**Speech Knowledge based Broadcast Audio Classification and Phone Recognition**




***BANRISKHEM K KHONGLAH***

# Speech Knowledge based Broadcast Audio Classification and Phone Recognition

*A*

*Thesis submitted*

*for the award of the degree of*

**DOCTOR OF PHILOSOPHY**

By

**BANRISKHEM K KHONGLAH**


DEPARTMENT OF ELECTRONICS AND ELECTRICAL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

GUWAHATI - 781 039, ASSAM, INDIA

Oct 2017

# Certificate

This is to certify that the thesis entitled "**Speech Knowledge based Broadcast Audio Classification and Phone Recognition**", submitted by **BANRISKHEM K KHONGLAH** (126102033), a research scholar in the *Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati*, for the award of the degree of **Doctor of Philosophy**, is a record of an original research work carried out by him under my supervision and guidance. The thesis has fulfilled all requirements as per the regulations of the institute and in my opinion, has reached the standard needed for submission. The results embodied in this thesis have not been submitted to any other University or Institute for the award of any degree or diploma.

Dated:

Guwahati.

Prof. S. R. Mahadeva Prasanna
Professor
Dept. of Electronics and Electrical Engg.
Indian Institute of Technology Guwahati
Guwahati - 781 039, Assam, India.

To

My **Parents**

for their blessings, love and support

# Acknowledgements

First of all, I would like to thank God for keeping me in good health and giving me strength and patience to work hard. This thesis would not have been possible without the immense help and support of several people in various measures. I would like to convey my acknowledgement to all of them. I attribute this achievement to my parents for their support in every aspect of my life. They worked really hard and sacrifice a lot in order to make sure that I come to this stage. Their motivations from time to time help me get through most of the difficulties that come my way.

I express my sincere gratitude to my research supervisor, Prof. S. R. M. Prasanna for providing me an opportunity to work under his guidance. It is very difficult to describe my feelings in words to acknowledge my supervisor for his continuous guidance in all aspects, constant motivation and support throughout the doctoral studies. I am thankful to him for providing me the Multimedia Analytics Laboratory for my research related work. This lab has been special to me and I have really enjoyed working in this lab. The memories of this lab will stay forever. His implementation of the weekly meetings to discuss our weekly progress has really helped the progress of my research. I disliked these weekly meetings but now I realized that without them I would not have made so much progress in my work. I would like to thank him for this and hope that he will not mind if I follow this idea of weekly meetings for my future purposes. He would always acknowledge any request made from my side and was ready to help in any way. His hard-working nature has been a constant motivation for me.

I am thankful to my doctoral committee members Prof. P.K. Bora, Dr. P. Guha and Dr. Ranbir Singh for their encouragement and valuable suggestions on my work. I would also like to thank Prof. S. Dandapat and Prof. Rohit Sinha for their critical comments and suggestions for my work. I would like to thank faculty members and the office staffs of the Department of Electronics and Electrical Engineering, IIT Guwahati, for their help in carrying out this research work. I would like to acknowledge E-Security Division, Department of Information technology, New Delhi for providing enormous funding to build an advanced computational facility in the Multimedia Analytics Laboratory.

I am thankful to my friends Nagaraj, Dr. Biswajit, Dr. Deepak, Vikram, Abhishek, Vivek, Matthew and Rajib for their assistance in my work. Special thanks to Raghvendra for his help, support and useful ideas for my work. I would also like to thank Udeshna for her assistance in the data preparation for my thesis related experiments. I am thankful to Syed, Haris, Jiss, Rohan,

Ramesh, Alex, Ato, Bhukya, Subhasis, Bidisha, Himakshi, Sishir, Akhilesh, Protima, Shikha, Moakala, Sarfaraz, Mrinmoy, Padhy, Anurag, Suman, Bhanu Priya, Tilendra, Ganji, Saswati, Sandeep and the rest that I have forgotten to mention, who have always been there for me someway or the other.

*BANRISKHEM K KHONGLAH*

# Abstract

This thesis describes the processing of speech data in multimedia context such as in broadcast audio. The broadcast audio data contains speech present in a variety of scenarios and these scenarios include the clean speech, speech with background music/noise and pure music. The speech with background music and pure music segments correspond to the voice-over cases and news headlines. The speech with background noise is due to the reporter speech. The main goal of this work is phone recognition of broadcast audio. However, some form of classification and enhancement is required considering the complexity of the broadcast audio data. In this work, the data mostly related to the anchor speakers' segments is considered and hence the scenarios such as clean speech, speech with background music and pure music has to be taken care in the classification and enhancement stages. The speech with background noise regions are not considered, since they are mostly present in the form of the reporter speech and recorded in outdoor environments.

The speech/music classification task is the first classification task attempted in this thesis. There have been several attempts in previous work on this task which uses the general temporal and spectral features. In this thesis, the task is attempted by exploring the speech-specific features. Defining features related to music is a difficult one considering the different kinds of music present. Alternatively, since the mechanism of speech production is uniform (speech is produced by exciting a time-varying vocal tract system by a time-varying excitation source), the features specific to speech can be explored. The speech-specific features related to the source, vocal tract system and syllable rate of speech are expected to behave differently in the music regions compared to speech. Since they are features derived in the context of speech production, they are able to characterize the speech regions better, thus achieving some form of discrimination between speech and music.

The speech regions thus obtained may be either clean speech or speech with background music considering the speech-specific nature of the features. The second level of classification involves classifying clean speech from the speech with background music segments. The reason for this step is because the phone recognizer to be used in this thesis is a single one trained on clean speech. As a result of this, the speech with background music may need to be separated (from the clean speech) and enhanced, prior to passing it through the phone recognizer to reduce the acoustic mismatch between the train and test speech data. The task of classifying clean speech from the speech with background music is performed by considering the average and relative spectral characteristics of the vocal tract system, which is again in terms of the speech-specific nature. The resonance features of the vocal tract system are exploited to achieve a discrimination between the mentioned regions. The presence of music in the background of speech will introduce frequency components due to the music in addition to speech, thus causing the relative spectral characteristics of the vocal tract system to be different between the clean speech and speech with background music regions.

The speech with background music is then passed through an enhancement module which consists of the temporal, spectral and perceptual enhancement systems. The temporal enhancement step involves emphasizing the high speech to music ratio (SMR) regions of the source signal and the perceptual enhancement step involves having a clean source signal representation as the excitation component. Based on the requirements of a robust source representation for enhancing the speech with background music, this thesis proposes a better representation of the source signal (a speech-specific information) which emphasizes the high SMR regions while de-emphasizes the other signal components for the temporal and perceptual enhancement steps.

The clean speech obtained from the clean speech/speech with background music classification step, and the enhanced speech obtained by enhancing the speech with background music, are next passed through a phone recognition system to obtain the phone transcription of the broadcast audio. The final goal of the task is to obtain an overall improved phone recognition accuracy compared to the case when the broadcast audio is passed directly through the phone recognizer.

The major contributions of this thesis are as follows:

- Speech/Music classification using speech-specific features in terms of source, vocal tract system, and syllable rate of speech production. The task was evaluated on the Scheirer and Slaney (S&S) database, GTZAN database and the Broadcast news audio database. The speech-specific features perform better than the existing temporal and spectral features since they are able to localize the speech regions better and deviate from their speech behavior in the music regions.
- Clean Speech/Speech with background music classification using average and relative spectral characteristics of the vocal tract system. This task was evaluated on the S&S and the Broadcast news audio database. The features derived in terms of the average and relative spectral characteristics of the vocal tract system complement each other in an effective way to provide good classification accuracy for the task.
- Enhancement of Speech with background music using temporal, spectral and perceptual processing exploiting the effect of source information. The database used for the evaluation consisted of the TIMIT samples added with the background music samples taken from the S&S and the GTZAN database. The task was also evaluated on the speech with background music samples from the Broadcast news audio database. The enhanced speech samples are passed through a phone recognition system and it is observed that the algorithm works well for the music degradations which are heavy in nature.
- The samples which contain clean speech, music and speech with background music are then passed through the various preprocessing steps. The output of the preprocessing steps which is the clean and enhanced speech, are passed through the phone recognition system to show the significance of the preprocessing stages. The samples were taken from the TIMIT database added with the background music samples (taken from the S&S and the GTZAN database) and the Broadcast audio database. It was observed that the preprocessing stages help in improving the accuracy of the phone recognition system.

**Keywords:** speech-specific, classification, speech enhancement, phone recognition

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| BN | Broadcast News |
| CCA | Canonical Correlation Analysis |
| COG | Center of Gravity |
| DC | Direct Current |
| DCT | Discrete Cosine Transform |
| DEGG | Differential Electroglottograph |
| DFT | Discrete Fourier Transform |
| DNN | Deep Neural Networks |
| DP | Dynamic Programming |
| EVFB | Energy Variance of Filter Banks |
| GCI | Glottal Closure Instants |
| GMM | Gaussian Mixture Models |
| HE | Hilbert Envelope |
| HMM | Hidden Markov Models |
| HNGD | Hilbert Envelope of Numerator of Group Delay |
| IDBM | Ideal Binary Masking |
| IR | Information Retrieval |
| KLT | Karhuen-Loeve Transform |
| KNN | K-Nearest Neighbor |
| LP | Linear Prediction |
| LPC | Linear Prediction Coefficients |
| LSA | Log Spectral Amplitude |
| MCCs | Mel Cepstral Coefficients |
| MFCC | Mel Frequency Cepstral Coefficients |

| | |
|---|---|
| MLP | Multilayer Perceptron |
| MLSA | Mel Log Spectral Approximation |
| MMSE | Minimum Mean Square Error |
| NAPS | Normalized Autocorrelation Peak Strength |
| NGD | Numerator of Group Delay |
| NSS | Non-linear Spectral Subtraction |
| OLA | Overlap Add |
| PER | Phone Error Rate |
| PLP | Perceptual Linear Prediction |
| PSR | Peak to Side Lobe Ratio |
| RBF | Radial Basis Function |
| SCD | Speaker Change Detection |
| SDCFG | stochastic dependency context free grammar |
| SDR | Spoken Document Retrieval |
| SFF | Single Frequency Filter |
| SGD | Stochastic Gradient Descent |
| SGMM | Subspace Gaussian Mixture Model |
| SMR | Speech to Music Ratio |
| SNR | Signal to Noise Ratio |
| SoE | Strength of Excitation |
| S&S | Scheirer and Slaney |
| SSEG | Story Segmentation |
| SSU | Story Summarization |
| STFT | Short Time Fourier Transform |
| STSA | Short Time Spectral Amplitude |
| SC | Spectral Contrast |
| SP | Spectral Peak |
| SV | Spectral Valley |
| SVD | Singular Value Decomposition |
| SVM | Support Vector Machines |

| | |
|---|---|
| UBM | Universal Background Model |
| VLRs | Vowel Like Regions |
| ZBF | Zero Band Filter |
| ZBFS | Zero Band Filtered Signal |
| ZCR | Zero Crossing Rate |
| ZFF | Zero Frequency Filter |
| ZFFS | Zero Frequency Filtered Signal |

# 1

# Introduction

Contents

*In this thesis, the significance of speech-specific knowledge for processing broadcast audio is demonstrated. The main focus is on the anchor speakers' segments present in broadcast audio. The scenarios present in the anchor speakers' segments include clean speech, speech with background music and pure music. The main goal is to obtain the phone transcription of the anchor speakers' segments. However, due to the complexity of the nature of the data due to the scenarios present, some preprocessing is necessary. Accordingly, the preprocessing task attempted in this work include the speech/music classification, clean speech/speech with background music classification and enhancement of speech with background music. The final task is the phone recognition task. The speech/music classification task involves the use of speech-specific features for segmenting the speech and music. Similarly, the clean speech/speech with background classification task involves the use of the features based on speech production in terms of the vocal tract system. The enhancement of speech with background music is performed by emphasizing the speech components without considering the background present.*

## 1.1 Introduction to Broadcast Audio Processing

Broadcast audio processing has been attempted in several ways in the past and is involved in different applications. Some of these include the spoken document retrieval [2], [3], speech summarization [4] and story summarization [5]. The transcription of the data in terms of the word or phoneme is obtained which is then used for further processing in most of these work. Hence the fundamental task of broadcast audio processing is the transcription of the broadcast audio. This involves either the signal to symbol transformation or the symbol-to-text transformation. In this thesis, the main focus is on the signal-to-symbol transformation or the phone transcription of the broadcast audio.

The process of phone transcription of speech is a relatively easier task, if a large amount of data is collected from a variety of speakers for training and the testing data consists of only clean speech. However, the phone transcription of data collected from TV broadcast audio, in particular, the ones from Indian news channels is a much more difficult task due to the presence of different types of scenarios which are complex in nature. These include clean speech which is mostly uttered by the anchor speaker, speech with background noise uttered by the reporter and pure music which may be played in advertisements and news headlines. The speech with background music constitutes the news headlines, advertisements, and voice over speech. An example audio of the scenarios can be seen in Figure 1.1. In order to obtain a phone transcription of the broadcast audio, some form of

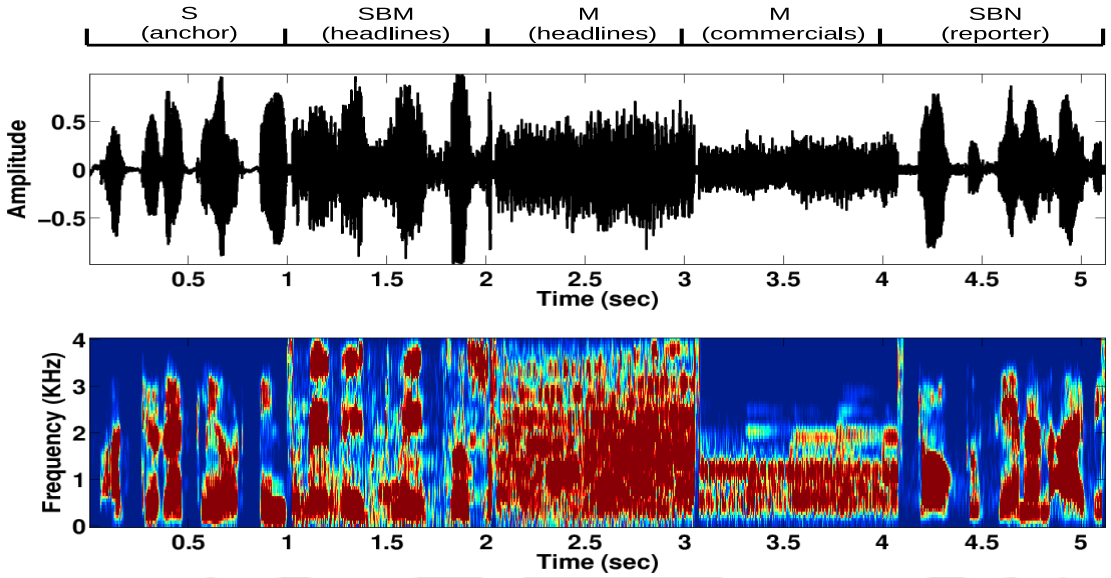


**Figure 1.1:** *(a) Audio Signal where the first 1 s correspond to clean speech related to anchor speaker, followed by 1 s of speech with background music related to news headlines, also followed by 1 s of music each related to headlines and commercials and final last 1 s of speech with background noise related to reporter. (b) Spectrogram of the audio signal*

preprocessing may be required prior to the phone recognition.

### 1.1.1 Preprocessing for Phone Transcription of Broadcast Audio

In previous work, the automatic transcription of broadcast audio follows the steps as shown in Figure 1.2, where initially the non-speech segments of the audio are filtered out. This is followed by iterative segmentation and labeling into different models based on gender and bandwidth. The final step is the speech recognition step. There are also works where the preprocessing stage includes speaker change detection [6, 7] and speaker clustering [6, 8, 9]. In such cases, the audio file obtained after preprocessing is passed either through the speaker adaptation module or the speech recognition module to obtain the textual transcription. The speech with background music and speech with background noises are processed in the speaker clustering stage.

In earlier work, the features for either the preprocessing or the recognition stage have been mel frequency cepstral coefficients (MFCC) [7, 8, 10]. Some have also explored features based on Linear Prediction Coding (LPC) [6, 7]. The general time and frequency domain features have also been used in some work for the classification stage [11–14]. In most of the earlier work, the focus has been on the general audio features with the help of powerful classifiers. In this work, the focus will be given on the nature of the audio signal and specific features will be defined based on the nature of the signal. The

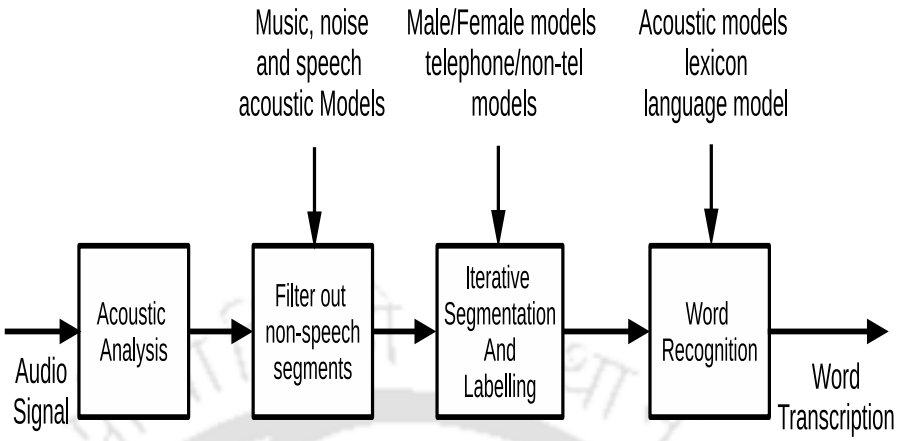


**Figure 1.2:** *Overview of transcription scheme for audio stream*

audio signal has speech components as well as other components like noise and music. The noise and music may be present in the background as overlapping or non-overlapping signals depending on the scenarios present. Defining features in terms of speech may be a better option, in particular, for the preprocessing stages considering the complexity of the scenarios present in broadcast audio. Hence the speech-specific knowledge will be used for processing broadcast audio in this work. The reason for using the speech-specific knowledge will be discussed in the later sections.

**Table 1.1:** *Statistics of human summaries averaged over* 20 *news shows [1]. An-% of anchor speaker sentences (An), % of sentences picked which indicates overlap (Ov), % of anchor speaker sentences in the overlap (An-Ov) in human summaries*

| type | An | Ov | An-Ov |
|---|---|---|---|
| % | **74%** | **63%** | **92%** |

The data which is useful for multimedia analysis purposes like audio summary may be contained in mostly the anchor speakers' audio segments. The significance of processing only the anchor speakers' audio segments can be understood from [1], where it is mentioned that humans give preference to anchor speakers' audio segments while constructing an audio summary. The statistics can be seen in Table 1.1, which shows the percentage of the anchor speaker sentences picked by the human during the audio summary. Hence in our work, only the anchor speakers' audio segments are processed, which may be useful for multimedia applications like the audio summary. The use of only the anchor

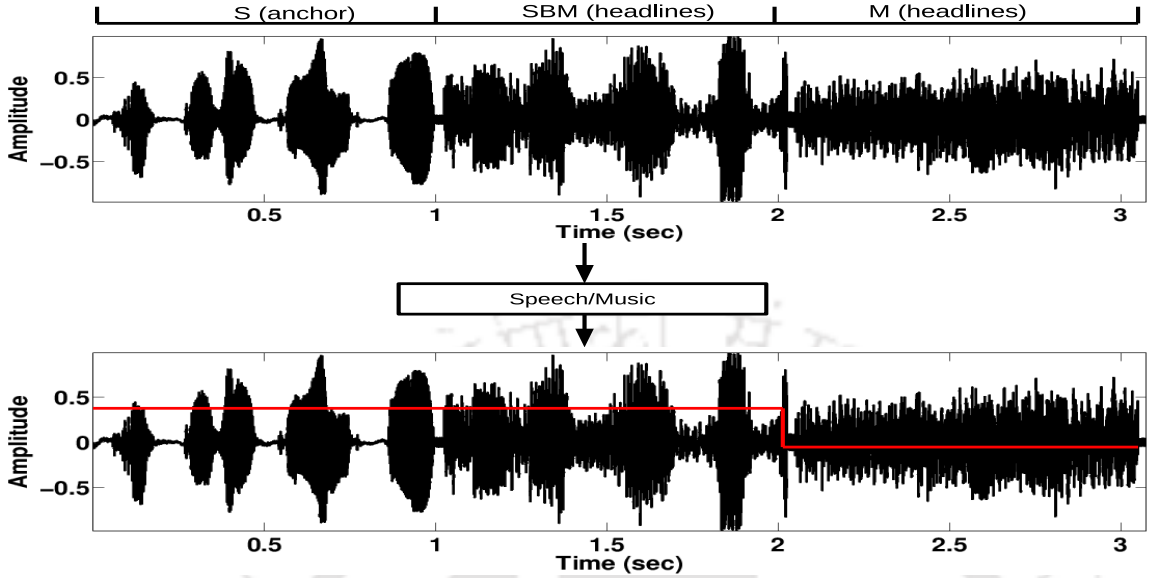


**Figure 1.3:** *Speech/Music Classification*

speakers' audio segments reduces the complexity since, in Indian Broadcast news channels, the number of anchor speakers is generally limited. Also, clean speech data constitutes the major part of the anchor speakers' audio data. However, the anchor speakers' data may also contain the other segments like pure music and speech with background music. These segments correspond to the news headlines and voice over speech. Classifying the anchor speakers' data into these segments is necessary. The music segments may be discarded since they may not contain any information for multimedia analysis purposes. However, the speech with background music segments may require transcription since they may have come from news headlines and voice over which may or may not contain useful information. If the speech with background music regions are required for recognition, their enhancement has to be performed prior to recognition in order to reduce the acoustic mismatch between the training and testing data. This has to be done because the models for the phone recognition stage in this work are assumed to be built on the clean speech case. Hence the modules required for processing the anchor speakers' data include the speech/music classification shown in Figure 1.3, clean speech/speech with background music classification shown in Figure 1.4, speech enhancement and phone recognition shown in Figure 1.5. These modules are explored in this work and the features for these modules will be proposed in terms of the speech-specific knowledge.

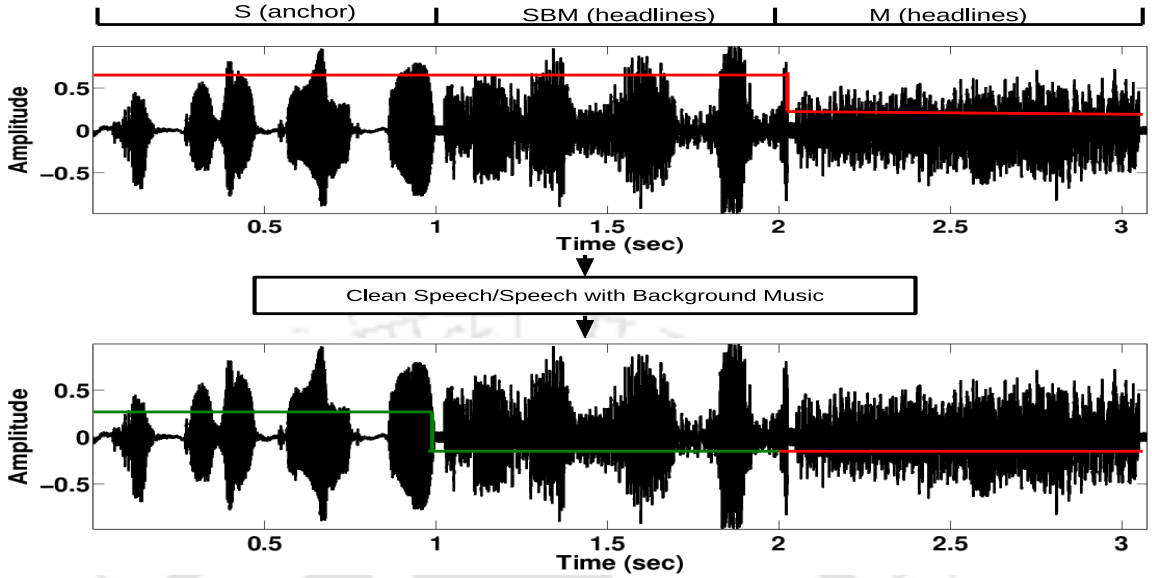


**Figure 1.4:** *Clean Speech/Speech with background Music Classification*

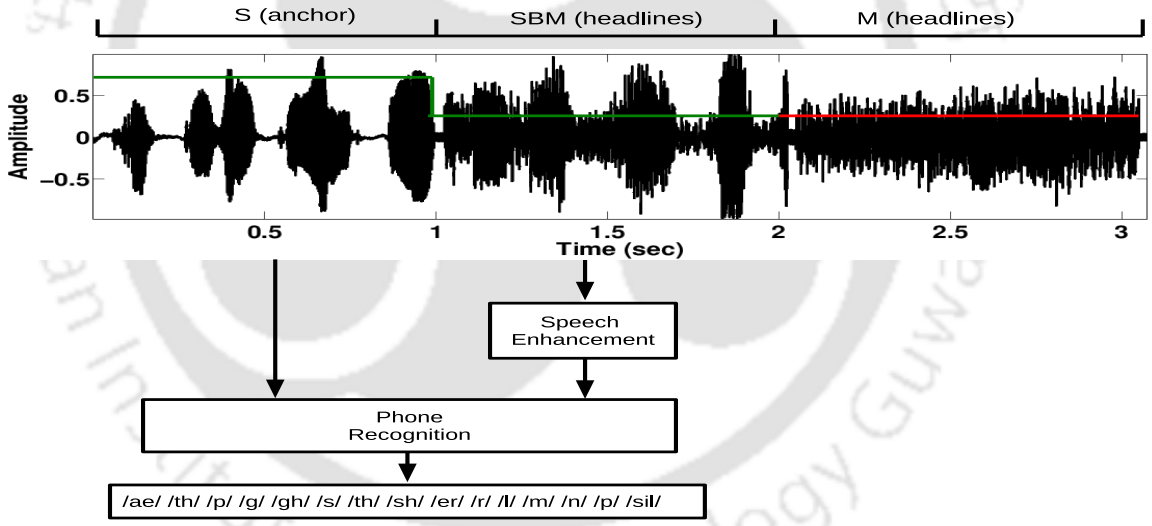


**Figure 1.5:** *Speech Enhancement and Phone Recognition*

### 1.1.2 Significance of Speech-Specific Knowledge

The use of speech-specific knowledge can be understood in some of the tasks performed earlier related to speech enhancement. In some speech enhancement task like spectral subtraction [15], the modeling of the background noise which is present in the degraded speech is necessary and this background noise is then subtracted from the degraded speech to obtain the enhanced speech. But these methods fail when the type of noise is unknown. Also, the noise may be coming from different

sources and hence it may be difficult to model the noise characteristics. In order to overcome these problems, methods which exploit the speech characteristics without considering the nature of the noise present are proposed. One such work can be found in [16] which uses the high signal to noise ratio (SNR) regions to enhance the speech signal. The LP residual, which is an excitation source based signal, is modified by enhancing the high SNR regions relative to the other regions present thereby causing the interfering noise regions to be suppressed. This modified residual is then used to excite the time-varying all pole filter in order to obtain the enhanced speech. In [16], only the speech characteristics in terms of the high SNR regions of the LP residual is exploited to perform the enhancement without considering the nature of the noise present. Similarly, the concept of exploiting the LP residual without considering the nature of the noise present was explored in [17] for the temporal based enhancement. In [18] the foreground speech enhancement was performed by also exploiting the speech characteristics in terms of the high SNR regions of the LP residual without considering the nature of the background noise present for the temporal based enhancement.

