some more hmm folders for the **re-estimation** of monophones. In this work, 22 folders are created and HERest command is used for re-estimation which uses Baum-Welch Re-estimation algorithm. Details of Baum-Welch Re-estimation algorithm can be found in [10]. Below snapshot shows the re-estimation script implemented.



-t option provides the beam width for re-estimation to avoid pruning of training data. Beam width is initialized to 250.0, it increments with 150.0 and ends at 1000.0. **sil.hed** is used to add extra transitions from state 2 to state 4. Contents of sil.hed are as shown.

Box A.3: File sil.hed

```
AT 2 4 0.2 {sil.transP}

AT 4 2 0.2 {sil.transP}

AT 1 3 0.3 {sp.transP}

TI silst {sil.state[3],sp.state[2]}
```

MU2.hed, **MU4.hed**, **MU8.hed** are used to model 2, 4 and 8 Gaussians, respectively, in the mixture model of a state. MU2.hed contains:

```
MU 2 {*.state[2-4].mix}
```

hmm22 folder contains the final HMM models for the phonemes.

A.5 Decoding of Data

Data is decoded using HVite command which uses Viterbi Algorithm to decode the observation sequence.

Box A.4: Commands to decode observation sequence

- HVite -A -D -T 1 -H hmm22/macros -H hmm22/hmmdefs -C config -S test.scp -l '*' -i recout.mlf -w wdnet -p 0.0 -s 5.0 dict monophones0
- HVite -A -D -T 1 -H hmm22/macros -H hmm22/hmmdefs -S test.scp -i recout_bigram.mlf -w outLatFile -p 0.0 -s 5.0 dict monophones0

wdnet is the word network and outLatFile file defines the bigram language model (LM). Further details on wdnet can be found in [10]. *5.0* denotes the weight of the LM. recout.mlf will store the output decoded through acoustic modeling only and recout_bigram.mlf will store decoded output when LM is also incorporated.

A.6 Testing the Decoded Output

For testing, HResults command is used. Script given below is used to get the testing results.

Box A.5: Commands to get testing results

- HResults -I testref.mlf monophones0 recout.mlf
- HResults -T 1 -e n en -e aa ao -e ah ax-h -e ah ax -e ih ix -e l el -e sh zh -e uw ux -e er axr -e m em -e n nx -e ng eng -e hh hv -e pau pcl -e pau tcl -e pau kcl -e pau q -e pau pcl -e pau dcl -e pau gcl -e pau epi -e pau h# -I testref.mlf monophones0 recout.mlf % mapping *61* phonemes to *39* phonemes
- HResults -T 1 -I testref.mlf monophones0 recout_bigram.mlf
- HResults -T 1 -e n en -e aa ao -e ah ax-h -e ah ax -e ih ix -e l el -e sh zh -e uw ux -e er axr -e m em -e n nx -e ng eng -e hh hv -e pau pcl -e pau tcl -e pau kcl -e pau q -e pau pcl -e pau dcl -e pau gcl -e pau epi -e pau h# -I testref.mlf monophones0 recout_bigram.mlf

The above files compare the output files recout.mlf/ recout_bigram.mlf with the testref.mlf file and displays the result (% Phoneme Recognition Accuracy).

References

[1] R. O. Duda, P. E. Hart and G. D. Stork, Pattern Classfication, John Wiley, 2002.

[2] http://en.wikipedia.org/wiki/Speech_recognition. {Last Accessed: 7th May, 2015}.

[3] https://support.google.com/chat/answer/3379791. {Last Accessed 7th May, 2015}.

[4] H. Hermanksy and S. Sharma, “Temporal patterns (TRAPS) in ASR of noisy speech,” in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, Arizona, USA, vol. 1, pp. 289-292, 1999.

[5] S. David and P. Mermelstein, “Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoustics, Speech and Signal Process.*, vol. 28, no. 4, p. 357–366, 1980.

[6] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[7] T. Chi, Y. Gao, M. C. Gutyon, P. Ru and S. A. Shamma, “Spectro-temporal modulation transfer functions and speech intelligibility,” *The Journal of the Acoustical Society of America*, vol. 106, no. 5, pp. 2719-2732, 1999.