In order to illustrate the significance of speech-specific features for broadcast audio processing, the task of foreground speech segmentation and enhancement [18] was performed. Foreground speech segmentation is the use of features to classify between the desired foreground speech and background interfering sources like noise and foreground speech enhancement is the use of features to enhance the desired foreground speech. The speech signal of a speaker speaking closer to the microphone is termed as *foreground speech* and rest of the interfering acoustic sources are categorized as *background noise*. The foreground speech segmentation is initially performed on the noisy speech taken from the broadcast audio. This process is performed by using the feature based on the production knowledge of speech which is the zero band filtered signal (ZBFS). The ZBFS is a variant of the zero frequency filtered signal (ZFFS) [19] and it also gives the epoch locations pertaining to the source nature of speech production. The energy of ZBFS is used to perform the foreground speech segmentation. The nature of the output localizes the foreground speech regions while discarding the background noise regions as seen in Figure 1.6 (b). The output of the foreground speech segmentation process will then be used as a gross weight function for the temporal enhancement of the foreground speech. The temporal enhancement requires the use of both the gross weight function as well as the fine weight function to derive a total weight function which will then be used to weight the LP residual of the noisy speech [17, 18]. The fine weight function is derived by using the epoch locations obtained from

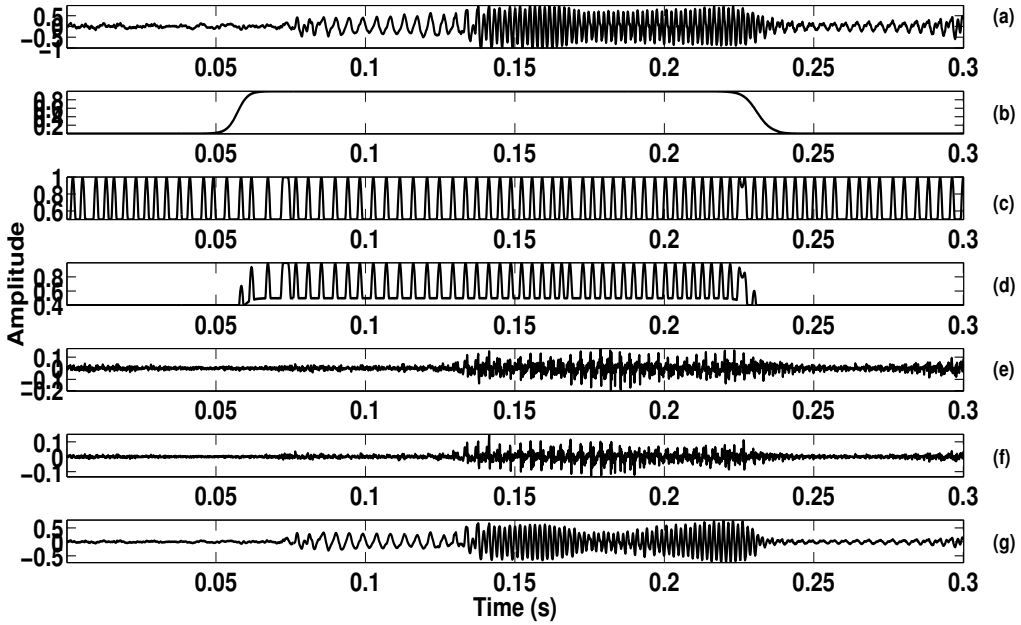


**Figure 1.6:** *Illustration of Foreground Speech Segmentation and Enhancement (a) Noisy Speech taken from Broadcast Audio where the foreground speech consist of the region from (0.05-0.25)s and the remaining is the background noise regions (b) Gross Weight Function obtained from foreground speech segmentation (c) Fine weight function (d) Overall Weight Function (e) LP Residual of Noisy speech in (a), (f) Weighted LP Residual (g) Temporally Enhanced Speech*

the ZBFS and is shown in Figure 1.6 (c). Finally, the total weight function is derived which used for weighting the noisy LP residual in Figure 1.6 (e). Figure 1.6 (f) shows the weighted LP residual and it can be observed that this residual is enhanced compared to the residual of noisy speech. In other words, the noise has been reduced in between the epoch locations of the foreground speech (0.05-0.25)s and it has also been reduced in the background noise regions. The weighted LP residual is then used to synthesize the temporally enhanced speech shown in Figure 1.6 (g). Note that in almost all the steps of foreground speech segmentation and enhancement, the nature of speech production was exploited in the form of the zero band filtered signal (ZBFS) by using its energy and its epoch locations. The nature of the background noise need not be determined. In other words, the foreground speech was segmented and enhanced based on the speech-specific features without worrying about the nature of the noise present. This illustration shows the potential of speech-specific features for foreground speech segmentation and enhancement.

Another illustration of the potential of speech-specific features can be shown for the classification of speech and music which is an important step for broadcast audio processing. When segmenting

speech and music regions, the production knowledge of speech may be used to define features which characterize the speech regions well while behaving differently in music segments thereby achieving a sort of discrimination between these two segments. To elaborate this statement, a plot is given in

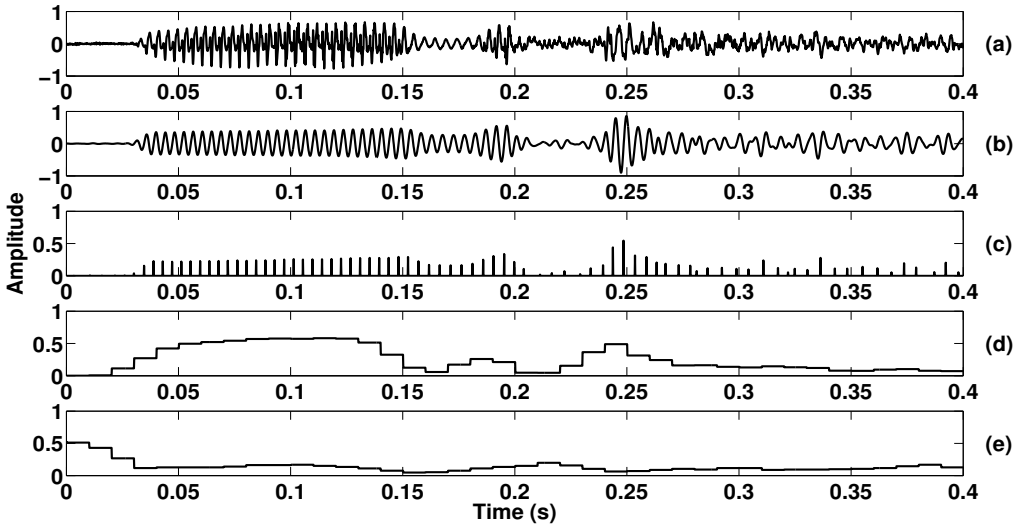


**Figure 1.7:** *Significance of Speech-Specific Features for Classification (a) Audio Signal where the first 0.2 s correspond to speech and the remaining 0.2 s correspond to music (b) Zero frequency filtered signal (c) Epoch Strength (d) Short term energy (e) Zero crossing rate*

Figure 1.7 which illustrates the use of a speech-specific feature. In Figure 1.7 (b), the zero frequency filtered signal (ZFFS) of the audio is shown which gives information about the epoch locations. As can be seen in the figure, the behavior of the ZFFS in music deviates from that of speech. The sinusoidal nature observed for speech is less observed for the music region. This shows the potential of the speech-specific feature, the ZFFS which is a feature explored for characterizing the speech production mechanism. The robustness of this feature has been compared to the short term energy and the zero crossing rate which are general audio features. The short term energy displays some discrimination between speech and music but there are regions of uncertainty (between 0.15 s to 0.2 s) where it is observed that the short term energy is low and has the same value as in the music region. Similarly, the zero crossing rate shows very little discrimination between the two segments. The epoch strength extracted from the ZFFS is also shown in Figure 1.7 (c) which shows the epoch locations and their strengths. The mechanism of producing speech is different from that of music. In this case, the music considered is guitar music and has a different mode of production compared to speech. Hence the epoch locations which are periodic at the fundamental frequency are expected to be clearly observed

for speech and different for music. This is exactly the case observed in the figure where it can be seen that for the music region, the epoch locations are randomly placed and have a low strength. This means that if the features which are developed for speech production are used for music will behave differently for the music and have a behavior which is far from speech. This provides a sort of discrimination between the two segments. This illustrates the significance of speech-specific features for the classification task.

## 1.2 Motivation for Present Work

The significance of speech-specific knowledge has been explored in various signal processing tasks like classification and enhancement as described in the previous section. The attempts infer that the processing can be done in the context of anchor speakers' segments of broadcast audio by exploiting the speech characteristics. The anchor speakers' segments contain scenarios such as clean speech, speech with background music and music. For the phone transcription of anchor speakers' segments, the classification and enhancement have to be performed as indicated earlier in Section 1.1.1. The main task would involve removing the music components and enhancing the speech with background music. The main interfering source of anchor speakers' segments of broadcast audio is music. However, for performing classification and enhancement with music as an interfering source, may require the use of speech-specific characteristics without worrying about the nature of the music present. The music may be of different types and may be originating from different sources, hence exploiting characteristics of music may be difficult. On the other hand, exploiting the speech-specific characteristics may be a better option since the nature of speech production in terms of the source and the vocal tract system is uniform across speakers. The speech characteristics have been extensively studied in terms of the source and vocal tract system and the potential of speech-specific features for classification and enhancement has also been demonstrated in the previous section for speech/music classification, and foreground speech segmentation and enhancement. This indicates that the processing of anchor speakers' segments in broadcast audio may be done by focussing mainly on the speech-specific characteristics.

On this context, for the speech/music classification task, the features in terms of the speech-specific knowledge can be used. These features will have a standard behavior in speech while deviating in music. Since the features are defined in the context of speech production, any other class of speech, for example, speech with background music will be classified as speech. Similarly, another set of speech-

specific features can be defined for the clean speech/speech with background music classification. These features will have a normal behavior in speech while deviating in the speech with background music segments due to the presence of music in the latter. The speech-specific features can also be proposed for enhancement of speech with background music segments, which will emphasize the high signal to music ratio regions relative to other regions. This will cause the background music regions to be suppressed so as to obtain the enhanced speech. Finally, the clean and enhanced speech can be passed through the phone recognition system to obtain an improved phone recognition accuracy.

## 1.3 Organization of the Thesis

To address the issues mentioned, this thesis work is organized into six chapters. The content of each chapter is summarized as follows:

- Chapter 2 reviews the literature related to the classification, enhancement and recognition methods. The review brings out the significance of speech-specific knowledge for the various tasks performed to obtain a solution for the phone recognition of broadcast audio
- In Chapter 3, the speech/music classification system is discussed which employs the use of speech-specific features. These features are defined in terms of the source, vocal tract system and syllabic rate of speech. These features characterize speech regions well while deviating their behavior in music regions thus achieving a sense of discrimination between speech and music.
- In Chapter 4, the clean speech/speech with background music classification system is presented which exploits the vocal tract system feature. The spectral characteristics of the vocal tract system in the average and relative sense is exploited to define features for this task.
- In Chapter 5, the speech enhancement of the speech with background music is explored expecting that the enhanced speech will give a good phone recognition accuracy. The enhancement methods are performed based on mostly the source characteristics which again exploits the speech-specific nature of the audio segment.
- Chapter 6 discusses the combination of the speech/music classification, clean speech/speech with background music classification and speech enhancement modules for effective phone recognition in the context of broadcast audio. The features defined for the different modules are utilized in this chapter to assess their overall performance on the task of phone recognition.

- Chapter 7 discusses the summary and conclusion

# 2

# Processing Broadcast Audio - A Review

Contents

*A description of the literature for the processing of broadcast audio is presented in the form of a review, where the specific focus will be on the phone transcription of broadcast audio. Broadcast audio comprises of complex scenarios which make their phone transcription not so straightforward as the phone transcription of clean speech. The preprocessing steps are required before the final phone recognition step and the details of the preprocessing will also be reviewed. Finally, based on the various advantages and disadvantages of the different methods in the literature, an alternative direction will be proposed which may give a better phone transcription accuracy and with a lesser complexity of preprocessing than the existing techniques. This framework will be particularly based on the speech-specific knowledge of processing the broadcast audio, wherein the source, system and suprasegmental information of speech will be exploited for processing.*

## 2.1 Introduction

The processing of broadcast audio is essential for applications like spoken document retrieval (SDR), story segmentation (SSEG) and story summarization (SSU). In most of these applications, the phone transcription of broadcast audio is an important step. The automatic phone transcription of broadcast audio is a complex task owing to the large variabilities present in the data which involves a large number of preprocessing steps before the final phone recognition step. The variabilities in the data are a result of the scenarios present in the broadcast audio which include indoor speech which generally corresponds to the anchor speech, outdoor speech which may be either the anchor or reporter speaking and advertisements and news headlines. The anchor speakers' speech is mostly free from interfering sources such as noise, but in cases where is a voice-over, which may be present in between the anchor speakers' speech, music may be present in the background of speech. Additionally, the news headlines which are also part of the anchor speakers' speech, contain music in the background as an interfering source for speech. The presence of these interfering sources may cause the phone transcription of speech to have low accuracy due to the presence of an acoustic mismatch between the training and testing data. The transcribed speech from the indoor scenarios may contain a lot of information for the multimedia applications. The outdoor speech, on the other hand, is mostly noisy in nature and the information which may be obtained from outdoor speech may be redundant and may already be available in the anchor speakers' speech. The significance of anchor speakers' speech has been shown in [1], where it is mentioned that humans give preference to anchor speakers'

audio segments while constructing an audio summary. Hence the phone transcription of only the anchor speakers' speech may be sufficient in applications like multimedia analysis such as an audio summary. The processing of only the anchor speakers' speech will tend to reduce the complexity of the preprocessing stage since the only interfering source which has to be taken care is the music which may or may not overlap the speech signal.

Some of the preprocessing steps involved include the speech/non-speech detection, bandwidth detection, and gender detection [8]. These preprocessing steps are required in order to partition the incoming audio data stream into homogeneous segments so that they can be used for either speaker adaptation or speech recognition. Since the focus is on the phone recognition of anchor speakers' speech, the desired segments will be mostly clean in nature so as to obtain a good level of phone transcription accuracy. In order to segment the clean speech segments from the anchor speakers' speech, the clean speech first needs to be segmented from the other scenarios present like the pure music and speech with background music. The speech obtained from the first level of classification may contain music in the background as well. Hence the clean speech may again be needed to be segmented from these types of segments. Finally, to obtain the clean speech from the speech with background music, the enhancement will be required to be performed. Finally, the clean speech obtained will be required to be converted into text using a phone recognizer.

The review for this work will be organized as follows. Initially, some broadcast audio applications will be discussed where the common tasks of automatic transcription of broadcast audio appearing in these applications will be identified. Next, the different types of preprocessing tasks which are the significant steps for automatic transcription of broadcast audio will be reviewed. There is not much literature present for the phone recognition of broadcast audio in general, but mostly it is mentioned as speech recognition even though the task of speech recognition involves the phone recognition as a prior step. In this work, the terms phone recognition and speech recognition may be used interchangeably, but overall, the goal of the literature survey is to provide a direction for phone recognition.

The speech/music classification task which is part of the preprocessing step will be reviewed in more detail since this task is essential for the direction to be proposed in this work. Similarly, some methods for segmenting clean speech and speech with background music will be discussed. The literature related to speech enhancement in the context of background noise will also be discussed expecting that these methods may work for the case when there is music in the background as well.

This review is necessary since the direction to be proposed in this work relies on the enhancement module to overcome the problem of having a lesser amount of data for training. Finally, the literature related to systems which employ speech recognition in the context of broadcast audio will also be reviewed since this is the essential step for processing the broadcast audio.

## 2.2 Tasks Involving Broadcast Audio Processing

The majority of the tasks which involve the processing of broadcast audio include the transcription. Some of these tasks consist of the spoken document retrieval (SDR), speech summarization (SSU) and story segmentation (SSEG). This section will highlight some of these tasks (SDR, SSU, and SSEG) followed by the review on some of the basic preprocessing steps which are the significant parts of the automatic transcription of broadcast news.

### 2.2.1 Spoken Document Retrieval

Spoken document retrieval (SDR) of broadcast audio has been performed in [2] and it is a task which involves the use of automatic transcription of the audio data. This task involves a phoneme based [2], [3] approach where phonetic transcription of the audio data is generated by the recognizer. This transcription is processed for producing subword unit representations which are useful for indexing and retrieval. There is also a word based approach [20], [21] where a large vocabulary speech recognition system is built for providing a word-level transcription. The information retrieval (IR) system treats the transcribed audio segment as a text document.

Documents and queries are normally represented in the form of vectors. Every vector component constitutes an indexing term. A term may consist of a word, a word fragment or a sub-word unit. Each term consists of an associated weight which is normally based on the term's occurrence statistics both within and across documents [2]. In the vector for document $j$, the weight of term $i$ is:

$$d_j[i] = 1 + log(f_j[i]) \tag{2.1}$$

and in the vector for query $k$, the weight of term $i$ is:

$$q_k[i] = (1 + log(f_k[i]))log(N/n_i) \tag{2.2}$$

where $f_j[i]$ is the frequency of term $i$ in document or query $j$, $n_i$ is the number of documents having term $i$, and $N$ is the total number of documents present in the collection. The second term is called the

inverse document frequency (idf) for term $i$. A normalised inner product similarity measure between document $d_j$ and query $q_k$ can be used to score and rank the documents during the retrieval

$$S(d_j, q_k) = \frac{d_j.q_k}{||d_j||||q_k||} \tag{2.3}$$

A spoken document retrieval (SDR) system is described in [22], for British and North American Broadcast News. This consist of a connectionist large vocabulary speech recognition system along with a probabilistic information retrieval system. In [22] an approach which is word based is adopted and hence the audio signal is represented in the form of text. Given a document $d$, a term-weighting function is defined which gives a weight for a term $t$. The within-document term frequency and the collection frequency weight are contained in this function. The number of occurences of term $t$ in document $d$ is the term frequency, $TF(t, d)$. The collection frequency weight of term $t$, $CFW(t)$, indicates a measure of the proportion of the collection in which the term appears:

$$CFW(t) = log\left(\frac{N}{n(t)}\right) \tag{2.4}$$

for a total collection of $N$ documents, where term $t$ appears in $n(t)$ documents.

The product of the collection frequency weight and the term frequency results in a combined weight as follows,

$$CW(t, d) = \frac{(K + 1).CFW(t).TF(t, d)}{K((1 - b) + b.NDL(d)) + TF(t, d)} \tag{2.5}$$

where $K$ is the discounting parameter on the term frequency, $NDL(d)$ represents the length of $d$ which is normalized by the length of the mean document across the collection and $b$ is a constant defined empirically which controls the influence of the length of the document ($0 \leq b \leq 1$).

The overall weight for a document $d$ corresponding to the query $Q$, $W(Q, d)$ can be calculated as

$$W(Q, d) = \sum_{t \in Q} CW(t, d) \tag{2.6}$$

where $CW(t, d)$ is the combined weights for each term present in the query and relevant to the document.

A Mandarin Chinese broadcast news retrieval system which is syllable-based was explored in [23]. The retrieval process is as follows. Consider a database $D$ and a query $q$, the retrieval problem is

nothing but a searching process which is used for finding the document $d^*$ mostly related to the query. The searching process is given as,

$$d^* = \text{argmax}_{d \in D} \text{Sim}(d, q) \tag{2.7}$$

where $\text{Sim}(d, q)$ indicates the measure of similarity between a document $d$ and a query $q$. Each spoken document $d$, which is based on the syllable lattice $l_d$, can be represented by two feature vectors $V_d^1$ and $V_d^2$. The standard Cosine measure [24] is used for estimating the similarity.

$$\text{Sim}(d, q) = w cos(V_d^1, V_q^1) + (1 - w) cos(V_d^2, V_q^2) \tag{2.8}$$

$$= w \frac{V_d^1 . V_q^1}{|V_d^1||V_q^1|} + (1 - w) \frac{V_d^2 . V_q^2}{|V_d^2||V_q^2|} \tag{2.9}$$

where $w$ indicates the weight used for adjusting the contributions from the single part and the syllable pair part in the overall similarity measure.

### 2.2.2 Speech Summarization

In [4] an automatic speech summarization system was proposed in which the set of words are extracted from automatically transcribed speech based on a particular summarization score. Dynamic Programming (DP) technique is used for extraction in accordance with the target compression ratio. The set of words extracted is then fused to form a summarization sentence. The summarization score is based on the word significance measure, a confidence measure, linguistic likelihood and a word concatenation probability. The dependency structure in the original speech which is given by the stochastic dependency context-free grammar (SDCFG) is used to determine the word concatenation score. The Japanese broadcast news speech transcribed using an LVCSR system is summarized in [4].

In [5], the speech and language technologies were described and an effort which integrates these technologies into a system is introduced. This system can be used for indexing speech data, creating structural summarization and providing tools in order to browse the stored data. The technologies explored in [5] consists of systems based on the speaker-independent speech recognition, segmentation, and identification of speaker, spotting of a name, classification of topics, segmentation of stories and retrieval of information. The continuous audio input stream by the speaker is first automatically segmented by the system. Next, the system is used for clustering the audio segments which belong

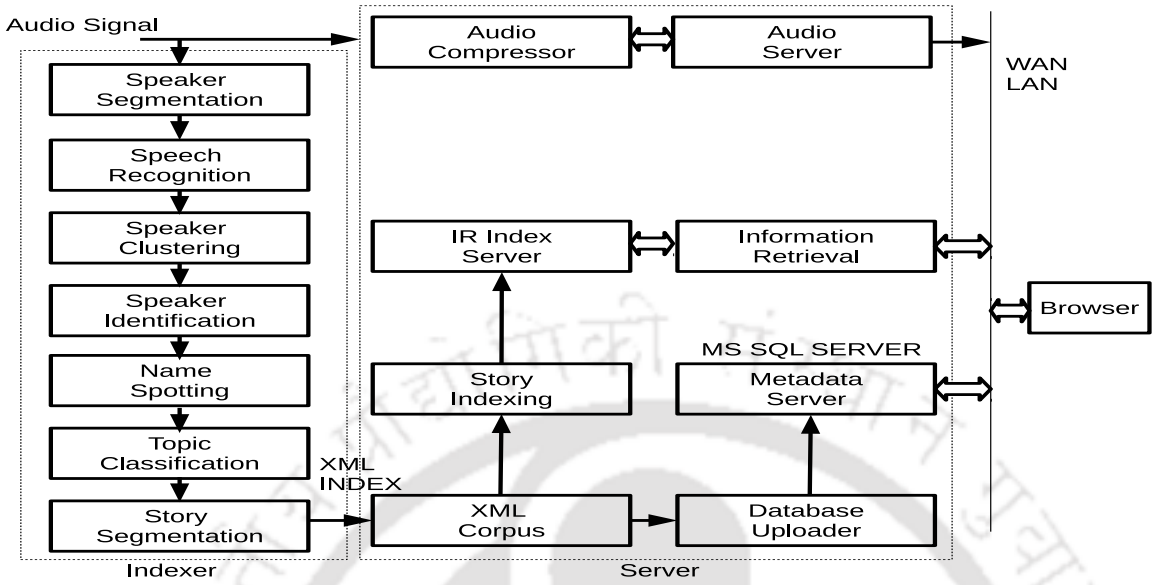


**Figure 2.1:** *Audio Indexing and Retrieval System*

to the same speaker. Identification of speakers which are known to the system is also done and finally, the spoken words are transcribed (see Figure 2.1 obtained from [5]). The input stream is also segmented into stories by the system, depending on the topic content. The names of persons, places, and organizations are also located by the system. The integration of the technologies mainly allows the system to produce a high-level structural summarization of the spoken language thereby allowing easy browsing of data. The transcription for this system is created by the Byblos large-vocabulary speaker-independent speech recognition system [25].

In most of the applications for processing broadcast news (for example the SDR, SSEG, and SSU), the automatic transcription of speech is a necessary step. The information retrieval system treats the transcribed audio as a text document. Similarly, the transcription is a crucial step for SSEG and SSU system as seen in Figure 2.1. In the automatic transcription of speech, the speech recognition is a general approach wherein phoneme models are created using the features extracted from the training data. The complexity of the scenarios present in broadcast audio poses challenges in terms of obtaining homogeneous segments which will improve the transcription accuracy. The preprocessing step is used to obtain the homogeneous segments and is an important step for the automatic transcription of broadcast news considering the complexities encountered in the broadcast audio. Accordingly, the previous work for preprocessing the broadcast news will be provided in the following sections.

### 2.2.3 Iterative Maximum Likelihood Segmentation/Clustering Procedure

In [8], the broadcast audio data is partitioned into segments which are homogeneous in nature. Identification and removal of the non-speech segments are performed by using the GMMs trained to detect speech, pure music and others (background). Labeling of the clustered speech segments according to bandwidth and gender is then performed. The advantages of using the partitioning or preprocessing before transcription are plenty. Firstly, speaker turn and speaker identity information can be extracted in addition to transcription. Secondly, problems which are caused by the linguistic discontinuity at speaker changes can be avoided. Thirdly, overall performance can be improved since the acoustic models will be trained on popular acoustic conditions. Finally, removing non-speech segments and using shorter segments for processing reduces the time required for computation and the decoding is simplified. An iterative procedure of maximum likelihood segmentation/clustering based data partitioning is performed in [8], where Gaussian mixture models are used along with agglomerative clustering. The approach does not depend on the language. The result of this process is a set of speech segments having speaker labels, gender labels, and telephone/wideband labels. In this method, a cluster of segments generally represents a speaker in a particular acoustic environment. Typically there are a slightly higher number of clusters than the true speakers for a particular show. For example, two clusters may be used to represent a given speaker's data, one which corresponds to speech with background music and another which does not contain music. A 38-dimensional mel frequency cepstral coefficient (MFCC) feature vector was used which is similar to the 39-dimensional MFCC popularly used for speech recognition but excluding the energy

### 2.2.4 Phone Decoding Segmentation Procedure

In [6], an algorithm is proposed which comprises of bandwidth detection, gender detection, speaker change detection and speaker turn clustering. In the bandwidth detection stage, a one-pass Viterbi decoder was used to detect the bandwidth by hypothesizing a sequence of phoneme classes tagged with bandwidth labels and their time offsets. Phonemes are grouped into 8 classes: voiced constituents, fricatives, and sibilants, obstruents, breath, laughter, lip smack, music, and silence. The first 5 classes are further tagged with wide bandwidth and narrow bandwidth labels, resulting in a total of 13 distinct symbols for band detection. Each acoustic training sentence is automatically labeled as wideband or narrowband based on the ratio of the average energy above and below 4 kHz. Its original

phonetic transcription and time-aligned labels are then transformed into band-specific phoneme-class symbols. A forwardbackward training procedure to train the model. The episode-length hypothesis of bandwidth-tagged phoneme class symbols, after being smoothed to suppress spurious bandwidth changes, is then separated into two sets of band-specific chunks. The result of this stage is two sets of band-specific long segments ready to be gender-classified. The gender detection stage involves phoneme symbols duplicated and tagged with male and female labels. The pronunciation of each gender-specific word now consists of corresponding gender-specific phonemes. Words of each acoustic training sentence are now tagged with appropriate gender and standard forward-backward training procedure is followed to train the dual gender models. The output of the decoding stage for each segment is a hypothesis of gender tagged words. The hypothesis is also smoothed to remove spurious gender changes. In the speaker change detection (SCD) and speaker turn clustering stage, each set of a band and gender-specific segments is first fed through the SCD component using the penalized likelihood ratio test and speaker clustering is performed to pool similar data together for consistency in cepstral normalization and for adaptation. Linear Prediction Coding (LPC) based mel warped cepstral of 45 dimensions are used as features in this work.

### 2.2.5 Combined GMM and Phone Decoding Segmentation Procedure

In [10] an algorithm was proposed which involves using gaussian mixture models for classifying audio stream into wideband speech, narrowband speech, and music. A model for speech with background music was also created and audio selected by this particular model is labeled as wideband speech. The separate inclusion of this model reduces the misclassification of speech as music. The music portion is rejected. Segmentation and gender labeling of the narrowband and wideband speech data are performed by using a phoneme recognizer which has a certain number of context independent phone models for a particular gender along with a silence/noise model. The output of the recognizer is a sequence of short segments with each having either a male, female or silence tag. Silence segments of duration greater than 3 seconds are usually classified as non-speech and discarded. A heuristic smoothing is also applied to eliminate spurious gender changes. The 39-dimensional MFCCs have been used as feature vectors.

### 2.2.6 Segmentation Procedure using Distance based Methods

In [9] the Symmetric KL (Kullback Leibler) distance metric was used for segmentation and clustering. KL (Kullback Leibler) distance between two random variables A and B represents an information theoretic measure which is equal to the additional bit rate accrued by encoding random variable B with a code that was originally designed for optimal encoding of A [26]. The larger this value, the more the distance between the two PDFs of 2 random variables. It is hence formulated as

$$KL(A;B) = E_A < log(P_A) - log(P_B) > \tag{2.10}$$

where $E_A$ represents the expectation operation performed with respect to PDF of A.

However, this expression is not symmetric which means that it is not a distance metric strictly. They defined the Symmetric KL (Kullback Leibler) distance metric designated KL2 metric as:

$$KL2(A;B) = KL(A;B) + KL(B;A) \tag{2.11}$$

The KL2 distance metric was used to perform segmentation without acoustic classes since the testing domain in the broadcast news may include different shows, both from television and radio. Means and variances were computed for every two-second window of the audio stream. When the KL2 distance between the adjacent windows reached a local maximum, a new segment boundary was generated. Classification of the segments into either full or half bandwidth was performed using a pair of gaussian mixture models. Given a segment of data, maximum likelihood selection of the class was used to classify incoming speech. The clustering stage involves grouping the segments having the same speaker identity and channel. Agglomerative clustering was chosen wherein the KL2 distance was used over the Mahalanobis distance. Depending on a threshold distance, an utterance was either clustered with an existing cluster or it was used as a seed for a new cluster.

### 2.2.7 Hypothesis Testing Segmentation Procedure

In [7] the use of Hotelling's $T^2$-test was explored for creating homogeneous segments. This test was used in the speaker change detection module of the segmentation process. This method considers a sliding window across the speech stream. A speaker break-point was hypothesized by splitting the frames into speaker A and speaker B within a particular window. For each hypothesized break-point, the $T^2$-statistic is computed

$$T^2 = (\mu_A - \mu_B)^T [S(\frac{1}{n_A} + \frac{1}{n_B})]^{-1} (\mu_A - \mu_B) \tag{2.12}$$

where $\mu_A$ and $\mu_B$ are the vectors of the parameter means for the two pieces,$n_A$ and $n_B$ are the frame counts and $S$ is a universal within speaker covariance matrix. The break points are decided based on the peak value of $T^2$. The features used consisted of 36-dimensional cepstral features which are (Perceptual Linear Prediction) PLP-based.

In most of the preprocessing task for the automatic transcription of broadcast news, MFCC or LPC based features are used. In addition to that most of the approaches assume a large amount of data is available for training the models. Some approaches can be explored which make use of the robust features for the classification by analyzing the signal characteristics. The problem of unavailability of large data can also be reduced by considering the enhancement of the segments containing background music and background noise. Accordingly, some features for the classification task similar to the classification task of the preprocessing stages of the broadcast audio will be discussed. Some enhancement strategies will also be reviewed which may be useful for the preprocessing stages of the broadcast audio transcription.

## 2.3 Features for Classification of Speech and Music

The classification of speech and music falls in the category of speech/non-speech detection in broadcast audio processing. There have been several features which have been explored in the literature along with some sophisticated classifiers. The description of some of the features is given below.

### 2.3.1 Temporal Based Features

The simplest time domain feature is the short term energy which has been employed for the speech/music classification task in [11, 12]. The short term energy in each frame is computed as

$$E_n = \frac{1}{N} \sum_{n=0}^{N-1} s^2[n] \tag{2.13}$$

where $s[n]$ is the discrete-time audio signal, $n$ is the time index and $N$ is the frame length. The energy contour of a certain audio waveform is capable of separating speech and music. For music signals, the energy contour shows a smaller number of dips and peaks than speech and also shows a little change over several seconds. Speech has an alternation between voicing and frication which

indicates a marked change in its energy contour [11]. The short term energy has also been used as a feature in terms of the percentage of low energy frames measure in [12]. The energy distribution for speech is more left-skewed than music since there are more silence frames and it is expected that this measure will be higher for speech than music.

The zero crossing rate [11] feature has also shown promising ability to classify speech and music. The speech shows a marked rise in the ZCR during the frication periods while music has no abrupt increases due to its tonal nature. The ZCR variation of music does not have the same nature as that of voiced and unvoiced speech. The nature of the ZCR distribution for speech tends to be skewed towards the high end causing several points in the distribution to lie at values higher than the mean. These events are due to the presence of unvoiced fricatives and affricates in speech. For discrete-time signals, it is considered that a zero-crossing has occurred if successive samples show different signs. The short-time average zero-crossing rate for each frame is defined as

$$Z_n = \frac{1}{2}\sum_{n=1}^{N-1} |sgn(s[n]) - sgn(s[n-1])| \tag{2.14}$$

where

$$sgn(s[n]) = \begin{cases} 1 & s[n] \geq 0 \\ -1 & s[n] < 0 \end{cases}$$

where $s[n]$ is the discrete-time audio signal, $n$ is the time index of short time ZCR and $N$ is the frame length. The probability of null zero crossings has been explored in [13]. The silent intervals are mostly present in the speech which causes the number of zero crossings to be null for speech in the silent intervals whereas this probability will be low for music due to the absence of silent intervals in music.

Speech has a characteristic energy modulation peak around the 4 Hz syllabic rate [12]. A portion of the MFCC algorithm is used for converting the audio signal into 40 perceptual channels. The energy in each band is extracted and each channel is bandpass filtered with a second order filter having a center frequency of 4 Hz. The short-term energy is then calculated by squaring and smoothing the result. Normalization of each channels 4 Hz energy by the overall channel energy in the frame is performed, and the result from all channels are summed. The modulation energy at 4Hz tends to be higher for speech than music.

### 2.3.2 Spectral Based Features

The spectrum centroid which represents a spectral based feature for classification of speech and music, has been used in [14] and is defined as the center of gravity (COG) of the discrete fourier transform (DFT) spectrum for a given frame of audio $s[n]$. It is computed as

$$S_c = \frac{k\sum_{k=0}^{K-1}|X(k)|}{\sum_{k=0}^{K-1}|X(k)|} \tag{2.15}$$

where $X(k)$ indicates the DFT of the frame of audio $s[n]$.

This measure indicates whether low or high frequencies are dominated in a power spectrum and can be regarded as an approximation of the perceptual sharpness of the signal.

The spectral flux is another spectral based feature which measures the fluctuations of the DFT spectrum between two consecutive audio frames and is defined as

$$S_f = \sum_{k=0}^{K-1}(|X_m(k)| - |X_{m-1}(k)|)^2 \tag{2.16}$$

where $X(k)$ indicates the DFT of the frame of audio $s[n]$.

The spectrum roll-off point which is also another robust spectral based feature, is defined as the boundary frequency $f_r$, where a certain percent $p$ of the DFT spectrum energy for a given frame of audio is concentrated below $f_r$ :

$$\sum_{k=0}^{f_r}|X(k)| = p\sum_{k=0}^{K-1}|X(k)| \tag{2.17}$$

where $X(k)$ indicates the DFT of the frame of audio $s[n]$.

The above three spectral based features are general based audio features and have been used even in some general audio tasks as well like audio classification and music genre classification.

The spectrum spread is used as a feature in [14] and this measure computes how concentrated the DFT spectrum is, around the perceptually adapted audio spectrum centroid. It is calculated as follows.

$$S_{sp} = \sqrt{\frac{\sum_{k=0}^{K-1}([log_2(f(k)/1000) - ASC]^2.|X(k)|^2)}{\sum_{k=0}^{K-1}|X(k)|^2}} \tag{2.18}$$

where $f(k)$ represents the frequency in each frequency bin, and ASC is the perceptually adapted audio spectral centroid, which is defined as

$$ASC = \frac{\sum_{k=0}^{K-1} log_2(f(k)/1000).|X(k)|^2}{\sum_{k=0}^{K-1}|X(k)|^2} \tag{2.19}$$

$X(k)$ indicates the discrete fourier transform (DFT) of the frame of audio $s[n]$.

In [27], the filter bank energy is explored for classification of speech and music. The energy variance of the filter bank (EVFB) coefficients tends to be higher for speech than music. The difference is more prominent in the lower frequency portions of the DFT spectrum. The reason is because speech has its signal energy which is concentrated in the lower frequency portions of the DFT spectrum. The presence of more number of silence segments in speech compared to music motivates the computation of the EVFB coefficients. When a speaker is speaking short pauses are made constantly between words and syllables and it is observed in [28] that more pauses are made within the pronounced words than between words. During short pauses, the signal has low energy. This effect also occurs when plosives are pronounced. Music has lesser silent moments than speech since there is always some instrument playing. Another motivation is that the duration of vowels is shorter in speech than music. High energy is present within the speech signal when a vowel is pronounced. Vowels are high energy carriers in speech while in music the vowels have longer durations, especially during choruses. The third motivation is the pitch variation. The pitch can vary by as much as 160 Hz when a speaker is speaking. Rapid variations in the tempo of the syllabic rate are typical. The pitch variation in music is less and the rate change is also usually slower than speech. The EVFB features are calculated by first sampling the audio signal, followed by windowing and FFT calculation. Filtering of the signal's magnitude spectrum is then performed by using a set of triangular filters. The logarithmic energies of each triangular filter are then calculated and the process in terms of mathematical equations is described below.

$$EN_g[i] = log\left[\sum_{k=1}^{M} |S[k,i]f_g[k]|^2\right] \tag{2.20}$$

in which $i$ represents the frame number, $g$ indicates the filter number, $k$ represents the frequency bin and $M$ represents the total number of bins.

$$S[k,i] = \sum_{n=0}^{N-1} s[n+iR]w[n]e^{\frac{-j2\pi nk}{N}} \tag{2.21}$$

where $s[n]$ is a frame of audio signal, $w[n]$ indicates a rectangular window, $R$ represents the frame shift and $N$ is the frame length. The term $f_g[k]$ defines the triangular filter having the central frequency $f_{cent}[g]$ on the linear scale, which depends on the distribution of filter. For the mel scale

filter distribution,

$$f_{\text{cent}}[g] = 700(10^{\frac{m}{2595}} - 1) \tag{2.22}$$

similarly for the inverse mel scale filter distribution,

$$f_{\text{cent}}[g] = 4031.25 - 700(10^{\frac{2195.28-m}{2595}} - 1) \tag{2.23}$$

and for the linear scale filter distribution,

$$f_{\text{cent}}[g] = m \tag{2.24}$$

where $m$ indicates the index number in mel filter, inverse mel filter or linear scale filter distribution and $g$ represents the filter number.

The variance of the log energies of each of the filter channels is calculated to form EVFB coefficients over the filter channels.

### 2.3.3 Posterior Probability Based Features

The mean entropy for each frame which is defined as

$$\frac{1}{N}\sum_{n=0}^{N-1}\sum_{k} -p(q_k^n).log(p(q_k^n)) \tag{2.25}$$

where $n$ represents the frame number (time index), $N$ indicates the number of frames in a particular segment, and $p(q_k^n)$ represents the posterior probability for label $q_k$ at particular time $n$ which is as calculated by the acoustic model, has been used as a feature for discriminating speech and music in [29] The posterior probabilities for a particular time represent a true pdf because of the mutually exclusive nature of the phone labels. The entropy of that pdf is used as a measure for the goodness-of-fit to the acoustic model of the current observations: A distribution in which there is a large probability for a particular single label tends to have an entropy near to zero, whereas the situation in which roughly equal probabilities are assigned to a number of labels will have a larger entropy.

Another posterior probably based feature has been explored in [29] called the average probability dynamism. This feature was developed based on the motivation that as normal speech transits between phones for every few tens of milliseconds, the estimated probabilities for speech segments which have been well-modeled, change abruptly and frequently. The signals which do not belong to speech rarely and gradually cross the phone boundaries. This results in the reduction of this feature value. It is

defined as follows

$$\frac{1}{N}\sum_{n}\sum_{k}(p(q_k^n) - p(q_k^{n-1}))^2 \tag{2.26}$$

The background-label energy ratio represents another posterior probably based feature and has been explored in [29]. Initially, a wide range of signals is used for training the background label. This label can catch all non-speech intervals, and it is mostly active for signals which do not belong to speech. For a clean speech segment, the segments with background labels should have much less energy compared to the other present segments (where speech has been identified). This causes the ratio of the expected energy of segments with background labels to the expected energy of segments with speech labels to be very small. The segments which are dominated by non-speech or speech segments which are noisy in nature will assign a background label to the high-energy frames, thus making the ratio equal to one or larger. It is defined as

$$\frac{\sum_n p(q_{sil}^n).e^n / \sum_n p(q_{sil}^n)}{\sum_n (1 - p(q_{sil}^n)).e^n / \sum_n (1 - p(q_{sil}^n))} \tag{2.27}$$

Here, $q_{sil}$ represents the specific classifier label associated with inter-speech gaps present in the training corpus, and $e^n$ represents the energy of the original speech signal present in a window around time step $n$.

The Phone distribution match also represents the class of posterior probability based features which is based on the motivation that music segments are observed to be classified as a single phone label (such as /n/). This distribution measure should be able to highlight wildly skewed instances defined. It is defined as

$$(S_r - \bar{S})^T C^{-1} (S_r - \bar{S}) \tag{2.28}$$

Where $S_r$ represents a vector of statistics computed for a total set of probability values for the segment $r$, in this case, the standard deviation along time of each label's probability, weighted to discount background/silence states, e.g.

$$S_r(k) = \sqrt{\frac{\sum_n (p(q_k^n) - p(\overline{q}_k))^2 (1 - p(q_{sil}^n))}{\sum_n (1 - p(q_{sil}^n))}} \tag{2.29}$$

where $p(\overline{q}_k)$ signifies the mean probability for label $q_k$ over a segment. For the whole set of labels

excluding background/silence, the values are arranged into $S_r = \{S_r(k)\}$ . The expected value of this vector is denoted by $\bar{S}$ and is based on a training set of known clean speech samples, and $C$ represents the covariance matrix of the training vectors, assumed to be diagonal for avoiding over-fitting. It should be possible to capture these underlying patterns in $\bar{S}$ to the extent that observed speech segments are long enough to be phonetically balanced, and to detect the arbitrarily-different behavior of non-speech segments.

### 2.3.4 Chroma Based Features

A feature has been used for music tonality based tasks like the chord or key identification. This feature is called the pitch class profiles or chroma vectors and has been represented as a feature in order to discriminate speech and music in [30]. The speech signal lacks musical scale. Further, the specific keys and tonal structures following strict patterns in the spectral domain dominate the music signal. This causes the feature to have a good discriminating ability. Pitch class profiles are based on the principle of octave invariance. This principle states that the musical notes which are separated by a doubling of frequency have no functional difference. Same characteristics are observed between a chord present in one octave and the same chord in another octave. This principle is utilized by the pitch class profiles to reduce the spectrum $X(f)$ by adding frequencies which are separated exponentially into the same bin, thereby causing folding of octaves of the spectrogram into the same range.

$$c(k) = \sum_{r=0}^{R-1} |X(2^{\frac{k+rK}{K}} f_{min})| \tag{2.30}$$

where the $k^{th}$ chroma bin is calculated. $K$ represents the total number of bins present in the chroma vector, $R$ denotes the number of octaves and $f_{min}$ indicates the lowest frequency in the summation process. The chroma feature captures chord and key information which varies between music segments, and this feature alone may not be effective for speech/music classification. For this purpose, the peakiness measure in the chroma vector was proposed in [30].

Musical tones are concentrated around certain frequencies more often than others. The music theory of the particular culture or music style determines the commonality of these frequencies and their relationships. However, most kinds of music have a specific set of rules regarding notes and note relationships. Speech, however, is far less strictly regulated in terms of the use of the pitch. These differences tend to allow music to have stronger and more separated peaks in their chroma vectors,

while the chroma vectors of speech become smoother with energy mostly concentrated around the bins which correspond to the fundamental frequencies and formant structure of speech. As a result, musical chroma vectors are expected to be more peaked as a function of k as compared to speech chroma vectors. Two metrics are used for this characteristic. First, the energy after differentiation of a normalized chroma vector is computed as follows.

$$CD = \sum_{k=0}^{K-1} |c(k) - c(mod(k+1, K))|^2 \tag{2.31}$$

where $CD$ denotes the chroma difference. Note that $c$ is the chroma vector and the differentiation is calculated circularly since the musical tones represented by the chroma vector are related circularly. Another feature which is the second one, is obtained by adding the high-frequency energy in the normalized DFT spectrum.

$$CHF = \sum_{l=l_{min}}^{l_{max}} |F\{c(k)\}(l)|^2 \tag{2.32}$$

where $CHF$ denotes the chroma high frequency, $F\{.\}$ represents the discrete fourier transform (DFT) and the feature is the total energy in the DFT spectral range $[l_{min}, l_{max}]$.

## 2.4 Features for Clean Speech/Speech with Background Music Classification

The clean speech/speech with background music classification task appears in the context of some audio classification task. The features used for this audio based classification may have similar characteristics with the temporal and spectral based features defined for speech/music classification earlier. Some of the other explored features for clean speech/speech with background music classification in the context of audio classification are as follows.

### 2.4.1 Spectral Peak Track

The characteristics of the type of sound can be revealed by the peak tracks [31] which are present in the audio signal spectrogram. For example, the spectral peak tracks stay at a same level of frequency and remain for quite some time in the case of musical instrument sounds. The peak tracks for sounds emerging from human voices are harmonic in nature and mostly align in the shape of a comb. The song segments have spectral peak tracks which vary over a fundamental frequency range from 87 Hz to 784 Hz. Songs have relatively long and stable tracks because the voice remains at a particular note over a

time period and often appears in the form of ripples due to the vocal chord vibration. Speech segments have spectral peak tracks which lie around the lower frequency regions. The spectral tracks are also much closer to each other in speech since the range of the fundamental frequency of speech is from 100 to 300 Hz. The length of spectral tracks for speech segments also tend to be shorter because of the intermissions present between voiced syllables, and the fluctuation may be slow because of the pitch change during the syllable pronunciation of various types. Separation of non-silence audio segments into two categories is done first, which are with or without music components. This separation is done by detecting frequency peaks from the power spectrum which have a continuous and stable nature. AR model parameters of order 40 are used for generating the power spectrum. The calculation is done once every 400 input samples. The signal frame for computing the spectrum has a size of 512 samples.

For a certain time period, if the peaks detected in consecutive power spectra stay at the same level of frequency, this particular period of time is normally indexed as having music related components. The speech harmonic peaks or the noise having low frequency is avoided by considering only the spectral peaks above 500 Hz. Most of the music components lie in this range. Signal frames which fall below a certain defined energy level are also discarded. For each segment of sound, an index sequence is generated where the index value is given a value of 1 if the sound is detected as having components related to music at that particular instant and to 0, otherwise. A measure called Zero Ratio is defined which is the ratio between the number of zeros in an index sequence and the total number of indices in the sequence. This is used as a measurement to indicate whether the sound segment has components pertaining to music or not. If there are lesser music components present in the sound, the ratio goes higher. The zero ratios of different types of audio are examined. It is observed in [31] that for the speech with background music if the speech signal present is strong, the music in the background is mostly hidden and cannot be detected. However, during the intermission periods of speech or if music signal gets stronger, the music components can be detected.

## 2.5 Methods for Enhancement

Several enhancement methods have been applied in the context of the presence of background noise as a degradation. Even though there are other kinds of degradation present, in most cases the background noise is a major form of degradation. Also, the methods for enhancing the other kinds of degradation other than background noise have almost a similar general approach. The methods of

speech enhancement in the context of background noise will be reviewed here and the possibility of their robustness for the presence of music in the background will be examined.

The methods generally assume the additive nature of the background noise. Consider a signal $s(n)$ degraded by background noise $d(n)$ which is additive in nature. The speech which is degraded can be represented by

$$x(n) = s(n) + d(n) \tag{2.33}$$

and in the spectral domain we have,

$$X(k) = S(k) + D(k) \tag{2.34}$$

where $k$ represents the frequency index, $X(k)$ denotes the average magnitude DFT spectrum of the degraded speech and $S(k)$ is the average magnitude DFT spectrum of the desired speech. $D(k)$ is the average magnitude spectrum of the degradation present. The speech enhancement method estimates $\overline{s}(n)$ which approximates the desired speech $s(n)$. In essence, the method of enhancement aims for the minimization of the error represented by

$$e(n) = s(n) - \overline{s}(n) \tag{2.35}$$

where $e(n)$ denotes the error introduced between the desired speech $s(n)$ and the speech estimated, $\overline{s}(n)$.