[8] M. Schadler, B. Meyer and B. Kollmeier, “Spectro-temporal modulation subspace spanning filter bank features for robust automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. 4134-51, 2012.

[9] K. Lee and H. Hon, “Speaker-independent phone recognition using hidden Markov Models,” *IEEE Trans. Acoustics, Speech and Signal Process*, vol. 37, no. 11, pp. 1642-1648, 1989.

[10] S. Young, G. Evermann, M. J. F. Gales and e. al., The HTK book for HTK version

3.4., Microsoft Corporation, 2006.

[11] D. A. Depireux, P. Ru, S. A. Shamma and J. Z. Simon, "Response-field dynamics in the auditory pathway," Computational Neuroscience, pp. 1-6.

[12] T. Chi, P. Ru and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America,* vol. 118, no. 2, pp. 887-906, 2005.

[13] http://www.babyhearing.org/hearingamplification/hearingloss/earworks.asp. {Last Accessed: 11th May, 2015}.

[14] T. Ezzat, J. Bouvrie and T. Poggio, "Spectro-temporal analysis of speech using 2-d Gabor filters," *INTERSPEECH,* Antwerp, Belgium, pp. 506-509, 2007.

[15] S. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," *INTERSPEECH,* Brisbane, Australia, pp. 898-901, 2008.

[16] H. Hermansky, D. Ellis and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," *ICASSP*, Istanbul, Turkey, vol. 3, pp. 1635-1638 , 2000.

[17] Y. Shao, Z. Jin, D. Wang and S. Srinivasan, "An auditory-based feature for robust speech recognition," in *ICASSP*, Taipei, Taiwan, pp. 4625–4628, 2009.

[18] T. F. Quatieri, Discrete-Time Speech Signal Processing, New Jersey: Prentice-Hall, 2001.

[19] M. Slaney, "Auditory Toolbox, version 2," https://engineering.purdue.edu/~malcolm/interval/1998-010/. [Last Accessed: 7th May, 2015].

[20] http://en.wikipedia.org/wiki/Mel_scale. {Last Accessed: 9th May, 2015}.

[21] http://en.wikipedia.org/wiki/Equivalent_rectangular_bandwidth. {Last Accessed: 11th May, 2015}.

[22] B. Meyer and B. Kollmeier, "Optimization and evaluation of Gabor feature sets for ASR," *INTERSPEECH*, Brisbane, Australia, vol. 2, 2008.

[23] J. R. Movellan, "Tutorial on Gabor Filters," 2008.

[24] A. Qiu, C. E. Schreiner and M. A. Escabı, "Gabor Analysis of Auditory Midbrain Receptive Fields: Spectro-Temporal," *Journal of Neurophysiology*, vol. 90, no. 1, p. 456–476, 2003.

[25] A. Nagpal and H. A. Patil, "Spectro-Temporal Gabor Filterbank-Based Features for Robust Phoneme Recognition," *submitted in the 16th Annual Conference of (ISCA), INTERSPEECH*, Dresden, Germany, 2015.

[26] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 578-589, 1994.

[27] A. Nagpal and H. A. Patil, "Novel Gammatone Filterbank based SpectroTemporal Features for Robust Phoneme Recognition," *submitted in 18th Int. Conf. on Text, Speech and Dialogue (TSD)*, Plzen, Czech Republic, 2015.

[28] http://www.voxforge.org/home/dev/acousticmodels/linux/create/htkjulius/tutorial/.{Last Accessed: 12th May, 2015}.

Publications from the Thesis

1. Ankit Nagpal and Hemant A. Patil, "Novel Gammatone Filterbank based Spectro-Temporal Features for Robust Phoneme Recognition," accepted in 18th Int. Conf. on Text, Speech and Dialogue (TSD), Lecture Notes in Artificial Intelligence (LNAI), Springer, Plzen, Czech Republic, Sept. 14–17, 2015.
2. Ankit Nagpal and Hemant A. Patil, "Spectro-Temporal Gabor Filterbank-Based Features for Robust Phoneme Recognition," submitted in the 16th Annual Conference of (ISCA), INTERSPEECH, Dresden, Germany, 2015.