The speech enhancement gives advantages like smooth communication between humans [32] due to the benefits obtained like better speech quality and intelligibility. In addition, these benefits may help in speech application tasks like speech and speaker recognition [33]. The speech enhancement methods are generally categorized into spectral and temporal enhancement methods.

### 2.5.1 Spectral Based Methods

The spectral based methods represent the simplest methods of enhancement and are also found to be very effective. The spectral processing methods rely on the fact that the perception of human speech is insensitive to short-time phase [34, 35]. These methods can be categorized into different kinds. The first kind involves magnitude spectral estimation of the degradation and this is subtracted from the magnitude spectrum of the degraded speech on a frame by frame basis to obtain the spectrum

of desired speech. This is called the spectral subtraction. The second method is based on a wavelet denoising method. The first two methods are also known as the non-parametric methods. The third method is based on the minimum mean square estimator (MMSE) which is a statistically based method and relies on the parametric model of the signal generation process.

#### 2.5.1.1 Spectral Subtraction

The background noise reduction can be performed by spectral subtraction. This method has been explored much earlier in the past and is still followed in recent works as a comparison due to the simplicity of the method and also because of the relative ease at which the algorithm can be implemented. Spectral subtraction can be done by first estimating the magnitude spectrum of the degradation and then this is subtracted from the magnitude spectrum of the degraded speech on a frame by frame basis, to obtain the spectrum of the enhanced speech [15]. The magnitude spectrum of the degradation is normally computed during the speech pauses to improve the degradation modeling. Mathematically, this is represented as follows [15]:

$$|\overline{S}(k)| = |X(k)| - |\overline{D}(k)| \tag{2.36}$$

where $\overline{S}(k)$ is the average magnitude spectrum of the estimated clean speech, $X(k)$ is the average magnitude spectrum of speech which has been degraded and $\overline{D}(k)$ is the average magnitude spectrum of the degradation.

Generally, the errors are introduced while estimating the spectrum of the degradation and these errors may introduce negative values for the enhanced spectrum in Equation 2.36. To obtain a non-negative magnitude spectrum, half wave rectification of the values can be performed and this process introduces peaks in the spectrum which are small and isolated and which occur at random frequency locations in each frame. On converting to the temporal domain, these peaks are perceived as tones with frequencies that change randomly from frame to frame. The tones are referred to as musical noise [35–37] and this can be annoying to the listeners. There have been several methods attempted in the literature to reduce the musical noise [37–40].

In [15], the reduction of musical noise was attempted by methods such as magnitude averaging, residual noise reduction and additional signal attenuation during non-speech activity. In [36], a method for musical noise reduction was proposed in which the noise spectrum is overestimated and a preset threshold value is used to avoid the resultant spectral components from going below that value.

Mathematically this is presented as [36]:

$$|\overline{S}(k)| = \begin{cases} |X(k)| - \alpha|\overline{D}(k)|, & |X(k)| - \alpha|\overline{D}(k)| > \beta|\overline{D}(k)| \\ \beta|\overline{D}(k)|, & otherwise \end{cases} \tag{2.37}$$

where $\alpha$ denotes the over-subtraction factor. This factor is a function of the noisy Signal to Noise Ratio (SNR) and calculated as

$$\alpha = \alpha_0 - \frac{3}{20}\text{SNR}, \quad -5dB \leq SNR \leq 20$$

where $\alpha_0$ is the value of $\alpha$ at $0dB$ SNR. The value of $\alpha$ has to be chosen in such a way that the musical noise and signal distortion are reduced and the $\beta$ value indicates the spectral floor which floors the values of the spectral components of the enhanced speech which fall below a preset lower value.

A method of spectral subtraction based on adapting the subtraction factor depending on the frequency is proposed in [38, 41]. The speech signal may not be affected by noise uniformly over the entire spectrum. Some frequency components may be affected more than the others. This motivated the proposal of the non-linear spectral subtraction (NSS) method which is based on the linear spectral subtraction method proposed in [36]. The NSS method comprises of varying the over-subtraction factor depending on the frequency in every frame of speech. Larger values of the over-subtraction factor are used to attenuate frequencies with low SNR and smaller values are used to attenuate the frequencies with high SNR.

In [38], a multi-band spectral subtraction method was proposed which is based on the concept of the non-linear spectral subtraction above. This multi-band approach involves dividing the signal into a set of non-overlapping bands. An independent over-subtraction factor calculated from the corresponding signal sub-band is then applied over each band. The above methods do not work well when the SNR is low. The reason being that it is difficult to suppress the noise without degrading the intelligibility and also without introducing other distortions and residual noises.

Perceptually based approaches were proposed in [42, 43], where the masking properties of the auditory system [42, 43] were exploited to mask the noise instead of eliminating the musical noise and introducing distortion. The masking effect gives the stronger signal the ability to make the weaker signal (which occurs at the same time) inaudible. If the noise is weaker compared to speech then the noise is masked by the speech signal based on the masking properties [42]. Hence lesser noise

subtraction can be performed which avoids distortion. The algorithms proposed in [42,43] attempt to attenuate the noise below the audible threshold, instead of removing all the noise components from the signal. In [43], the over-subtraction factor $\alpha$ and noise floor parameter $\beta$ are controlled depending on the audible threshold. The parameters $\alpha$ and $\beta$, are dependent on the frequency components, denoted as $\alpha(\omega)$ and $\beta(\omega)$ which is similar to the non-linear spectral subtraction method [41]. The parameters are related to the audible threshold as follows

$$\alpha_m(\omega) = F_\alpha[\alpha_{min}, \alpha_{max}, T(\omega)] \tag{2.38}$$

$$\beta_m(\omega) = F_\beta[\beta_{min}, \beta_{max}, T(\omega)] \tag{2.39}$$

where $m$ represents the frame index, $T(\omega)$ is the threshold based on noise masking and is obtained by modeling the human ear's frequency selectivity and its masking property. $\alpha_{min}$, $\beta_{min}$, and $\alpha_{max}$, $\beta_{max}$ represent the minimum and maximum values of over subtraction and spectral flooring, $F_\alpha$ and $F_\beta$ represent the functions leading to a maximum residual noise reduction

Although these methods are able to reduce the musical noise, the main disadvantage is the complexity and heavy computational requirements associated with psychoacoustic modeling.

#### 2.5.1.2 MMSE Estimator

There is no statistical property assumption of either the noise or the speech spectral components in the spectral subtraction method of speech enhancement. There are some methods which incorporate the use of the probability distributions and these have been explored in [44, 45]. The speech and noise spectral components are observed to be statistically independent zero mean Gaussian random variables in [44]. From the noisy signal, the clean speech is estimated, and this is mapped perceptually to the original clean speech by using the minimum mean square error (MMSE). The short-time spectral amplitude (STSA) was used for enhancement. The MMSE-STSA estimator is mathematically represented as,

$$\widehat{M} = E\{(A_k - \widehat{A_k})^2\} \tag{2.40}$$

where $\widehat{M}$ denotes the MMSE-STSA estimator, $A_k$ represents the STSA of the clean speech signal, and $\widehat{A_k}$ denotes the STSA of speech signal estimated from the noisy signal.

The mean square error between the short time magnitude of the clean and enhanced speech signal is generally minimized in the method of MMSE-STSA estimator. A good level of enhancement with reduced musical noise is obtained with a certain optimality criterion, which is based in terms of the mean square sense and without considering the non-linear characteristics of human audible perception. Another method called the MMSE log-spectral amplitude (MMSE-LSA) that considers the non-linear nature of human perception is proposed in [45]. This method computes the logarithm of the short time DFT spectral amplitude of clean and estimated speech and minimizes the mean square error between the two. This MMSE-LSA method is given as,

$$\widehat{L} = E\{(logA_k - log\widehat{A_k})^2\} \tag{2.41}$$

where $\widehat{L}$ denotes the MMSE-LSA estimator, $logA_k$ represents the logarithmic STSA of the clean speech signal, and $log\widehat{A_k}$ is the STSA of speech signal estimated from the noisy signal. MMSE-LSA showed a better reduction of musical noise than MMSE-STSA. However, very little improvement is achieved in the speech quality for the MMSE-LSA compared to the MMSE-STSA.

Generally, discrete cosine transform (DCT) is known to have a better energy compaction property compared to the discrete Fourier transform (DFT). This property of DCT is utilized to get better results for the MMSE estimator in [46]. In MMSE-STSA estimator both the real and imaginary parts of the DFT coefficients are assumed to have Gaussian probability distribution function. However, this assumption is valid mostly for the speech frame of longer duration. Since the frame size taken for the analysis is around 20-30 ms, the assumption of Gaussian distribution may be valid only for the noise components of the signal but not for the speech components. Hence different probability distributions are used to model the real and imaginary parts of the DFT coefficients, and particularly Gamma and Laplacian distributions have been used in [47, 48].

Another method called the ideal binary masking (IDBM) technique was proposed, which utilizes a priori SNR measurements to obtain clean speech from the noisy speech signal. A time-frequency transformation is applied along with the binary masking which is based on a priori SNR and the time-frequency analysis is performed by using DFT or gamma-tone filter bank analysis which is a frame based approach [49]. The noise components and speech components are suppressed by different gain values. This method seems to perform well. However, we know that in natural recording, a priori SNR is not available and the binary masking will have to be calculated on the basis of a posteriori

SNR. This error in estimating SNR causes the quality of the IDBM method to degrade and gives poor enhancement quality. A continuous gain function was proposed in [49], which is found to be superior compared to the binary masking function.

A DNN based method was employed in [50] which is a supervised approach for speech enhancement. A mapping function is obtained between the clean speech and the estimated one using deep neural networks (DNN). A simulated version of the database is created to incorporate all combinations of speech and noise signal. This database can be created by adding noise of different types and of different levels to speech. However, the problem is these methods degrade when the noise is of unseen type. A global variance equalization has been used to overcome such problems. These supervised methods show good performance compared to the MMSE approaches. The degradation in the performance can be seen when there is large deviation of the noise characteristics. The supervised methods also have a disadvantage that training data is required and they are computationally more intensive.

Although the performance of the MMSE and the MMSE-LSA methods are inferior compared to several methods, these approaches are found to be the most popular methods. The MMSE method has been used in hearing-aid applications [51]. These methods are also used as the state-of-the-art techniques so that the newly proposed methods can be compared to these techniques in terms of the performance.

#### 2.5.1.3 Wavelet Denoising

Most of the speech enhancement algorithms which work in the spectral domain are analyzed using the short-time Fourier transform (STFT). The STFT presents a compromise between the time and frequency resolution, but the time resolution is fixed for all frequency components once the frame length is chosen. There are enhancement algorithms based on the wavelet transform which gives a flexible time-frequency representation of speech [52]. Wavelet shrinkage [53] is a simple denoising method which uses thresholds to define a limit between the target signal and noise wavelet coefficients. However, simple thresholding may not give a good separation of the components of the target signal from the noise, which means that uniform thresholding to all wavelet coefficients will not only suppress the additional noise but also some speech components, in particular, the unvoiced segments, since the energies of the unvoiced segments are comparable to that of noise [52]. Hence the quality of the filtered speech is affected. Other methods like combining wavelet transform with signal processing tools consisting of Wiener filtering in the wavelet domain and wavelet filter bank have been proposed

in [54]. Perceptually motivated wavelet decompositions have also been attempted in [55].

### 2.5.2 Subspace Approaches for Enhancement

The speech and noise for the case of noisy speech can be separated by using a method of subspace filtering. The speech and noise components are divided into subspaces having a mutually orthogonal nature and it is assumed that the speech and noise can be represented in terms of a low-rank linear model. The correlation assumption is given for speech while the noise is assumed to be uncorrelated [56]. The enhancement is performed by removing the subspace of the noise and reconstructing the signal back into the subspace of speech. The decomposition of the noisy signal can be performed by either Singular Value Decomposition (SVD) [57] or Karhuen-Loeve Transform (KLT) [58].

The signal and noise subspaces are represented in terms of the eigenvectors and the corresponding eigenvalues in the SVD approach. They are assumed to be separable in the eigensubspaces and the reconstructed signal using the dominant eigenvectors and the eigenvalues result in the clean speech signal. There is also a quotient SVD method which has been proposed in [59]. In KLT method, the noisy signal is separated into the speech and noise subspaces and a gain function derived aids in differentiating the speech and noise components. The KLT is inversed using the modified KLT coefficients by the gain function. This helps to enhance the speech signal. Since the noise is assumed to be stationary in such approaches, the performance for the degradations which are non-stationary is poor. Also, the computational requirements are high in these subspace based approaches.

### 2.5.3 Temporal Based Methods

The previous methods of speech enhancement which are also mentioned earlier, mostly involve modeling the background noise present in the degraded speech signal and subsequently, the noise is eliminated from the signal based on the model. These kind of approaches are simple and seem to be very effective. However, there are certain drawbacks for these kinds of methods. For example, if the type of noise is unknown then modeling such noise becomes a very tedious task. Also, there are many kinds of noises which are produced by different types of sources and the sources are plenty which means that modeling all these noise sources is also very difficult and cumbersome. The are also cases where the noise present in the signal may be non-stationary and modeling non-stationary noises is also a very difficult task. Some non-stationary noises include background speakers and the use of conventional approaches may fail under these conditions. Hence an alternative approach is necessary to take care

of these scenarios and the approach which is based on directly exploiting the desired speech signal characteristics can be explored. Some evidence has been shown that humans generally perceive the information from the noisy speech by focusing on the high SNR regions and then extrapolating to the low SNR regions in order to fill the gaps which are created by the interfering noise [60]. This evidence led to the exploration of some work like the one in [16], which effectively uses the high SNR regions to enhance the speech signal. The high SNR regions in terms of the excitation source are mostly exploited to enhance the degraded speech. The Linear Prediction (LP) residual represents the excitation source information. The LP residual is modified and then used to excite the time-varying all pole filter in order to obtain the enhanced speech. The reason for using the LP residual is due to the fact that it has a random polarity and also exhibits noise-like characteristics. This nature of the LP residual makes it possible to modify the residual by enhancing the instants of significant excitation relative to the other regions present and thereby causing the interfering noise to be reduced. The advantage of using this method is that it causes minimum distortion as compared to the earlier mentioned methods like spectral subtraction and MMSE based approaches. The Frobenius norm of the Toeplitz matrix computed from 2 ms frame size of LP residual is used in order to exaggerate instants of significant excitation. There have also been similar approaches that use a different weighting scheme for the LP residual based on a constrained optimization criteria [61].

The concept of LP residual modification has been extended to the case of noisy and reverberant speech enhancement for multiple microphone recordings [62]. In this method, the time delay between the recordings of two microphone is generally computed by the cross-correlation of two residual signals. In order to enhance the significant excitation instants, the Hilbert Envelope (HE) of LP residual is first computed and this is used as a weighting function for modifying the LP residual. All the enhanced LP residual signals are added together synchronously to synthesize the enhanced speech using multiple microphone recordings. The main idea is based on the fact that the noise and the reverberant components of the recordings of the multiple microphones are added incoherently, while the speech components are added coherently. The addition of the speech components coherently by compensating for the delay helps to exaggerate the high SNR regions and thereby aiding in the enhancement of the speech signal.

The combined temporal and spectral processing method for speech enhancement has been proposed in [17]. This method also employs the idea of modifying the LP residual for the temporal case. The

glottal closure instant (GCI) locations or the instants of significant excitation are obtained using the Hilbert envelope of LP residual and these locations are further boosted in LP residual relative to other present regions. The modified LP residual is then used to excite an all-pole filter to obtain the enhanced speech. Next spectral subtraction is performed on the temporally enhanced speech in order to obtain a better-enhanced speech. The combined temporal and spectral techniques present advantages like better quality and absence of musical noise for the enhanced speech.

## 2.6 Speech Recognition for Broadcast Audio

Continuous Hidden Markov Models (HMMs) with gaussian mixture models (GMM) are generally used for speech recognition of broadcast audio where 39 dimensional Mel Frequency Cepstral Coefficients (MFCC) are used as the feature vectors [4, 8, 10]. In [8] the acoustic models for American English were trained on 150 hours of broadcast news data which has been distributed by LDC. A subset 10 hours from the 600 hours of unpartitioned, unrestricted American English broadcast data was randomly selected and used for testing the models. Similarly in [6] a Continuous HMM is used for modeling the phonemes with the GMM for acoustic modeling but a 45-dimensional feature vector is used wherein the features are extracted from overlapping frames of audio data. A window of 29 ms is used with a frame rate of 100 frames/sec. A Hamming window is used for framing the segments. A 36-pole Linear Prediction Coding (LPC) smoothed power spectrum is computed for a particular frequency band. The 14 mel-warped cepstral coefficients are computed from this. For each segment of speech, the mean cepstrum and peak energy are removed non-causally from the appropriate sub-vector thus removing any long term bias due to the channel. The feature vectors are scaled and translated so that for each speaker turn, the data has zero mean and unit variance. The cepstral features are taken along with the first and second derivatives and the energy to form the 45-dimensional feature vector. Multilayer Perceptron (MLP) Based features have been explored in [63]. The features consisting of Perceptual Linear Prediction (PLP) features are fed as input to the MLP. Temporal context information is also incorporated for the PLP features which means a high dimensional feature is given as input to the MLP. Next, the PCA transformation is performed on the MLP output and this is given as input to the HMM for training and testing the speech recognizer.

## 2.7 Discussion and Direction for the Work

Most of the speech/music classification task in the literature rely on the use of general temporal and spectral based features. Some may describe these features as being general audio features. Speech is produced by exciting a time varying vocal tract system by a time varying excitation source. The mechanism of producing speech remains same across different speakers. Additionally, there have been a lot of work which has studied the production and perception aspects of speech production. The music especially the instrumental ones are produced in a different manner and the sources of music can be of many types. Defining features in terms of music may be difficult although an attempt has been made in terms of the chroma based features but then again in terms of production, the features for music cannot be clearly defined. On the other hand defining features in terms of the source, vocal tract system and suprasegmental characteristics of speech may be a better option since a lot of work for speech has been studied in terms of these characteristics. These speech-specific features may perform well for speech but their behavior for the music regions may deviate from their normal characteristics. This deviation in the music regions may provide some sort of discrimination for speech and music regions. The discussion about speech/music classification is further detailed in Chapter 3.

The clean speech/speech with background music classification task is also performed using the general audio features in the previous literature. For the speech with background music regions, in most cases, the music is added after the speech has been produced. The fact that it is possible to perceive speech even in the presence of background music shows that the source and vocal tract characteristics are intact. However, the overall signal characteristics of that speech with background music segment may be affected and cause their nature to be different compared to their nature in clean speech. The difference in nature is mostly due to the presence of music. By deriving features based on the source and vocal tract characteristics, some kind of discriminative information for the clean speech and speech with background music may be achieved. Even though the feature may capture the source and vocal tract characteristics of the speech and speech with background music segments, the presence of music in the latter may cause deviation in the value of these features and this deviation is mostly due to the effect of the music in the signal. This deviation provides a sense of discrimination for the clean speech and speech with background music segments, thereby allowing for an effective classification of the two classes. Chapter 4 discusses the clean speech/speech with background music classification in more detail.

The speech enhancement in the previous literature relied on the use of background noise modeling. Such methods are feasible when the type of noise is known and also if the noise is stationary. Similarly, other approaches like temporal based approach are explored which exploit the speech characteristics directly without considering the noise present. The combined temporal and spectral based approaches showed good enhancement performances for the background noise cases. The combined temporal and spectral processing based methods can also be attempted when the background present is music since it was shown in the literature that the combined temporal and spectral based methods do not take into account the background noise types. The speech-specific features in terms of the source, system, and suprasegmental characteristics can be used for deriving the gross and the fine weight functions to enhance the speech with background music segments. In most of the previous works on speech recognition of broadcast audio, the GMM-HMM system has been used. Other explorations such as the ANN-HMM or the DNN-HMM systems can be explored for the modeling purpose. Recent methods consisting of the SGMM-HMM system can also be explored. The issues on enhancement and recognition of speech with background music are discussed in Chapter 5.

Chapter 6 discusses the overall phone recognition of broadcast audio by combining the methods developed in Chapter 3, 4 and 5. The given anchor speaker audio signal containing clean speech, speech with background music and music is passed through the preprocessing steps and finally, the segmented and enhanced portions are passed through a phone recognizer. The impact of the preprocessing techniques will be discussed in detail in that Chapter.

# 3

# Speech/Music Classification using Speech-Specific Features

Contents

### Objective

*This work proposes the use of speech-specific features for speech / music classification. Features representing the excitation source, vocal tract system and syllabic rate of speech are explored. The normalized autocorrelation peak strength of zero frequency filtered signal and the peak-to-sidelobe ratio of the Hilbert envelope of linear prediction residual are the two source features. The log mel energy feature represents the vocal tract information. The modulation spectrum represents the slowly-varying temporal envelope corresponding to the speech syllabic rate. The novelty of the present work is in analyzing the behavior of these features for the discrimination of speech and music regions. These features are non-linearly mapped and combined to perform the classification task using a threshold-based approach. Further, the performance of speech-specific features is evaluated using classifiers such as Gaussian mixture models, and support vector machines. It is observed that the performance of the speech-specific features is better compared to existing features. Additional improvement for speech / music classification is achieved when speech-specific features are combined with the existing ones, indicating the different aspect of information exploited by the former.*

## 3.1 Introduction

Audio data obtained from the broadcast news channels generally consists of complex scenarios. Some of these include speech recorded in studio which is of good quality, speech recorded in the field which is mostly in outdoor environments and may contain background noise, speech with background music, vocal and non-vocal music. Hence processing audio data for different multimedia applications is a challenging task. Among different issues, the fundamental one to pursue is speech / music classification, needed for separation of speech and music regions for further processing. The current work explores the task of speech versus music classification.

The speech / music classification task has been explored in several ways in literature using different features and classifiers [11–14,27,29,64,65]. This work proposes to explore the speech-specific features for speech / music classification motivated from the use of music-specific features explored in [30]. There are several reasons for looking at this task from the speech-specific point of view. The music signal cannot be generalized so easily due to the presence of different types of music sources. Hence selecting robust features relating to music is a difficult task. Speech is produced by humans and

extensive work has been done to study the speech production and perception systems in terms of the excitation source, vocal tract system, and the dynamics associated with them. The gross mechanism for producing speech remains the same across the human race. In order to produce a particular sound unit, the shape of the vocal tract (lowering jaw) and glottal vibration as excitation source remains mostly the same for a particular speaker. Even though there are other factors like fundamental frequency, pronunciation and speaker's individual anatomy that influence the production of a particular sound unit across different speakers, the major factors involved in the production are the shape of the vocal tract and the nature of the excitation source. Hence, exploring the behavior of speech-specific features which exploit the characteristics of the excitation source, vocal tract system, and syllabic rate of the speech signal may be a better option for speech / music classification. The behavior of these speech characteristics in music segments is expected to be different compared to the speech segments.

The quasi-periodic and impulsive nature of the glottal vibration (a major excitation source in speech production) is unique to speech production. The normalized autocorrelation peak strength (NAPS) [66,67] of the zero frequency filtered signal (ZFFS) represents the quasi-periodic nature of the excitation source information of speech. The peak-to-sidelobe ratio (PSR) [68] of the Hilbert envelope (HE) of linear prediction (LP) residual feature represents the impulsive nature of the excitation source information. The majority of energy in case of speech is in the vowel-like sounds and concentrated in the low frequency region of the audio spectrum. The log mel energy feature can be used to exploit this property, and hence to represent the vocal tract information. Due to the physical limitation of the speech production system, the number of sound units that can be generated per unit time is also limited. The rate of speech production can be measured using the modulation spectrum. This feature, which has already been exploited in [12], represents the changes in the slowly varying temporal envelope corresponding to the speech syllabic rate [69]. These speech-specific features are different and carry independent evidence for speech / music classification.

The main idea of the work is the use of speech-specific features for the speech / music classification task. The NAPS of ZFFS has been explored for the task of foreground speech segmentation in [66], where the periodicity of the ZFFS for foreground speech is higher compared to the background speech and noise. The NAPS was used to measure the periodicity. The ZFFS was extracted based on the average pitch period of speech [19]. The pitch period of speech is lower than the pitch period of music and a study related to finding the pitch period of speech and music can be found in [70], which is an

autocorrelation based method. This method has no upper frequency limit search range and hence it is possible to use this algorithm to find the pitch period of music. Extracting ZFFS of music according to the method in [19] will be interesting considering the higher pitch of music and the nature of the ZFFS of music may be different compared to speech as seen in the case of background speech and noise in [66]. This difference may be exploited for classifying speech from music signals.

The PSR of HE of LP residual has been explored in the case of processing degraded speech [71], where the spurious instants of significant excitation detected from small random peaks in the Hilbert envelope are eliminated using this measure. It is also used as a quantity to compare the cross-correlation function of different methods in [68]. There is impulse-like excitation for speech signals, but such an excitation is completely different for the music signal. The HE of LP residual of speech has been used to find the impulse-like excitation in the speech in earlier works. However, since the nature of excitation in music may be different, it will be interesting to study the behavior of this feature in music. Hence this feature is explored for the speech / music classification task. The best way to measure the differences of the HE of LP residual in speech and music is in terms of the PSR which is found to act as an effective measure for different tasks explored in [68, 71].

There is an alternating nature of vowel and non-vowel regions in speech and this kind of nature may not be present in the music. In [72], the gross vocal tract information was represented in terms of the sum of ten largest peaks of the DFT spectrum for the task of vowel onset point detection. The log mel energy can be considered to be an extension of this feature. However, for the log mel energy feature, the source information is smoothed out by passing the DFT spectrum through the mel filter banks. The difference of the vowel nature in speech and music regions can be captured in terms of the log mel spectrum energy, which represents the vocal tract system information for the speech / music classification task.

Except for modulation spectrum, to the best of our knowledge, the other features have not been explored in speech / music classification task. In particular, their behavior in music needs to be studied. Since these features are speech-specific, their behavior in the music regions may deviate from speech regions. This gives a kind of discrimination between the speech and music regions. Accordingly, the novelty of the work may be summarized as follows:

- The overall concept of using the speech-specific features for speech / music classification. These features may have been explored in earlier literature in the context of speech. The behavior of

these features in music is analyzed in this work. Their ability to discriminate speech from music is examined, so as to achieve an effective speech / music classification system.

The significance of the proposed speech-specific features may be explained as follows: For instance, the NAPS feature is estimated (to be described later) using *a priori* knowledge of human pitch range. As a result, the NAPS will be able to localize speech regions better compared to music regions. This may result in a distinct behavior of NAPS for speech and music regions. Alternatively, existing features for speech / music classification like zero crossing rate banks on the generic characteristics of the audio signal in the time domain, and not on any specific properties of speech production and perception. Thus NAPS may be more effective compared to zero crossing rate. These reasons motivate us to explore the speech-specific features for speech / music classification.

Considering the complexity of the audio data in broadcast news, and since the task involves only a two-class classification, certain regions in the audio signal are defined as to be either speech or music. This task involves obtaining the speech regions as much as possible and hence speech in all kinds of scenarios (indoor, outdoor and with music background) belongs to the speech class. The music class contains either vocal or non-vocal music, where a majority of the broadcast news music segments considered consists of non-vocal or instrumental music. The speech with background music refers to news headlines and advertisements, and the vocal music category includes songs and advertisements having the singing voice. Although, there are different scenarios within the speech or music class, the focus of this work is mainly on classifying audio[1] into speech and non-vocal music segments. The music segments which contain singing mixed in with musical instruments are also considered. The music segments which do not involve musical instruments but contain only singing are not considered in this work.

Initially, the classification task is performed using a threshold based approach and non-linear mapping on the proposed speech-specific features. The classifiers such as Gaussian mixture models (GMMs) and support vector machines (SVMs) are then considered with the speech-specific features given as the input. Existing features like zero crossing rate (ZCR), spectral roll-off, spectral flux, spectral centroid and percentage of low energy frames which are popularly used in the literature have also been considered for the classification. The speech-specific features are then concatenated along with the existing features and the effect of this combination on the classification task is studied. The

[1] http://www.iitg.ernet.in/cseweb/tts/tts/Assamese/speech_music.php

rest of the work is organized as follows: The description of the speech-specific features and their significance are given in Section 3.2. Speech / music classification using the speech-specific features is described in Section 3.3. Section 3.4 describes the results and discussions. Finally the conclusion of the work is given in Section 3.5.

## 3.2 Speech-Specific Features for Speech/Music Classification

This work focuses on the use of speech-specific features which will be described in a detailed manner. The other existing features consisting of spectral flux, spectral centroid, spectral roll-off [14], zero crossing rate (ZCR) [13], and percentage of low energy frames [12] have been used extensively for the speech / music classification task. Therefore they will not be described in detail here due to their availability in the literature.

### 3.2.1 Speech-Specific Excitation Source Features

The quasi-periodic and the impulsive nature of the excitation source of speech are exploited for deriving features which are described below.

#### 3.2.1.1 Normalized Autocorrelation Peak Strength

The ZFFS gives information about the epoch locations in the speech signal [19]. The speech signal is passed through a resonator located at the zero frequency which preserves signal energy around zero frequency and significantly attenuates all other information, mainly due to the vocal tract resonances. The trend in the output of zero frequency resonator is removed further by considering a window of length one or two pitch periods. The trend removed signal is termed as the ZFFS [19]. The positive zero crossings of ZFFS are demonstrated to give the location of epochs. The ZFFS is obtained as follows [19]:

- Difference the speech signal $s[n]$

$$x[n] = s[n] - s[n-1] \tag{3.1}$$

- The differenced speech signal $x[n]$ is passed through a cascade of two ideal zero frequency (digital) resonators, i.e,

$$y[n] = -\sum_{k=1}^{4} a_k y[n-k] + x[n] \tag{3.2}$$

where $a_1$ = -4, $a_2$ = 6, $a_3$ = -4, $a_4$= 1

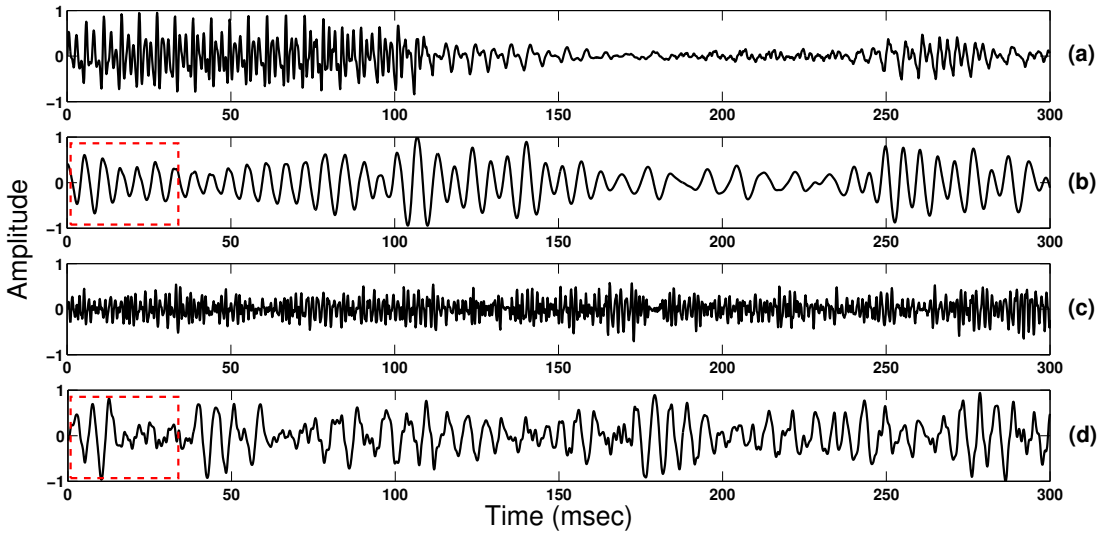


**Figure 3.1:** *(a) Speech signal, (b) ZFFS of speech, (c) Music signal, and (d) ZFFS of music.*

- Remove the trend i.e.,

$$y_1[n] = y[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y[n-k] \tag{3.3}$$

$$\widehat{y}[n] = y_1[n] - \frac{1}{2N+1}\sum_{k=-N}^{N} y_1[n-k] \tag{3.4}$$

  where $2N+1$ corresponds to the average pitch period over a longer segment of speech

- The trend removed signal $\widehat{y}(n)$ is termed as ZFFS.

In this work, the above method is applied to obtain the ZFFS of the audio signal. Figure 3.1 displays the ZFFS plots for speech and music. It may be noted that the nature of the ZFFS for speech and music is different. The periodic nature of the ZFFS is more evident in speech compared to music and is unique for speech. The short term autocorrelation analysis is performed to exploit the differences in the periodic nature of the ZFFS of speech and music. The ZFFS is processed in blocks of 30 ms with 1 ms frame shift. The value of the first peak (after the central peak) in the autocorrelation sequence is an indication of the level of correlation in the frame. The central peak is the peak of the autocorrelation sequence at the origin and is indicated by the arrows in Figure 3.2. The value of the first peak is normalized with the central peak resulting in normalized autocorrelation peak strength (NAPS) feature. The marked rectangles in Figure 3.1 represent the region over which the NAPS is computed. It may be noted from Figure 3.2 (a) and (b) that the NAPS of ZFFS of speech and music

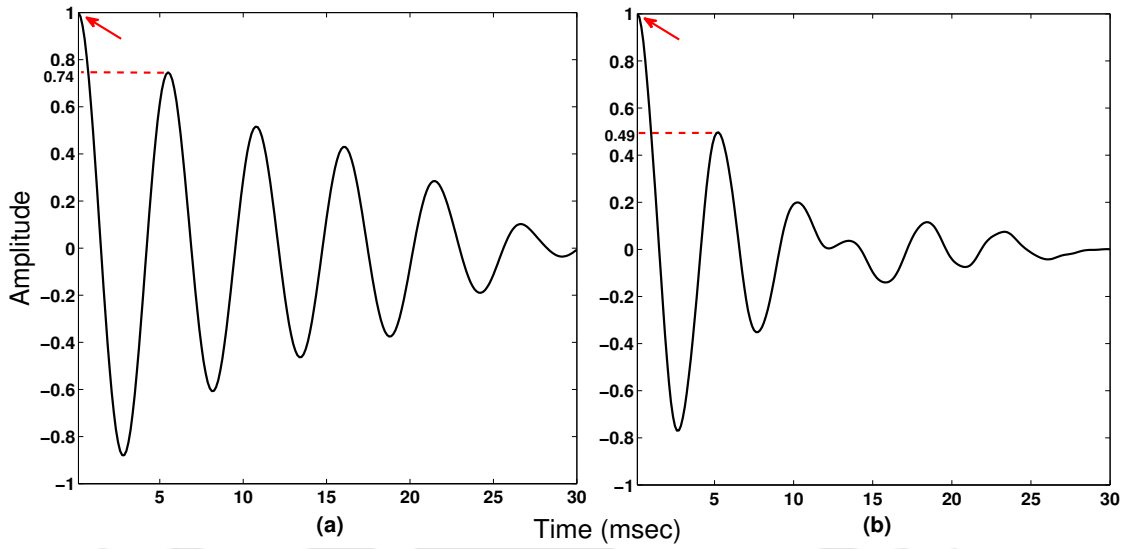


**Figure 3.2:** *Normalized autocorrelation plot for a selected portion of ZFFS of (a) speech and (b) music, respectively. The selected regions are shown in Figure 3.1.*

is 0.74 and 0.49, respectively.

The NAPS is higher for speech compared to music reflecting better the periodic nature of the ZFFS of speech compared to music. The presence of glottal activity in the speech regions gives a nearly periodic nature of the ZFFS and this type of periodic nature may not be present in music regions due to the different glottal activity like action in music compared to speech. This feature was developed as a speech-specific feature, not taking into account the other signals like music. Hence the behavior of this feature in music may be different compared to its behavior in speech as is evident in Figure 3.1.

The reason for the difference is due to the pitch. The extraction of ZFFS may be viewed as a low pass filtering process (DC resonator) followed by a high pass filtering process (trend removal) to form a band pass filter. The center of the band pass filter depends on the average pitch period. The average pitch period considered is around the typical range for speech. Since the bandpass filter is a function of the pitch period of speech, it will result in an extraction of a periodic signal corresponding to the pitch period of speech. In the case of music, the band pass filter will emphasize a portion of the spectral energy around the pitch of speech. This spectral portion does not carry any information relating to the pitch of music since the pitch of music is higher than that of speech in most cases. There may be cases of down-tuned instruments used in modern rock music, where the pitch of music may be lower than that of speech, but such cases are very few in our considered database. The spectral

energy of music emphasized by the band pass filter has a different nature from the spectral energy of speech emphasized by that same filter. This results in a lesser periodic nature of the ZFFS for music compared to speech. Figure 3.6 (b) shows the NAPS of ZFFS of the audio signal sample, Figure 3.6 (a), taken from the Indian broadcast news where the first 5 s of the audio signal corresponds to speech while the next 5 s corresponds to music. It can be seen that the NAPS of ZFFS gives larger values in the speech regions compared to the music regions. Thus NAPS may be used as a feature to discriminate between speech and music regions.

#### 3.2.1.2 Peak-to-Sidelobe Ratio

The LP analysis is a method of extracting the vocal tract and excitation source information in speech [73]. The effect of this analysis on the audio signal is studied in this work, where each frame size of 30 ms is processed with a frame shift of 1 ms. For each block of 30 ms, $10^{th}$ order LP analysis (audio is sampled at Fs = 8 kHz) is performed to estimate the LP coefficients. The audio signal is passed through the inverse filter to extract the LP residual signal. In speech, the time-varying changes in the excitation source characteristics are smeared in the LP residual due to its bipolar nature [74]. These changes are further enhanced by computing the HE of LP residual [74].

The HE ($h_e[n]$) of LP residual ($e[n]$) is defined as [75]

$$h_e[n] = \sqrt{e^2[n] + e_h^2[n]} \tag{3.5}$$

where $e_h[n]$ is the Hilbert transform of $e[n]$,and is given by

$$e_h[n] = \text{IDFT}(E_h[k]) \tag{3.6}$$

where

$$E_h[k] = \begin{cases} -jE[k], k = 0, 1, ..., (\frac{N}{2}) - 1 \\ jE[k], \quad k = \frac{N}{2}, (\frac{N}{2}) + 1, ...., (N-1) \end{cases} \tag{3.7}$$

and IDFT denotes the inverse discrete Fourier transform with $E[k]$ computed as the DFT of $e[n]$, and $N$ is the number of points used for computing DFT.

The HE of LP residual for the audio signal is computed in this work. The HE of LP residual contains peaks corresponding to the excitation source information along with side-lobes around the peak. These peaks are higher in the case of speech compared to music as shown in Figure 3.3 (b) and (d). The side-lobe variation of the speech and music regions is almost the same although in most cases

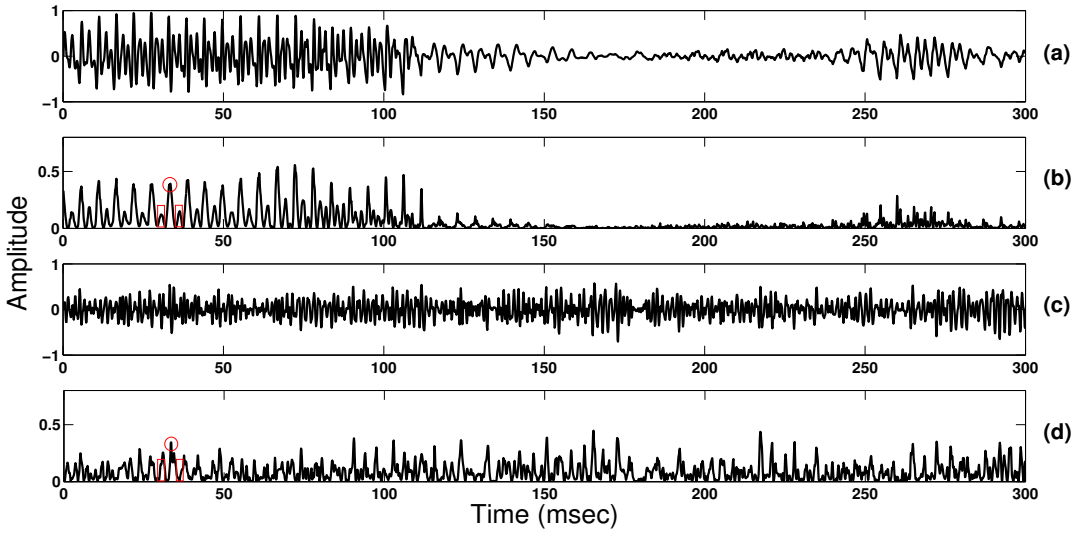


**Figure 3.3:** *(a) Speech signal, (b) HE of LP residual of speech, (c) Music signal and (d) HE of LP residual of music. The marked circles indicate the peaks while the marked rectangles indicate the region over which the side lobe variance is computed.*

this variation is higher in music as compared to speech. This motivated the idea of computing the peak-to-sidelobe ratio (PSR) [68] of the HE of LP residual. The PSR is computed by first obtaining the peaks of HE of LP residual marked as circles in Figure 3.3 (b) and (d). These peaks can be located by searching around the epoch locations obtained from the ZFFS [19]. A frame size of 3 ms around the epoch is considered for searching the peaks and the maximum value of the peaks in that frame is considered as the peak of the HE of LP residual. The side-lobe variance is computed over a frame size of one pitch period, which consists of half pitch period before 4 samples to the left and half pitch period after 4 samples to the right of the peak marked as rectangles approximately in Figure 3.3 (b) and (d). Dividing the peak of HE of LP residual of the speech signal by the side-lobe variance gives the ratio of peak-to-sidelobe. For the marked segments in the Figure 3.3 (b) and (d), the PSR for speech and music, respectively, is 53.34 and 11.76. Thus the speech regions have a higher PSR compared to the music regions as shown in Figure 3.6 (c), where the PSR in the figure has been normalized over the entire duration of the 10 s audio clip.

The HE of LP residual has higher peaks in the speech regions which represent the impulse-like excitation of the glottal source. This impulse-like excitation may not be present in the music signals and is evident from the nature of the HE of LP residual wherein the peaks in the music region are much lower or even absent (Figure 3.3). The main reason for this is attributed to the estimation of

the LP residual signal. In LP analysis, each sample is predicted as a linear weighted sum of past $p$ samples, where $p$ is the order of prediction. The residual which is the difference between the predicted sample and the actual sample gives a high value for the speech signals at the instant of significant excitation. In music, the error is nearly the same throughout the signal since the excitation source for music is different compared to speech. The side-lobe variance of the speech regions also tend to be lower than the music regions due to the noise-like nature of some of the music signals like the one in Figure 3.3. The PSR of HE of LP residual hence has a higher value in the speech regions compared to the music regions and this feature can be used for discriminating speech from music regions.

### 3.2.2 Speech-Specific Vocal Tract System Features

The presence of a majority of high energy vowel-like sounds in speech is exploited to define a feature in terms of the vocal tract system which is described below.

#### 3.2.2.1 Log Mel Spectrum Energy

Speech is produced as a sequence of sound units. These sound units are produced as a result of changes in the vocal tract shape. At a gross level, the sound units can be grouped into vowel and non-vowel-like sounds. Thus there will be continuous change in the vocal tract shape from the production of vowel to non-vowel-like sounds and vice versa. Distinct vocal tract shapes are associated with the production of vowels in speech. The vowel sound units are high energy regions and have most of their energy concentrated in low frequency range ($\leq$ 2.5 kHz). The dominance of vowel-like sound units in speech having energy concentrated in the low frequency range makes it different compared to music components. The energy in the low-frequency range can be represented in terms of the mel filterbank energies. Due to the multiplication of the magnitude spectrum by the mel filterbank and summing the values obtained in each filter, most of the source information is smoothed out while computing mel filterbank energies. Hence the resulting evidence may be treated as a representation of vocal tract shape information. The vocal tract shape is manifested in the log mel spectrum energy values of the speech signal.

In this work, the audio signal is processed in blocks of 30 ms with a shift of 1 ms. For each block of 30 ms, a 512 point DFT is computed to obtain the spectrum of each block. The spectrum is then passed through 22 triangular filters (audio is sampled at Fs = 8 kHz) having central frequencies on the linear scale converted from the evenly distributed central frequencies on the mel scale to obtain

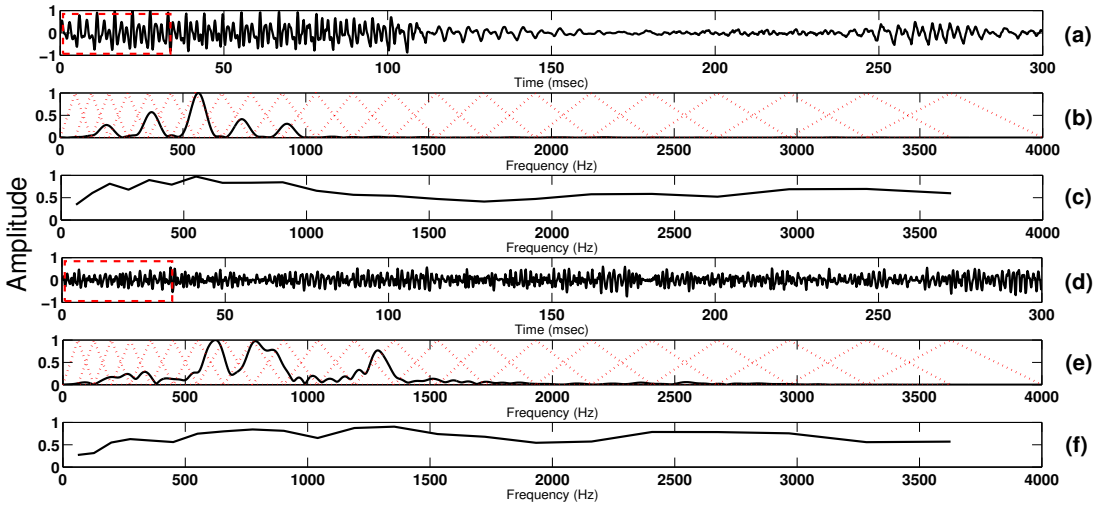


**Figure 3.4:** *(a) Speech signal (b) Fourier transform spectrum of 30 ms (marked as dotted rectangle) of speech (c) Log mel filter energy values of speech (d) Music signal (e) Fourier transform spectrum of 30 ms (marked as dotted rectangle) of music (f) Log mel filter energy values of music. The marked rectangles indicate the regions over which the log mel filter energy values are computed. The marked triangles indicate the distribution of the mel filter banks. The first 18 filters cover the 0 to 2.5 kHz range of frequencies.*

the mel filter energy values. The logarithm of mel-filter energy values is then calculated. The sum of the first 18 log mel filter energy values are computed which covers about 2.5 kHz, the range of first 2 to 3 formant frequencies of the vowel sound units of speech, and this sum represents the log mel spectrum energy. Mathematically this is expressed as:

$$E[i] = \sum_{g=1}^{18} \log \left[ \sum_{k=1}^{M} |S[k,i] f_g[k]|^2 \right] \tag{3.8}$$

where $i$ is the frame number, $g$ is the filter number, $k$ is the frequency bin and $M$ is the total number of bins.

$$S[k,i] = \sum_{n=0}^{N-1} s[n + iR] w[n] e^{\frac{-j2\pi nk}{N}} \tag{3.9}$$

where $s[n]$ is the speech signal, $w[n]$ is a rectangular window, $R$ is the frame shift, and $N$ is the total number of points for computing the Fourier transform. $f_g[k]$ is the triangular filter which has the central frequency $f_{\text{cent}}[g]$ on the linear scale converted from the mel scale as,

$$f_{\text{cent}}[g] = 700(e^{\frac{m}{1125}} - 1) \tag{3.10}$$

where $m$ is the index number in the mel scale and $g$ is the filter number.

Figure 3.4 (c) and (f) represents the log mel spectrum energy values obtained from the 22 triangular filters, which are marked as triangles in the figure for speech and music, respectively. These values have been computed for a frame of speech as well as music which is indicated as rectangles in the figure. It can be seen that for speech, the marked rectangle represents a vowel-like region, which is also evident from the magnitude spectrum. The log mel energy values are high for the lower order mel filters indicating the high energy concentration in the low frequency range for speech segments containing vowel-like regions. There is a continuous change in shapes from vowel to non-vowel-like units production while moving from one frame to the next. The vowel-like regions exhibit higher energy and the non-vowel-like units exhibit lower energy. Thus there is a high variation of the log mel spectrum energy as seen in Figure 3.6 (d). The log mel spectrum energy in the figure has been normalized over the entire duration of the audio clip. Alternatively, for music, the energy distribution is random for a particular frame. The nature of the energy distribution which is evident in vowel-like regions in speech, may not be present in the music regions (Figure 3.4 (f)). There is also a lower variation in the log mel spectrum energy of music as seen in Figure 3.6 (d).

The high variation of the log mel spectrum energy in the speech regions compared to the music regions may be attributed to the fact that the speech regions contain an alternative nature of the high energy vowel-like regions and other types of non-vowel-like regions. This kind of nature may not be present in the music regions. It may be observed that the log mel spectrum energy is related to the first mel frequency cepstral coefficient (MFCC) $c_0$. The first MFCC, $c_0$ is computed by summing the log mel filter energy values of all the filters covering the entire frequency range. However, for the log mel spectrum energy in our work, the sum of the first 18 log mel filter energy values is taken and these filters cover about 2.5 kHz, the range of first 2 to 3 formant frequencies of the vowel sound units of speech. The variation of the log mel spectrum energy is higher for speech compared to music and this variation is unique in the case of speech. The variance in the log mel spectrum energy feature may, therefore, act as a good discriminator for the speech / music classification task.

### 3.2.3 Speech-Specific Modulation Spectrum Features

The syllable rate of speech is exploited to define a feature in terms of the modulation spectrum as described below.

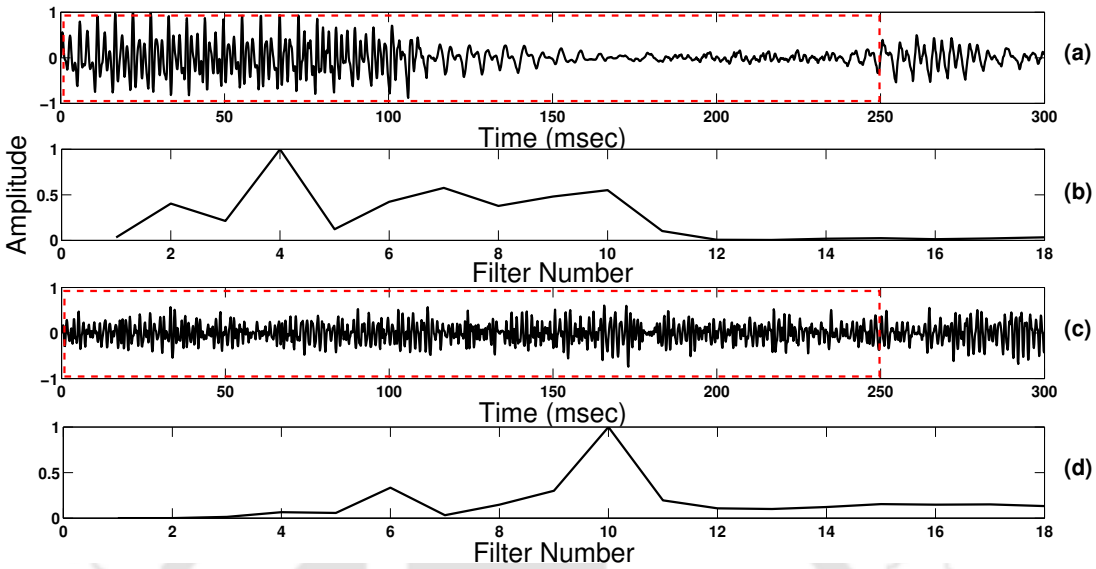


**Figure 3.5:** *(a) Speech signal (b) 4 Hz Modulation spectrum energy from the critical band filters for speech (c) Music signal (d) 4 Hz Modulation spectrum energy from the critical band filters for music. The marked rectangles indicate the regions over which the 4 Hz modulation spectrum energy is computed.*

#### 3.2.3.1 Modulation Spectrum Energy

The slowly varying temporal envelope components in the speech signal generally represent the modulation spectrum [76]. Low frequency components of several Hz mostly constitute the temporal envelope of speech signal. This kind of representation has compelling parallels to the speech production dynamics, where the articulators move at the rates of 2 to 12 Hz [77], and to the sensitivity of auditory cortical neurons to amplitude modulations at the rates below 20 Hz. Several studies have been explored earlier to show the importance of modulation spectrum in speech related tasks [78, 79]. The use of modulation spectrum in speech / music classification has also been explored in [12] which exploits the idea that speech has a characteristic energy peak around the 4 Hz syllabic rate and music does not have this kind of nature. A detailed focus on the modulation spectrum including the development of the modulation spectrogram has been demonstrated in [76, 80].

Given the audio signal, the modulation spectrum energy is computed as follows [76,80]: The audio signal is first analyzed into 18 critical band filters between 0 and 4 kHz frequency band. The filters are generally trapezoidal in shape, and the overlap between adjacent bands is minimum. Half-wave rectification and filtering with a low pass filter having cutoff frequency of 28 Hz are performed in each band to obtain an amplitude envelope signal. Down-sampling of each amplitude envelope signal to 80 samples/s is performed. Each down-sampled amplitude envelope signal is then normalized by the

average envelope level in that channel, measured over the entire audio signal clip. In order to capture the dynamic properties of the signal, the modulations of the normalized envelope signals are analyzed by computing DFT over a Hamming window of length 250 ms with a window shift of 12.5 ms. Finally, the 4 Hz components are summed together, across all critical bands. Mathematically the modulation transfer function energies are expressed as

$$\mathrm{MTF}[m] = \sum_{c=1}^{18} \left[ \left| \widehat{X}_c[k1, m] \right|^2 \right] \tag{3.11}$$

where $m$ is the frame index, $c$ represents the critical band number, and $k1$ represents a frequency index of 4 Hz. $\widehat{X}_c[k, m]$ is computed as

$$\widehat{X}_c[k, m] = \sum_{n=0}^{N-1} \widehat{x}_c[n + mR] w[n] e^{-\frac{j2\pi nk}{N}}; c = 1, 2, ..18. \tag{3.12}$$

where $\widehat{x}_c[n]$ represents the normalized envelope of $c^{th}$ filter output, $w[n]$is a Hamming window, $R$ is the frame shift and $N$ is the number of points used for computing the DFT. The modulation energy components computed are up-sampled to 8000 samples/s.

The distribution of the 4 Hz modulation energy is shown in Figure 3.5 (b) and (d) for speech and music, respectively, computed for a frame of speech and music, shown as rectangles in the figure. It can be clearly observed that there is higher energy at the 4 Hz frequency in speech compared to music. Figure 3.6 (e) shows the plot of the 4 Hz modulation spectrum energy, where its values have been normalized over the entire duration of the audio clip. The 4 Hz modulation spectrum energy feature represents the slowly varying temporal envelope corresponding to the speech syllabic rate. Speech is more characterized by the 4 Hz syllable rate compared to music and hence this feature is higher in speech regions compared to the music regions. The 4 Hz modulation spectrum energy has been used for the speech / music classification task in earlier work. It has been used in the current work, since it represents the long term aspect of speech which is different from the source and the system.

So far, the behavior of the features for a single frame and for a single audio file has been described. Their behavior over a larger number of frames computed over all the audio files from the Scheirer and Slaney database [12] is seen by the histogram plot in Figure 3.8, where the thick line indicates speech and the dashed line indicates music. It can be seen that there is higher separability in the speech and music distribution for the log mel energy variance feature with minimum overlap than the rest of the speech-specific features, although there is visible separation for the other speech-specific features as

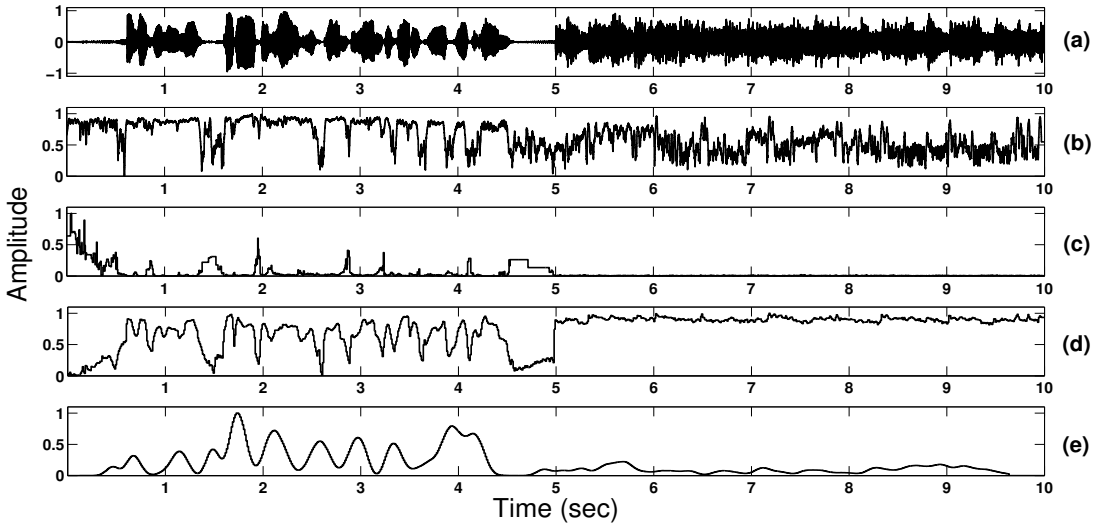


**Figure 3.6:** *(a) Audio signal, where the first 5 s correspond to speech and the next 5 s correspond to music (b) NAPS of ZFFS (c) PSR of HE of LP residual (d) Log mel energy (e) 4 Hz Modulation spectrum energy*

well. The nature of the histogram for the log mel spectrum energy variance and the PSR of HE of LP residual is similar to the existing features. The nature of the histogram for the NAPS mean and modulation spectrum mean is different compared to the other features.

## 3.3 Overall Speech/Music Classification System

The previous section described the different speech-specific features that are considered for speech / music classification. The illustrations indicated that each of the features indeed shows discrimination between speech and music regions. The evidence from each of these features needs to be effectively combined for speech / music classification. The explorations for the same are presented in this section.

### 3.3.1 Speech/Music Classification by Non-linear Mapping and Combining

The speech / music classification task involves assigning a particular label to speech and music. In this work, speech is given a label as *one* and music as *zero*. With this objective in mind, smoothing and non-linear mapping of the features is performed. Ideally, the value of the feature is mapped to one for speech and zero for music, hence performing the classification task. It can be seen in Figure 3.6, the NAPS of ZFFS, PSR of HE of LP residual, and modulation spectrum have mostly high values in the speech regions compared to music regions. However, there are some speech regions in which the feature values may be lower than the music regions which are categorized to be spurious. This

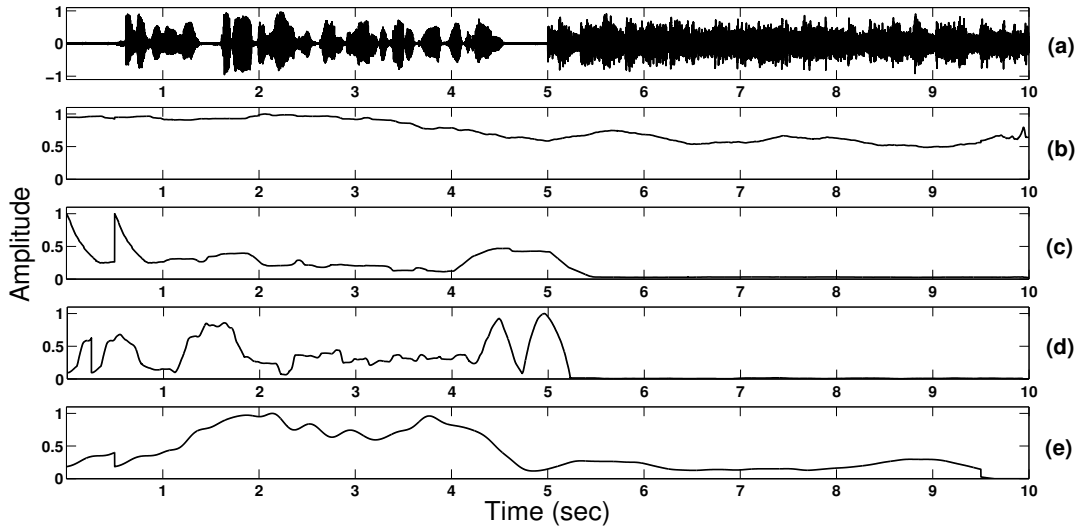


**Figure 3.7:** *(a) Audio signal, Smoothed (b) NAPS of ZFFS (c) PSR of HE of LP residual (d) Log mel energy and (e) 4 Hz Modulation spectrum energy. The smoothed contours have been computed from the features in Figure 3.6*

spurious can be reduced by smoothing the features. It may be noted that the above features have been computed for a window size of 30 ms with a shift of 1 ms. Before the smoothing process, interpolation of the missing samples is required. The interpolation process is performed by duplicating the single value obtained for every frame to the missing samples caused by the frame shift to the next frame. A small shift has been chosen to reduce the number of samples required for interpolation. If a larger shift is chosen, more samples need to be interpolated and may affect the accuracy of the smoothing process and thereby reduce the overall accuracy. The mean over 1 s frame is computed with every sample shift for the NAPS, PSR, and modulation spectrum and the smoothed values are shown in Figure 3.7. The variation of the interpolation method does not significantly change the final smoothed value of the features. Even the standard linear interpolation method results in a very similar smoothing effect on the features as the interpolation method mentioned earlier. However, the linear interpolation method is not done here since it is computationally more intensive than the interpolation method followed in this work.

For the log mel spectrum energy, the variance over 1 s frame for every sample shift is computed for smoothing. It can be observed from Figure 3.6(d), the variation of the feature in the speech regions is very high compared to the music regions. The window size is experimentally chosen for the best values for computing smoothed mean and variance contour. It was observed that there is not much difference

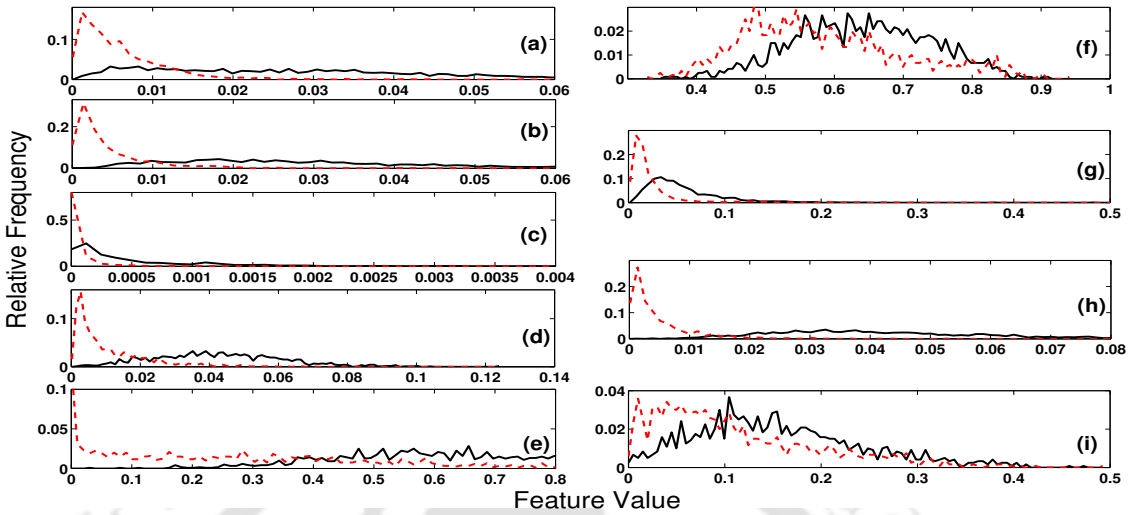


**Figure 3.8:** *Histogram plot for (a) ZCR variance (b) Spectral centroid variance (c) Spectral flux variance (d) Spectral roll-off variance (e) Percentage of low energy frames (f) NAPS of ZFFS Mean (g) PSR of HE of LP residual mean (h) Log mel energy variance (i) 4 Hz Modulation spectrum energy mean. Note that the continuous line represents speech and the dashed line represents music.*

in the mean and variance contours while smoothing with window sizes in the range of 500 ms to 1200 ms. If the window size is chosen beyond this range, severe degradation in the smoothed contours is observed. It can be seen that the features having values lower in the speech regions compared to the music regions shown in Figure 3.6, are now having their values smoothed to their nearest higher values as shown in Figure 3.7, thus reducing spurious.

The smoothed evidence is then mapped using the non-linear mapping function given by

$$P_m = \frac{1}{1 + e^{-(P_s - \Theta)/\tau}} + \alpha \tag{3.13}$$

where, $P_m$ is non-linearly mapped value, $P_s$ is the smoothed evidence value, $\Theta$, $\tau$ are the slope parameters and $\alpha$ is the offset which is the minimum value of the function. The values of $\tau$ and $\alpha$ are set to 0.001 and 0, respectively, since the binary mapping of either 0 or 1 is required. The main tunable parameter is $\Theta$ which is set experimentally, and this value is kept in the range of 0.3 to 0.6 based on the experiments performed on the databases. The overall performance does not change significantly if the value of $\Theta$ is varied in this range. The non-linearly mapped plots are shown in Figure 3.9 (b)-(e), where it is seen that nearly all the speech regions have a value of one and the music regions have a value of zero.

A classification framework is presented, wherein these non-linear mapped values are combined

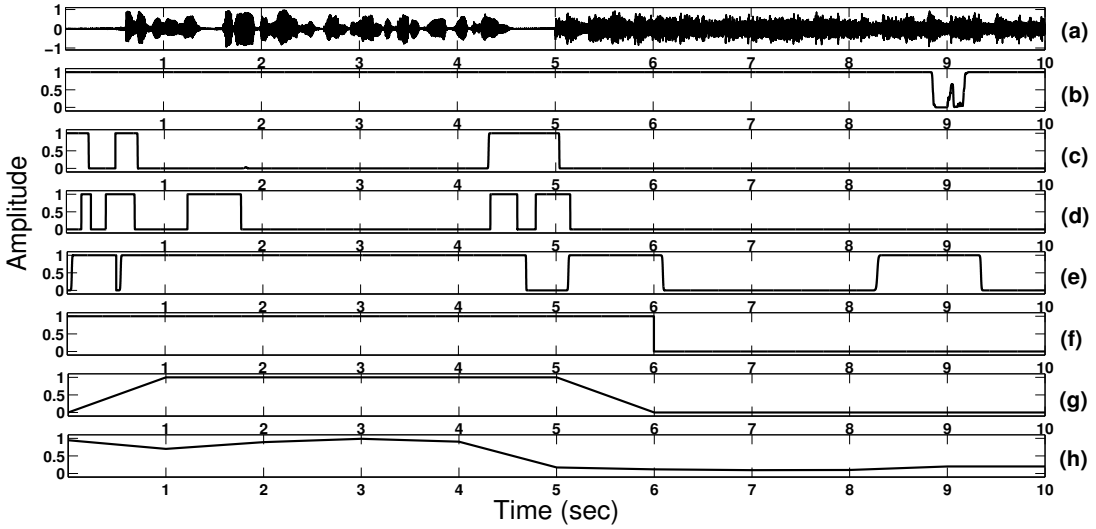


**Figure 3.9:** *(a) Audio signal, non-linear mapped value of smoothed (b) NAPS of ZFFS (c) PSR of HE of LP residual (d) Log mel energy (e) 4 Hz Modulation spectrum energy. The non-linear mapped values have been computed from the smoothed contours in Figure 3.7. Classification result using (f) non-linear mapping (g) Gaussian Mixture Models (GMM) (h) Support Vector Machines (SVM)*

by summing them together to produce an evidence for the speech / music classification task. The summed evidence is next non-linearly mapped using the same mapping function and a fixed threshold of 0.4. After that, a 1 s window with a non-overlapping 1 s shift of the audio segment is considered and the mean is computed. This window size is chosen in most of the segment-based speech / music classification task [12,30]. If the value of the mean is greater than a threshold which is 0.08 (this value is not crucial for the task), the segment is classified as speech, otherwise, it is classified as music. The final result of the classification is shown in Figure 3.9(f) where all the 5 speech segments are classified correctly whereas only 4 out of 5 music segments are classified correctly. Misclassification of 1 music segment is observed in the figure.

It can be seen that the non-linear mapping technique requires thresholds to be defined. However, the advantage of this method is that there is no requirement of training data as required by the classifiers. This method can also be described as the signal processing approach for classification, where the accuracy will be decided by the linear separability of the speech-specific features. This kind of signal processing approach has been followed in other tasks like voiced / unvoiced detection [81]. Since those tasks are similar to the speech / music classification task, we have employed this kind of approach for classification in this work.

The non-linear mapping technique gives an overall improved accuracy. To illustrate this, the classification is directly performed on the smoothed values. The smoothed log mel spectrum energy value of Figure 3.7(d) is taken as an example and its mean is computed for 1 s window with a non-overlapping 1 s shift. If the mean is greater than a threshold of 0.5, the segment is classified as speech, otherwise, it is classified as music. Similarly, the classification on the non-linear mapped value of this feature is performed by computing the mean of the non-linear mapped value ($\Theta$=0.5) of the log mel spectrum energy shown in Figure 3.9(d), compared the mean to a threshold of 0.08 as earlier, and the same kind of segment classification is performed. Overall accuracies of 68.46 % and 74.34 %, respectively, are obtained for the two cases, on the Broadcast News database (to be described later), indicating the significance of using the non-linear mapping technique.

### 3.3.2 Speech/Music Classification using Gaussian Mixture Models and Support Vector Machines

Gaussian mixture models (GMMs) have been explored earlier for the speech / music classification task [12]. The models are trained for the speech and music signals by using the expectation maximization algorithm. A new feature is assigned to a particular model which has a higher likelihood estimate. Diagonal covariances for the GMM have been used in this work.

Support vector machines (SVMs) are well suited for binary classification tasks and have shown considerable success in a variety of domains. The use of SVMs for speech / music classification has been explored in [82, 83]. All the experiments using SVM in this work, were carried out using the libSVM [84] with a radial-basis function (RBF) kernel of the form,

$$K(x, y) = exp(-Y||x - y||^2) \tag{3.14}$$

The classification result using GMM and SVM for the audio file in Figure 3.9 (a) is shown in Figure 3.9 (g) and (h), respectively. The statistics of the raw speech-specific features shown in Figure 3.6 are computed for a window size of 1 s with a non-overlapping shift of 1 s [12, 30]. These statistics of the features are concatenated to form a feature vector. This feature vector is given as input to the classifiers. It can be seen that using GMM, the first segment of speech has been misclassified and all the other remaining segments are classified correctly whereas using SVM all the speech and music segments have been classified correctly. The classifiers have been trained with a particular database and the details of the training process will be given in the next section.

## 3.4 Results and Discussion

The proposed method for speech / music classification is first evaluated on a database which has been recorded at random from the radio during the summer of 1996 by Scheirer and Slaney [12] which will be referred to as Scheirer and Slaney (S&S) database. The training examples taken from this database include 80 files of speech and 80 files of music without vocals which are of 15 seconds each. Another database evaluated in this work is the GTZAN database which has been explored in [30, 85]. This database contains 64 files of speech and 64 files of non-vocal music each of length 30 seconds. The audio data in both of the databases has a sampling frequency of 22050 Hz and has been down-sampled to 8000 Hz for the task. The evaluation is also done on the database containing audio data recorded from the Indian broadcast news channels having a total of 104 files of speech and 104 files of non-vocal music each length of 5 seconds with 8000 Hz sampling frequency.

### 3.4.1 Non-linear Mapping and Combining

First, the results using non-linear mapping of the speech-specific features are shown in Table 3.1. For the GTZAN database, the performance in music is higher, however, for the case of S&S and Broadcast news database, the performance in speech is higher. This depends on the threshold. However, varying the threshold does not change the overall performance for the three databases to larger extent. The results shown in the table are evaluated for a threshold ($\theta$ of the non-linear mapping function) of 0.5 for all the speech-specific features.

**Table 3.1:** *Results using non-linear mapping in terms of classification accuracy (%).*

| **Features →** | **Speech-Specific Features** | | |
|---|---|---|---|
| **Database ↓** | **Speech** | **Music** | **Overall** |
| **S&S** | **84.33** | **64.58** | **74.45** |
| **GTZAN** | **70.26** | **78.80** | **74.53** |
| **Broadcast News** | **96.34** | **60.96** | **78.65** |

**Table 3.2:** *Performance in terms of classification accuracy (%) using the different individual features on the Scheirer and Slaney (S&S) database and the GTZAN database. In the table, the abbreviations, GMM indicates the Gaussian mixture model classifier and SVM indicates support vector machines. For the different features, the statistics are computed on the raw features and not on the smoothed features.*

| Database → | S&S database | | | | | | GTZAN database | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier → | GMM | | | SVM | | | GMM | | | SVM | | |
| Features ↓ | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall |
| ZCR Var. | 74.25 | 90.25 | 82.25 | 70.00 | 93.91 | 81.95 | 75.78 | 72.81 | 74.29 | 58.12 | 85.62 | 71.87 |
| Spec. Centroid Var. | 89.91 | 87.25 | 88.58 | 86.00 | 91.50 | 88.75 | 84.94 | 80.15 | 82.55 | 72.03 | 89.68 | 80.85 |
| Spec. Flux Var. | 80.91 | 80.66 | 80.79 | 55.91 | 92.00 | 73.95 | 68.33 | 70.36 | 69.34 | 48.59 | 85.00 | 66.79 |
| Spec. Roll-off Var. | 85.75 | 85.58 | 85.66 | 85.50 | 85.91 | 85.70 | 78.38 | 79.53 | 78.95 | 74.16 | 83.75 | 78.95 |
| Percent. of Low Energy Frames | 83.75 | 77.75 | 80.75 | 88.58 | 72.25 | 80.41 | 78.17 | 77.08 | 77.63 | 83.54 | 72.23 | 77.89 |
| NAPS Mean | 69.08 | 56.00 | 62.54 | 73.75 | 52.83 | 63.29 | 61.77 | 56.09 | 58.93 | 62.39 | 58.07 | 60.23 |
| PSR Mean | 82.50 | 72.16 | 77.33 | 74.66 | 80.00 | 77.33 | 74.37 | 67.55 | 70.96 | 57.39 | 80.05 | 68.72 |
| Log Mel Spec. Energy Var. | 94.75 | 94.58 | 94.66 | 94.08 | 94.91 | 94.50 | 85.62 | 85.46 | 85.54 | 82.96 | 88.33 | 85.65 |
| Modulation Spec. Energy Mean | 71.58 | 53.66 | 62.62 | 69.25 | 56.58 | 62.91 | 45.78 | 63.33 | 54.55 | 29.94 | 83.33 | 56.64 |

### 3.4.2 Classifiers

The use of thresholds for the task may not give optimal performance. Hence classifiers like GMM and SVM are used for the classification task on the speech-specific features. The width parameter $Y$ and the cost parameter $c$ of the SVM as well as the mixture $k$ of GMM is varied to achieve optimal performances. The cost parameter $c$ is set to 1 and the width parameter $Y$ is set to 3 in this work. The number of mixtures for the GMM has been set to $k = 8$. The SVM parameters ($c = 1, Y = 3$) and the number of mixtures of the GMM ($k = 8$) have been fixed at their optimal values based on the results for the test data across the different databases. These parameters have been fixed for all the features and across different databases to show the impact of the speech-specific features for the classification task. The existing features like the ZCR, spectral centroid, spectral flux and spectral roll-off as well as the percentage of low energy frames are also considered for evaluation in order to compare the performances of the speech-specific features. The statistics of speech-specific features, as well as the existing features, are computed using a window size of 1 s with a non-overlapping shift of 1 s. As mentioned in Section 3.3.2, the statistics are computed on the raw features and not on the smoothed features. The individual features are evaluated and their performances are shown in Table 3.2 and Table 3.3. A 4-fold cross-validated scheme was used for evaluation with separate files in

**Table 3.3:** *Performance in terms of classification accuracy (%) using the different individual features on the Broadcast News (BN) database. In the table, the abbreviations, GMM indicates the Gaussian mixture model classifier and SVM indicates support vector machines. For the different features, the statistics are computed on the raw features and not on the smoothed features.*

| Database → | BN database | | | | | |
|---|---|---|---|---|---|---|
| Classifier → | GMM | | | SVM | | |
| Features ↓ | Speech | Music | Overall | Speech | Music | Overall |
| ZCR Var. | 51.15 | 84.42 | 67.78 | 60.00 | 79.80 | 69.90 |
| Spec. Centroid Var. | 89.03 | 76.53 | 82.78 | 77.30 | 84.42 | 80.86 |
| Spec. Flux Var. | 67.30 | 69.61 | 68.46 | 57.50 | 81.53 | 69.51 |
| Spec. Roll-off Var. | 89.03 | 85.76 | 87.40 | 87.11 | 88.26 | 87.69 |
| Percent. of Low Energy Frames | 90.38 | 80.57 | 85.48 | 90.38 | 80.00 | 85.19 |
| NAPS Mean | 69.23 | 73.07 | 71.15 | 72.69 | 71.15 | 71.92 |
| PSR Mean | 81.53 | 78.07 | 79.80 | 75.76 | 85.38 | 80.57 |
| Log Mel Spec. Energy Var. | 92.50 | 82.69 | 87.59 | 88.65 | 85.19 | 86.92 |
| Modulation Spec. Energy Mean | 77.69 | 59.42 | 68.55 | 65.00 | 70.57 | 67.78 |

training and testing datasets. It can be observed that the variance of log mel spectrum energy feature, which is the speech-specific feature representing the vocal tract system shows superior performance on all the three databases. The existing features like the variance of spectral centroid, variance of spectral roll-off and percentage of low energy frames also show good performances.

It is interesting to see that individually, most of the existing features show better performances than the source and modulation spectrum features which belong to the category of speech-specific features. However, on combining the speech-specific features, we get a better performance than combining the existing features. This is reflected in the $4^{th}$ and $5^{th}$ row of Table 3.4, where the feature combination is performed by concatenating the features together to form a feature vector and fed as input to the classifiers. The reason for the better performance of the combined speech-specific features could be due to the capturing of complementary information by each of the speech-specific features since these features represent the different aspects of speech production. The existing features, being general audio features may not be able to characterize the speech regions as much as the speech-specific features.

It can also be seen from Table 3.1 and 3.4, that the overall performance of speech-specific features increased gradually from 74.45 % when using threshold based approach to 95.12 % using GMM and finally 95.87 % using SVM on the S&S database. Similar trends are observed for the GTZAN and

**Table 3.4:** *Performance in terms of classification accuracy (%) using the existing, speech-specific and combined set of features on the Scheirer and Slaney (S&S) database, the GTZAN database and the Broadcast News (BN) database. In the table, the abbreviations, GMM indicates the Gaussian mixture model classifier and SVM indicates support vector machines.*

| Database → | S&S database | | | | | | GTZAN database | | | | | | BN database | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Classifier → | GMM | | | SVM | | | GMM | | | SVM | | | GMM | | | SVM | | |
| Features ↓ | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall | Speech | Music | Overall |
| Existing | 91.58 | 87.16 | 89.37 | 90.50 | 89.83 | 90.16 | 87.70 | 81.66 | 84.68 | 87.29 | 88.43 | 87.86 | 90.76 | 81.53 | 86.15 | 92.11 | 89.42 | 90.76 |
| Speech-Specific | 95.08 | 95.16 | 95.12 | 95.08 | 96.66 | 95.87 | 89.68 | 84.01 | 86.84 | 88.64 | 87.55 | 88.09 | 89.61 | 85.38 | 87.50 | 92.88 | 88.07 | 90.48 |
| Combined | 96.91 | 94.66 | 95.79 | 96.91 | 96.58 | 96.75 | 91.77 | 87.08 | 89.42 | 93.48 | 90.57 | 92.03 | 91.02 | 90.00 | 90.51 | 93.26 | 91.34 | 92.30 |

Indian broadcast news database. On combining the existing features with the speech-specific features, the best performances are obtained and are shown in the last row of Table 3.4. The best overall performance obtained is for the S&S database, which is 96.75 % using SVM. The earlier work [12] on this database showed the best overall performance of 94.2 %. Similarly, for the GTZAN database, the best performance obtained is 92.03 % which is comparable to the best performance reported in [30] which is 93.5 %.

From Table 3.4 it can be observed that the difference of the performances in speech and music is higher, for the existing features compared to the speech-specific features when using GMM. This reflects the inability of the existing features to reduce the confusion between speech and music. The speech signal is more controlled in terms of its production and hence the signal characteristic of speech is similar. The music signal has a complex nature which may be produced by different instruments and some characteristics of the music signal may be similar to speech. The existing features are able to characterize the speech segments to a certain extent due to the similar signal characteristics of speech. However, since the existing features are not derived from the speech-specific knowledge, they are not able to discriminate the speech like music segments and tend to describe these segments as speech. On the other hand, the speech-specific features which represent the source, vocal tract system and syllabic rate aspects of speech are able to capture the speech segments of the audio signal successfully. For those music segments which have a nature similar to speech, the speech-specific features deviate significantly from their normal behavior since those speech-like music segments may not be completely described by the speech-specific features, thereby reducing the confusion between speech and music.

**Table 3.5:** *Level of Canonical Correlation*

| NAPS of ZFFS | | | PSR of HE of LP | | |
|---|---|---|---|---|---|
| Existing | Speech-specific excluding NAPS of ZFFS | All excluding NAPS of ZFFS | Existing | Speech-specific excluding PSR of HE of LP | All excluding PSR of HE of LP |
| 0.1318 | 0.1331 | 0.2040 | 0.5508 | 0.5090 | 0.5895 |
| Log Mel | | | Modulation Spectrum energy | | |
| Existing | Speech-specific excluding Log Mel | All excluding Log Mel | Existing | Speech-specific excluding Modulation Spectrum energy | All excluding Modulation Spectrum energy |
| 0.7773 | 0.5165 | 0.7952 | 0.2797 | 0.3267 | 0.4003 |

### 3.4.3 Canonical Correlation Analysis (CCA)

Canonical-correlation analysis finds vectors $a$ and $b$ in such a way that the random variables $a'X$ and $b'Y$ maximize the correlation $\rho = corr(a'X, b'Y)$. The random variables $U = a'X$ and $V = b'Y$ represent the first pair of canonical variables. Then one finds vectors which maximize the same correlation subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; this gives the second pair of canonical variables. This procedure is continued up to $\min\{m, n\}$ times.

In order to measure the correlation of each speech-specific feature to the combined cases, canonical correlation analysis (CCA) is performed, initially between each speech-specific feature with the existing features consisting of spectral flux, spectral centroid, spectral roll-off, zero crossing rate and percentage of low energy frames. Next, CCA is performed with the other speech-specific features. Finally, CCA is performed with the overall set of features consisting of the existing and the speech-specific features. The result of this analysis is shown in Table 3.5. This analysis was performed for the features computed on the S&S database. In the Table 3.2 and 3.3, it shows that the performance of the log mel energy variance feature is best individually than the other speech-specific features. However, CCA analysis shows that its value is greater than the other speech-specific features. This means that it is more correlated to the overall combined set of features than the other speech-specific features. CCA analysis also shows that the NAPS of ZFFS mean feature is the most uncorrelated feature to the combined

set of features, followed by the PSR of HE of LP residual mean and the modulation spectrum mean features. This shows that the speech-specific features are mostly uncorrelated and combine effectively for the speech / music classification task.

**Table 3.6:** *Performance in terms of classification accuracy (%) of first three features on the three databases using SVM classifier*

| Database → | S&S database | | |
|---|---|---|---|
| Features ↓ | Speech | Music | Overall |
| Log Mel+NAPS Mean | 95.33 | 96.75 | 96.04 |
| Log Mel+NAPS Mean+Spec. Roll-off Var. | 96.25 | 97.25 | 96.75 |
| Database → | GTZAN database | | |
| Features ↓ | Speech | Music | Overall |
| Log Mel+Spec. Centroid Var. | 85.93 | 90.10 | 88.02 |
| Log Mel+Spec. Centroid Var.+ PSR Mean | 91.04 | 87.44 | 89.24 |
| Database → | BN database | | |
| Features ↓ | Speech | Music | Overall |
| Log Mel+Spec. Roll-off Var. | 93.46 | 88.46 | 90.96 |
| Log Mel+Spec. Roll-off Var.+ NAPS Mean | 92.69 | 90.19 | 91.44 |

### 3.4.4 Feature Selection

An experiment is performed to find the minimum subset of features having performances close to the combined case in Table 3.4. Table 3.6 shows the result of the subset of those features. Since the log mel energy variance feature performs the best, this is used as the base feature. For the S&S database, the addition of NAPS mean provides the best additive improvement compared to the other features followed by the addition of the spectral roll-off variance feature. The performance saturates after the addition of the third feature. Similarly, for the other databases, the subset of features giving good performances are shown in Table 3.6.

### 3.4.5 Mismatched Training and Testing data

In order to study the performances of the mismatched training and testing data cases, an experiment is performed which involves one database as the training set and the other database as the

**Table 3.7:** *Performance in terms of classification accuracy (%) on the Broadcast News database using the models trained on the GTZAN database and S&S database.*

| Broadcast News Test | Classifier → | GMM | | | SVM | | |
|---|---|---|---|---|---|---|---|
| Models Trained on | Features ↓ | Speech | Music | Overall | Speech | Music | Overall |
| GTZAN database | Existing | 94.23 | 70.19 | 82.21 | 85.76 | 75.57 | 80.67 |
| GTZAN database | Speech-Specific | 93.65 | 79.03 | 86.34 | 95.00 | 80.57 | 87.78 |
| GTZAN database | Combined | 97.30 | 77.11 | 87.21 | 87.11 | 82.30 | 84.71 |
| S&S database | Existing | 93.65 | 75.00 | 84.32 | 92.11 | 83.07 | 87.59 |
| S&S database | Speech-Specific | 94.03 | 81.73 | 87.88 | 93.46 | 84.23 | 88.84 |
| S&S database | Combined | 95.96 | 80.19 | 88.07 | 94.42 | 84.23 | 89.32 |

testing set. Table 3.7 shows the results of testing the broadcast news data on models trained on the GTZAN as well as the S&S database. The performances of the combined speech-specific features are better than the combined existing features. When the speech-specific and the existing features are combined, the best performance is obtained. Hence similar trends in the results are obtained even for the mismatched training and testing data.

### 3.4.6 Analysis on Vocal Music

An analysis of the behavior of the speech-specific features for the vocal music is briefly discussed here. The vocal music considered here involves singing mixed in with musical instruments. Figure 3.10 shows the behavior of the speech-specific features for an audio signal which consists of speech for the first 5 s and vocal music for the next 5 s. It can be seen that there is some kind of discrimination between speech and vocal music especially for the NAPS of ZFFS, PSR of HE of LP residual, and log mel spectrum energy. The present work mostly focuses on the discrimination between speech and non-vocal music to understand the behavior of the speech-specific features and their discrimination for speech and non-vocal music regions. The work can be extended to the task of discriminating speech against vocal music. In particular, the behavior of the features for the vocal music segments which contain singing (with or without the mixing of musical instruments) can be explored in detail. Figure 3.10 shows that there is potential for exploration of this case in the future.

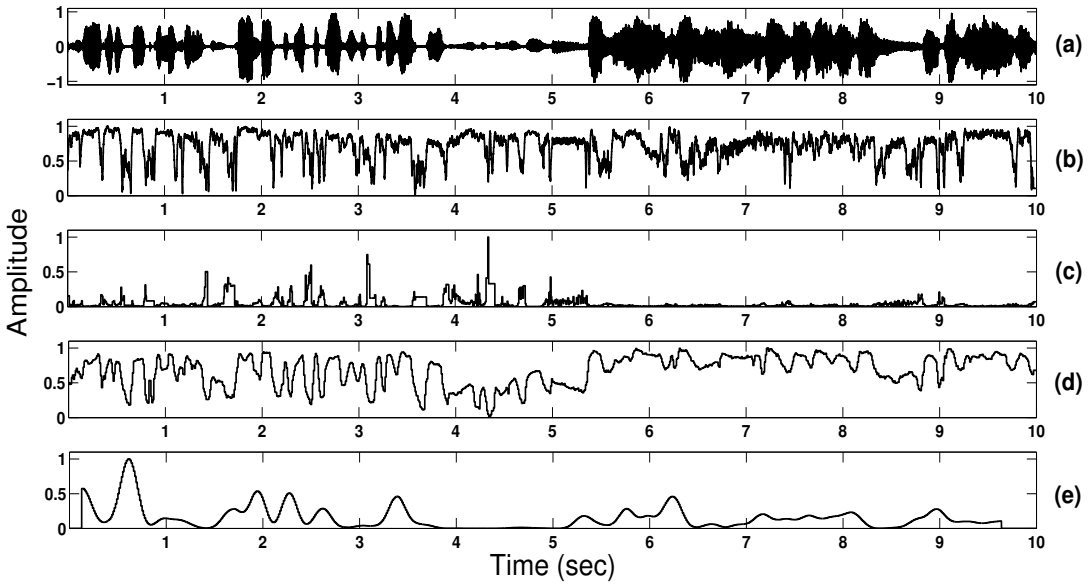


**Figure 3.10:** *(a) Audio signal, where the first 5 s correspond to speech and the next 5 s correspond to vocal music (b) NAPS of ZFFS (c) PSR of HE of LP residual (d) Log mel energy (e) 4 Hz Modulation spectrum energy*

## 3.5 Summary

The use of the speech-specific features for the task of speech / music classification is explored. The NAPS of ZFFS, PSR of HE of LP residual, log mel spectrum energy, and modulation spectrum energy are considered as speech-specific features. The behavior of each feature is studied independently to demonstrate its potential for speech / music classification. Non-linear mapping of the features is done initially for the speech / music classification task. The speech-specific features are then classified using GMM and SVM. Their performances on the S&S database, GTZAN database, and the Indian broadcast news database are tabulated. The performance of the combined speech-specific features has been compared to the combined existing features where it is observed that the performance of combined speech-specific features is better. The existing features are then combined with the speech-specific features for the speech / music classification task and the best performance is achieved with this feature combination. Similar trends in the performances were obtained when testing the broadcast news database on the models trained with either the S&S or the GTZAN database.

# 4

# Clean Speech/Speech with Background Music Classification using HNGD spectrum

Contents

### Objective

*This work explores the characteristics of speech in terms of the spectral characteristics of vocal tract system for deriving features effective for clean speech and speech with background music classification. A representation of the spectral characteristics of the vocal tract system in the form of Hilbert envelope of the Numerator of Group Delay (HNGD) spectrum is explored for the task. This representation complements the existing methods of computing the spectral characteristics in terms of the temporal resolution. This spectrum has an additive and high resolution property which gives a better representation of the formants especially the higher ones. A feature is extracted from the HNGD spectrum which is known as the spectral contrast across the sub-bands and this feature essentially represents the relative spectral characteristics of the vocal tract system. The vocal tract system is also represented approximately in terms of the mel frequency cepstral coefficients (MFCCs) which represent the average spectral characteristics. The MFCCs and the sum of the spectral contrast on HNGD can be used as features to represent the average and relative spectral characteristics of the vocal tract system, respectively. These features complement each other and can be combined in a multidimensional framework to provide good discrimination between clean speech and speech with background music segments. The spectral contrast on HNGD spectrum is compared to the spectral contrast on Discrete Fourier Transform (DFT) spectrum, which also represents the relative spectral characteristics of the vocal tract system. It is observed that better performances are achieved on the HNGD spectrum than the DFT spectrum. The features are classified using classifiers like Gaussian Mixture Models (GMM) and Support Vector Machines (SVM).*

## 4.1 Introduction

There are several tasks which involve the processing of broadcast news. Some of the literature which mentions about the tasks involving the broadcast news processing can be found in [6–10, 22, 86, 87]. The common task in some of these works is the automatic transcription of broadcast news. Generally, there is an involvement of certain steps of preprocessing before the final speech recognition step owing to the complex scenarios present in broadcast news. The preprocessing task is required to account for the different variabilities present in the broadcast audio data. The task of classifying clean speech from speech with background music is one such preprocessing step. This task has not been explored

much in previous literature. However, there are some tasks in which these classes have appeared like in the audio classification tasks in [31, 88, 89]. The reason for choosing these two classes explicitly in our work is described as follows.

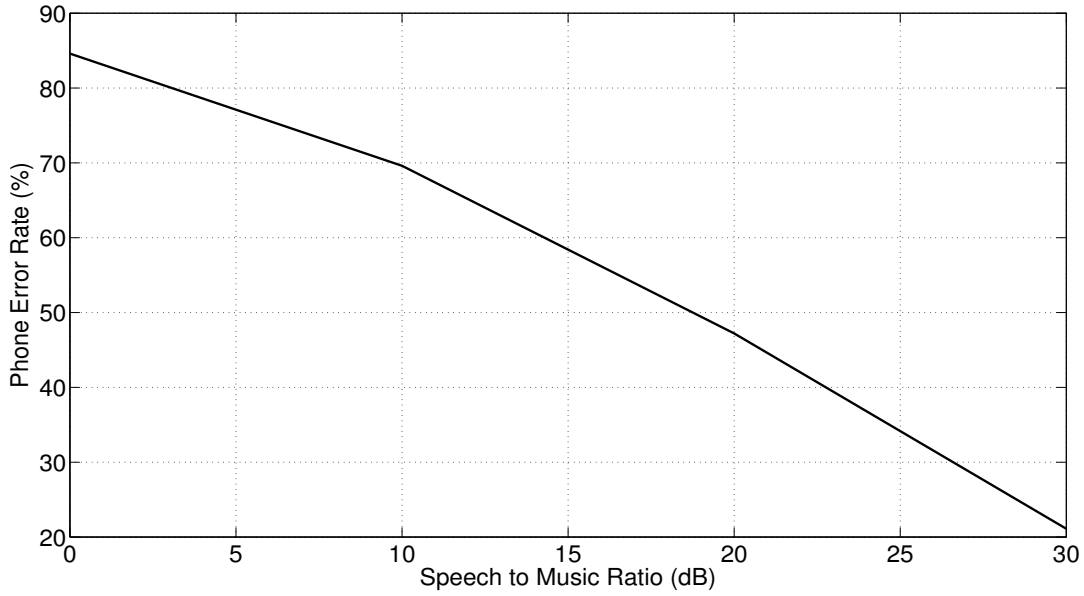


**Figure 4.1:** *Figure showing the degradation in the performance of the phone recognizer when rock music is added to speech TIMIT database samples. The rock music has been taken from the GTZAN database.*

Assuming that there are models for the phone recognition trained on the clean speech data, such models will work well only for the clean speech case. For the speech with background music case (news headlines and voice over in broadcast news), such models may not produce a required level of phone recognition accuracy and this can be seen in Figure 4.1, where it is observed that the addition of music to clean speech causes degradation in the overall phone recognition accuracy. The Phone Error Rate (PER) increases with the decrease of the speech to music (SMR) ratio. In order to properly transcribe the speech with background music data, either models for this case need to be trained or some form of enhancement may be required before passing the speech with background music through the phone recognition system having models built on the clean speech case. Either way, the speech with background music portions need to be segmented so that models may be created using them or enhancement of these portions may be performed. There may also be cases where the majority of the broadcast news data consists of clean speech along with a small percentage of speech with background music data. This small percentage of the data may not be required for further transcription and can be discarded. Based on the above, there may be a requirement to separate the clean speech from

the speech with background music. This motivated to explore the clean speech versus speech with background music classification technique.

As mentioned earlier, this task has been rarely explored, as a result, the features that will be defined in this work will be mostly looked at in terms of the speech characteristics. Speech is produced by exciting a time varying vocal tract system by a time varying source. If the music had been produced simultaneously along with the speech, some characteristics of the vocal tract system may have been affected. However, most of the time the music is added to speech at particular dB level after the speech is produced, and the resultant speech segment present in the signal can still be perceived. This indicates that the source and the vocal tract characteristics of speech remain intact. Any processing of the resultant signal can be looked at in terms of the source and vocal tract characteristics.

In this work, it is assumed that background music in speech is mostly additive and hence the vocal tract characteristics remain intact. But the overall signal characteristics of that speech with background music segment may be affected and cause their nature to be different compared to their nature in clean speech. The difference in nature is mostly due to the presence of music. By deriving features based on the vocal tract characteristics, some kind of discriminative information for the clean speech and speech with background music may be achieved. Even though the feature may capture the vocal tract characteristics of the speech and speech with background music segments, the presence of music in the latter may cause deviation in the value of this feature and this deviation is mostly due to the effect of the music in the signal. The deviation provides a sense of discrimination for the clean speech and speech with background music segments, thereby allowing for an effective classification between the two classes. The basic feature which represents the vocal tract characteristics is the mel frequency cepstral coefficients (MFCCs). MFCCs have been chosen as a base features for this task since they have been used in different types of speech related tasks and showed promising performances. The MFCCs are expected to deviate from the clean speech behavior when music is present in the background of speech which may provide discriminative information in terms of the vocal tract system. The classification system using MFCC features will be treated as a baseline system in this work.

The MFCCs represent the average spectral characteristics of a signal [90] since the energy in each band is summed during the MFCC computation. The average spectral characteristics may not be able to completely characterize the signal for the clean speech and speech with background music

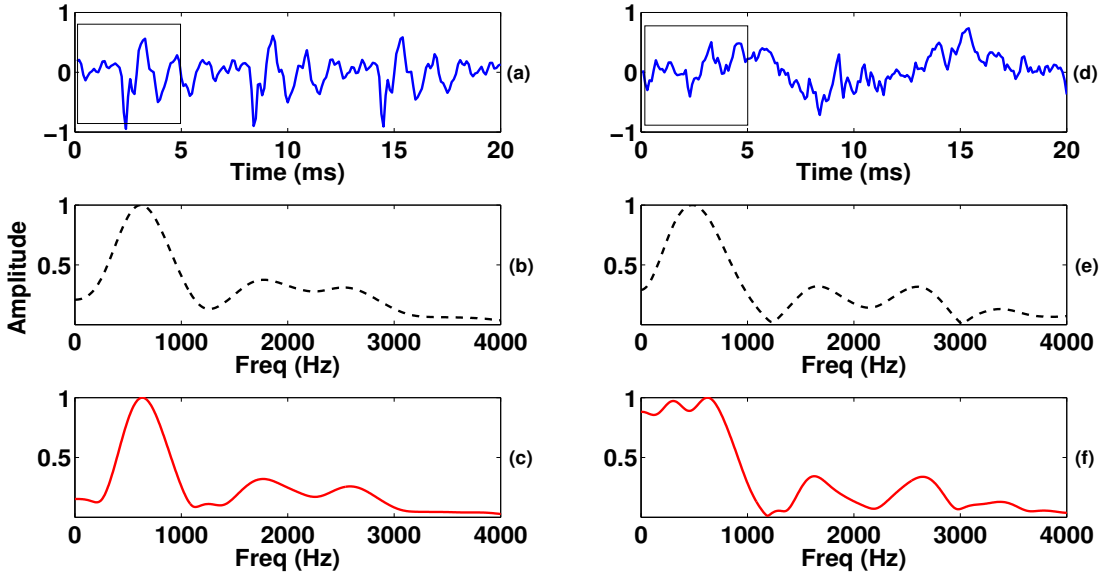


**Figure 4.2:** *Figure showing the difference between DFT and HNGD spectrum for a segment of 5ms (marked as black rectangle) of speech and speech with background music (a) Clean Speech (b) its DFT spectrum (dotted black) and (c) HNGD spectrum (continuous red) (d) Speech with background music (e) its DFT spectrum (dotted black) and (f) HNGD spectrum (continuous red)*

classification. Some fine fluctuations in the spectrum may be required to be captured for this task, since the presence of music for the speech with background music segments may be reflected in the fine spectral characteristics. To illustrate this point consider Figure 4.2. It can be seen in the figure that there is deviation of the overall signal characteristics, where it is observed that on adding music to the clean speech, the magnitude spectrum (marked as dotted black in Figure 4.2(b)& (e)) of a small segment (5 ms) of clean speech and speech with background music is different and this difference can be observed mostly in the fine spectral fluctuations. The specific changes are that some of the peaks get slightly shifted, the magnitude at 0 Hz is higher for the speech with background music and the introduction of additional peaks in the higher frequency regions for the speech with background music segment. Overall, it can be observed that there is a different nature for the peaks and valleys in clean speech and speech with background music segments. The difference of the fine spectral fluctuations in terms of the peaks and valleys indicates that there is discriminative information present between the two classes and if proper features are derived, a good level of classification may be achieved. The characteristics of the peaks and valleys can be captured by the spectral contrast which has been explored in [90]. The spectral contrast finds the difference between the spectral peaks and spectral

valleys in each sub-band. This feature has already been used in music classification task [90] and has been shown to perform well. This feature represents the relative spectral characteristics as opposed to MFCCs which represent the average spectral characteristics. The use of this feature may be complementary to the MFCC features for the classification task in this work.

The spectral contrast is computed on the DFT spectrum of a short (5 ms) segment of speech in order to capture the relative spectral differences of the clean speech and speech with background music observed in Figure 4.2. This DFT spectrum represents the spectral characteristics of the vocal tract system [91]. It was observed in [91], that the peaks due to the formants generally appear to be less prominent for the DFT spectrum since the spectral characteristics of the vocal tract system are averaged over the analysis window. In [91], Hilbert Envelope of the Numerator of Group Delay (HNGD) spectrum [91] was proposed which provides the instantaneous spectral characteristics of the vocal tract system and this method complements the existing DFT method in terms of the temporal resolution. It was chosen in [91] to take advantage of the additive and high resolution property of the group delay function [92–94]. The HNGD spectrum is based on zero time liftering which is analogous to the zero frequency filtering [19, 95]. This spectrum has been shown to give prominent peaks and valleys and computing the spectral contrast on the HNGD may provide better discrimination for the classification task than on the DFT. The features may need to be analyzed in terms of the frame level, utterance level, and the histogram level in order to properly understand their characteristics and to assess their contribution to the clean speech/speech with background music classification. The overall novelty can be described as follows:

- The spectral contrast has been shown to perform well for the music classification task. Its effectiveness for the classification between clean speech and speech with background music is explored
- In addition to computing the spectral contrast on DFT, the spectral contrast on HNGD is also computed and its ability to capture the relative spectral characteristics is investigated
- The combination of the average spectral characteristics represented by MFCCs and the relative spectral characteristics represented by the spectral contrast on DFT or HNGD, for the classification task are explored
- The analysis of the features at the frame level, utterance level, and histogram level is performed.

The features in this work hence consist of the spectral contrast of the HNGD spectrum which represents the relative spectral characteristics, the MFCCs which is considered as the baseline system and represents the average spectral characteristics and the spectral contrast of the DFT spectrum for comparison. The parameters for computing the spectral contrast which are the spectral peaks and spectral valleys are also considered as additional features since it was mentioned in [90] that these features, especially the spectral valleys give an additional spectral information. These features are evaluated individually to show their contribution to the task and are also combined by concatenating them to show their complementary nature with respect to each other. All the features (individual and combined) are passed through the classifiers consisting of the Support Vector Machines (SVM) and Gaussian Mixture Models (GMM).

The rest of the paper is organized as follows. The feature description and feature analysis are given in Section 4.2 and 4.3. Section 4.4 describes the classification framework. The results are given in Section 4.5. Finally, the conclusion is given in Section 4.6.

## 4.2 Spectral Contrast on DFT and HNGD spectrum representing Vocal Tract System Characteristics

This section provides the description of the various features used for the classification task in this work. The MFCC features which are considered as the baseline in this work are widely used in different speech and speaker application tasks. The description of their extraction can be found in a wide variety of literature related to speech related applications like speech recognition and speaker recognition. In this work, the 13-dimensional MFCCs are computed over a window size of 20 ms with a shift of 10 ms.

The spectral contrast feature will be described in detail here and note that this feature is computed on both the DFT and the HNGD spectrum. Hence the description of these two kinds of spectrums will be done first.

The DFT spectrum is defined as follows:

$$S[k,i] = \sum_{n=0}^{N-1} s[n+iR]w[n]e^{\frac{-j2\pi nk}{N}} \tag{4.1}$$

where $s[n]$ is the speech signal, $w[n]$ is a rectangular window, $R$ is the frame shift, and $N$ is the total number of points for computing the Fourier transform. The DFT spectrum is computed on a 5

ms window with every sample shift and the number of points for computing it is 2048.

The HNGD spectrum [91], on the other hand is obtained as follows:

(a) Consider $M$ samples of the differenced audio signal $s[n]$. That is $s[n]$ is defined for $n = 0, 1, ..., M-1$.

(b) Choose the DFT length $N >> M$ so that there is sufficient sampling in the frequency domain. The signal $s[n]$ is appended with the appropriate number of zeros to make its length equal to $N$.

(c) The windowed signal $x[n]$ is obtained as $x[n] = s[n]w_1[n]$, for $n = 0, 1, ..., N-1$, where

$$w_1[n] = \begin{cases} 0, & n = 0 \\ 1/(4sin^2(\pi n/(2N))), n = 1, 2, ..., N-1, \end{cases} \tag{4.2}$$

where $N$ is the window length.

Since the window function is highly decaying in nature, masking of the formant peaks may occur due to over-smoothing. This effect can be reduced by using the Fourier transform phase spectrum called the Group-Delay spectrum in place of the magnitude spectrum.

(d) The Numerator of Group Delay (NGD) function of $x[n]$ is then computed as,

$$g[k] = X_R[k]Y_R[k] + X_I[k]Y_I[k], k = 0, 1, ..., N-1 \tag{4.3}$$

where $X[k] = X_R[k] + jX_I[k]$ is the N-point DFT of the sequence $x[n]$, and $Y[k] = Y_R[k] + jY_I[k]$ is the N-point DFT of the sequence $y[n] = nx[n]$.

(e) The NGD function is double-differenced and sign reversed to obtain a function referred to as the DNGD function. This is done to enhance the spectral resolution and sharp peaks will be obtained at the formant locations.

(f) Finally, the Hilbert envelope (HE) [75] of the DNGD function is computed to obtain the HNGD spectrum which highlights the peaks obtained above.

The Hilbert envelope $a[n]$ of a sequence $e[n]$ is obtained as

$$a[n] = \sqrt{e^2[n] + e_h^2[n]} \tag{4.4}$$

where $e_h[n]$ is the Hilbert transform of $e[n]$,and is given by

$$e_h[n] = \text{IDFT}(E_h[\omega]) \tag{4.5}$$

where

$$E_h[\omega] = \begin{cases} -jE[\omega], 0 < \omega < \pi \\ jE[\omega], -\pi < \omega < 0 \end{cases} \tag{4.6}$$

and $E[\omega]$ is the DTFT of $e[n]$.

The HNGD spectrum is computed on a 5 ms window with every sample shift, and the number of points for computing the DFT is 2048 [91].

Finally, the spectral contrast feature [90] which estimates the strength of the spectral peaks, valleys and their differences in each sub-band is computed on the DFT and HNGD spectrum as follows.

Suppose the DFT or HNGD vector of the $k^{th}$ sub-band is $x_{k,1}, x_{k,2}, L, x_{k,N}$. Sorting in descending order we obtain $x'_{k,1}, x'_{k,2}, L, x'_{k,N}$ where $x'_{k,1} > x'_{k,2} > L > x'_{k,N}$.

The strength of the DFT or HNGD spectral peaks and valleys are estimated as:

$$Peak_k = log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\} \tag{4.7}$$

$$Valley_k = log\{\frac{1}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,N-i+1}\} \tag{4.8}$$

and their difference is:

$$SC_k = Peak_k - Valley_k \tag{4.9}$$

where $N$ is the total number of bins in the $k^{th}$ sub-band.

The value of $\alpha$ for the spectral peak and valley computation was set to 0.3. This value is set according to the best performance achieved in the experiments while varying $\alpha$ in the range of 0.1 to 0.5. The triangular overlapping filters are used, placed uniformly over the frequency band with a total number of six sub-bands. The spectral contrast over each sub-band is summed to obtain a single dimensional feature. The spectral peaks and valleys over the sub-bands are also summed. The summing process is done to reduce the dimension of the feature resulting in lesser computational requirements. This is also done so as to obtain a better representation of the feature since the feature can now be expressed as a single dimension and the behavior of this feature can be better displayed and analyzed. The behavior of the single dimensional feature, for example, the spectral contrast can be seen in Figure 4.3(b). A better analysis of the features will be provided in the next section. A

discussion on the effect of with and without summing the spectral contrast feature will be discussed in the results and discussion section.

## 4.3 Frame-wise, Utterance-wise and Histogram-wise Characterization of Vocal Tract System Features

The MFCCs represent the average spectral characteristics of the vocal tract system. The presence of music in the background will tend to change the average spectral characteristics and this difference can be captured by the MFCCs for performing the classification. However, some of the fine level changes in the spectrum may not be captured by the MFCCs since the spectral magnitudes in each band are averaged out. The MFCCs give an approximate spectral envelope characterizing the formants of the vocal tract system. The other frequency components which may also consist of components of the music may be averaged out while computing MFCCs thus giving a lower level of discrimination between the clean speech and speech with background music.

The spectral contrast represents the relative spectral characteristics and captures the fine level changes of the spectrum pertaining to the vocal tract system. In addition to capturing the formant peaks, the relation of the peaks to the valleys of the spectrum is captured thus, in essence, giving a better measure of the frequency components present in addition to speech, since the presence of the music in the background of speech may affect both the formant peaks and valleys of the spectrum compared to the formants and peaks when only speech is present.

The analysis of the features in this section will be described in three ways. The first is in terms of the frame wise characteristics of the DFT and HNGD spectrum and how it may affect the spectral contrast. The second is the utterance wise characteristics. The behavior of the spectral contrast on both the DFT and the HNGD spectrum will be studied over an utterance. The third analysis will consist of the behavior of the spectral contrast feature over a larger number of utterances. This study will be in terms of the histogram analysis, to explore the amount of separability of the features and their comparison.

#### 4.3.0.1 Frame-wise Characteristics

Figure 4.2 shows the frame wise behavior of the DFT and the HNGD spectrum for the clean speech and speech with background music. Note that the DFT and HNGD spectrum has been normalized in the range from 0 to 1 in the figure to visualize the difference between the two kinds of spectrums.

The speech with background music consists of the same clean speech segment but with the addition of music (guitar music taken from the GTZAN database which has been explored in [30,85,96]) at a signal to music ratio of 0dB. It is mentioned earlier that the addition of music to speech tends to change the overall signal characteristics of clean speech and it can be clearly seen in the figure (marked as dotted black for DFT) that the peaks and even valleys are affected due to the addition of music. This causes a deviation of the signal characteristics from their original clean speech behavior. This difference is captured for discriminating clean speech and speech with background music. It can also be observed in Figure 4.2(c)&(f) (marked as continuous red) that the signal characteristics obtained from HNGD spectrum are also affected by adding music which shows that computing the spectral contrast on this spectrum may also be feasible for discriminating clean speech and speech with background music.

An interesting observation can be seen by comparing the continuous red and dotted black plots for the HNGD and the DFT spectrum, respectively, in Figure 4.2. It is to be noted that the same audio segments are used for computing the DFT (dotted black) and the HNGD (continuous red) so as to easily compare their characteristics. The HNGD displays better spectral resolution compared to the DFT spectrum as observed. For the clean speech case (Figure 4.2(b) and (c)) the peaks around 700 Hz, 1700 Hz, and 2700 Hz are sharper for the HNGD (continuous red) compared to the DFT (dotted black) spectrum. This sharpness causes the spectral contrast which is the difference between the peaks and valleys, to have a higher value when computing on HNGD as compared to when computing on DFT. On the other hand, for the speech with background music regions (Figure 4.2(e) and (f)) it is seen that the peaks around the 1700 Hz, 2700 Hz, and 3400 Hz are only slightly sharper for the HNGD compared to the DFT. This may cause the spectral contrast to having almost the same values when computed on the HNGD and the DFT. In addition, the peaks around the 700 Hz are sharper for the HNGD compared to DFT, but for the region around (0-500 Hz), the amplitude of the HNGD has been drastically increased and a peak has been introduced around the 300 Hz region. The reason for the introduction of the peak around 300 Hz is because of the higher resolution of the numerator of group delay (NGD) function. This property of NGD causes the resonance features of the spectrum to be highlighted [91]. For the speech with background music segments, the music component, in particular, the instrumental ones may have a resonance like nature and this will cause the NGD function to highlight the resonance features of music as well. If we consider computing the spectral contrast (say for a band from 0 to 1000 Hz), its value will be lower for the HNGD compared to the DFT case.

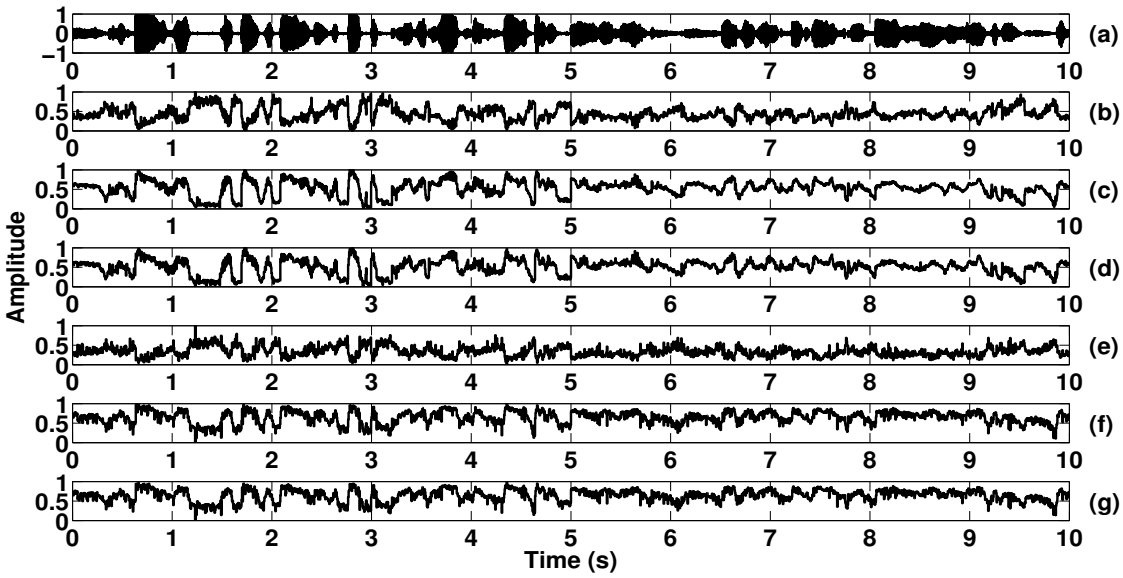


**Figure 4.3:** *Figure to demonstrate the behavior of the raw features for a single utterance for DFT and HNGD spectrum (a) Audio signal, where the first 5 s correspond to clean speech and the next 5 s correspond to the speech with background music taken from the Scheirer & Slaney database (b) Sum of Spectral Contrast (c) Sum of Spectral Peaks (d) Sum of the Spectral Valleys, of DFT (e) Sum of Spectral Contrast (f) Sum of Spectral Peaks (g) Sum of the Spectral Valleys, of HNGD*

The region around (0-500 Hz), in Figure 4.2(e) and (f), which was a sharp valley for the DFT has become flatter for the HNGD, thus in essence giving a smaller peak to valley difference for the HNGD particularly if it is computed in a band say around (0-1000 Hz). Overall, for the clean speech, it is expected that the spectral contrast value will be higher on the HNGD compared to the DFT, and for the speech with background music, its value may be equal or lower on the HNGD compared to the DFT. This implies that the spectral contrast computed on HNGD spectrum may provide a better discrimination for clean speech and speech with background music regions.

#### 4.3.0.2 Utterance-wise Characteristics

The frame wise characteristics showed that both the DFT and HNGD spectrums are capable of discriminating clean speech and speech with background music and it is expected that the spectral contrast on the HNGD will provide a better discrimination. In this section, an utterance is considered and this utterance consists of first 5 s of clean speech followed by next 5 s of the speech with background music taken from the Scheirer & Slaney database (the details of the database will be described in the results section). The spectral contrast sum is computed on this utterance for both the DFT and the

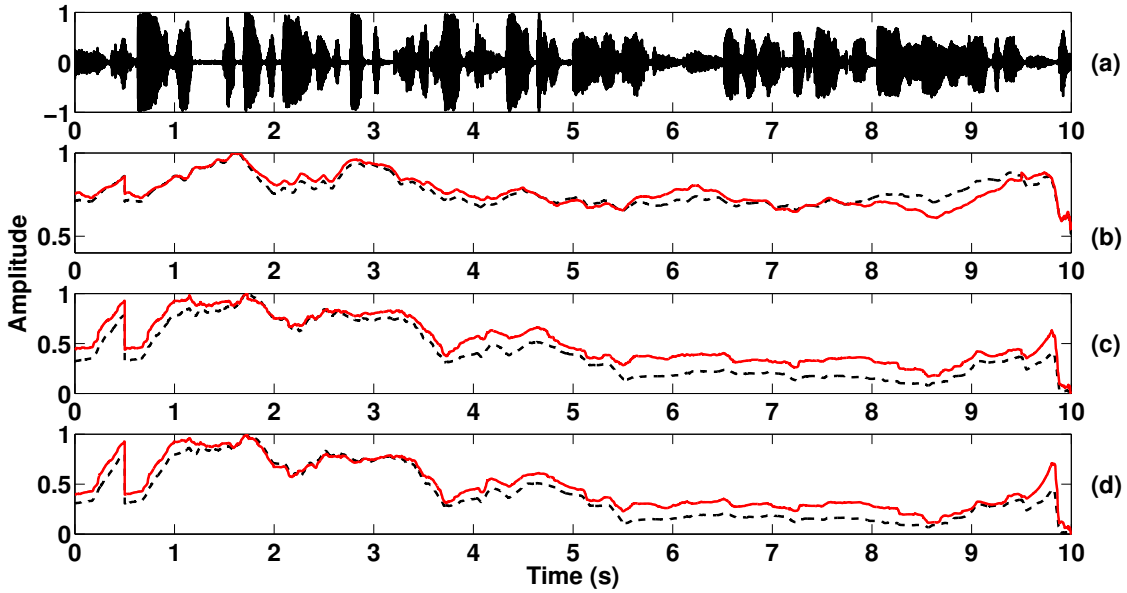


**Figure 4.4:** *Figure to demonstrate the behavior of the smoothed features for a single utterance for DFT and HNGD spectrum (a) Audio signal, where the first 5 s correspond to clean speech and the next 5 s correspond to the speech with background music taken from the Scheirer & Slaney database (b) Smoothed Sum of Spectral Contrast of DFT (dotted black) and HNGD (continuous red) (c) Smoothed Sum of Spectral Peaks of DFT (dotted black) and HNGD (continuous red) (d) Smoothed Sum of the Spectral Valleys of DFT (dotted black) and HNGD (continuous red)*

HNGD case and it is plotted in Figure 4.3. The spectral contrast sum on DFT is clearly seen to be higher for the clean speech segments and lower for the speech with background music segments (Figure 4.3 (b). The sum of the spectral peaks and valleys on DFT are also plotted and shown in (c) and (d) of Figure 4.3. The sum of the spectral peaks and valleys also show discrimination between the two classes. The sum of the spectral contrast on DFT in Figure 4.3(b) is smoothed by computing the mean over a 1 s frame with a shift of 1 s and plotted in Figure 4.4 (b) marked as dotted black. The smoothed version is computed to clearly show the discrimination between the two classes. The sum of the spectral peaks and valleys on DFT are also smoothed by computing the variance over a 1 s frame with a shift of 1 s and shown in Figure 4.4 (c) and (d) marked as dotted black. The variation of the sum of the spectral peaks and valleys are higher for the clean speech segments compared to the speech with background music segments which is why the smoothing is performed by computing the variance. Similarly the sum of the spectral contrast, peaks and valleys are computed for the HNGD case and plotted in Figure 4.3 (e),(f) and (g). The smoothed versions for the HNGD are plotted as continuous red lines in Figure 4.4 (b),(c) and (d).

On comparing the plot of Figure 4.4(b) in which the dotted black line shows the sum of the spectral contrast computed on the DFT and the continuous red line shows the sum of the spectral contrast computed on the HNGD, it can be seen that the smoothed sum of the spectral contrast is higher for the HNGD case compared to the DFT spectrum for the clean speech and this can be observed clearly around the (2-4.5) s segments of the clean speech. The value of the smoothed sum of the spectral contrast for the speech with background music segment is almost the same for both the HNGD and the DFT spectrum (6.5-8)s. For the region from (8-9.5)s, the value is lower for the HNGD compared to the DFT. The reason for this is attributed to the occurrence of amplitude increase around the (0-500 Hz), which has been observed earlier for the HNGD on speech with background music segments while performing the frame-wise characteristics (Figure 4.2(d)). The increase in amplitude may have caused the spectrum to be flatter which gives a lower peak to valley difference for the HNGD compared to the DFT. There are also cases where the value for DFT is lower than for the HNGD (around 6 s). For the case around 6 s of Figure 4.4(b), the HNGD for the particular frame around that region may have caused the amplitude to increase. However, this increase may have caused the peak to be sharper in some other frequency band different from the (0-500 Hz) band observed earlier, considering the non-stationarity of music. It may also happen that the resonant frequency of a certain kind of music may be same as that of the resonant frequency of the vocal tract system of speech, thus it may further sharpen the existing peaks corresponding to speech. This sharpness of the peak may cause the peak to valley difference to be higher for the HNGD compared to the DFT.

Overall, it is expected that there should be a larger discrimination between the feature values of the two classes for the HNGD compared to the DFT. This shows that the evidence observed for the frame-wise characteristics is reflected for a single utterance as well which means that computing the spectral contrast on the HNGD may give better discrimination than the DFT spectrum for the two classes. The frame-wise and utterance-wise characteristics display some evidence for the discrimination of clean speech and speech with background music segments. However, some statistical analysis is required to obtain the actual feature distribution and to gain more insight of the overlap between the features. For this purpose, the histogram of the features is computed and is discussed in the following section.

The sum of the spectral peaks in Figure 4.3(c) of the HNGD (continuous red) are higher than DFT (dotted black) for both the clean speech and speech with background music regions. The reason is due

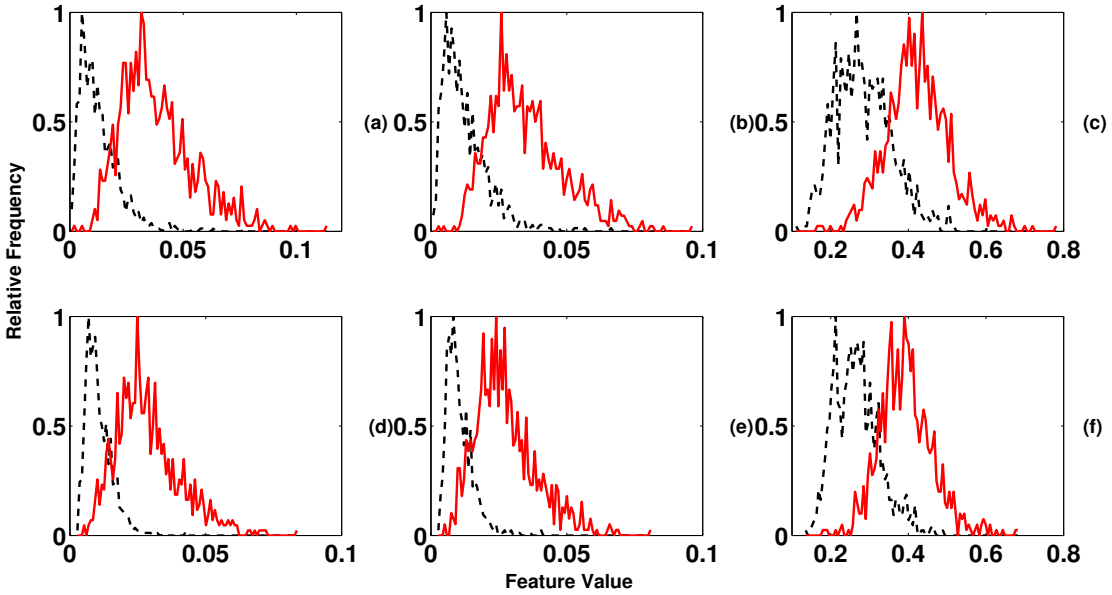


**Figure 4.5:** *Histogram plot for S&S database to display the degree of overlap for computing the features on DFT and HNGD. The distribution for clean speech is represented by the continuous red line while for the speech with background music it is represented by the dotted black line. (a) Sum of Spectral Peaks (b) Sum of Spectral Valleys (c) Sum of Spectral Contrast, of DFT (d) Sum of Spectral Peaks (e) Sum of Spectral Valleys (f) Sum of Spectral Contrast, of HNGD*

to the sharpness of the peaks which is caused by the HNGD, due to the high resolution properties of the NGD function used for computing HNGD. Similarly, the sum of the spectral valleys in Figure 4.3(d) is also higher for the HNGD than the DFT for both the clean speech and speech with background music regions due to the sharpness of the valleys caused by the additive and high resolution properties of the NGD function. Although the values are higher for the sum of spectral peaks and sum of the spectral valleys for the HNGD in the clean speech, their values in the speech with background music are also high which shows that there may not be advantages of using HNGD for the spectral peaks and valleys. In contrast, for the spectral contrast, the values in the clean speech for the HNGD are higher than the DFT, while the values are lower for the HNGD compared to the DFT for the speech with background music regions which may give more discrimination for reasons mentioned earlier. However, the sum of the spectral peaks and valleys will be included since they may give additional spectral information.

#### 4.3.0.3 Histogram-wise Characteristics

The analysis over a larger number of utterances is performed by computing the histogram for the sum of the spectral contrast on both the HNGD and the DFT spectrum. A total of 60 utterances

which consists of 5 s of clean speech followed by 5 s of speech with background music are taken from the Scheirer and Slaney database to compute the histogram and shown in Figure 4.5. The distribution for the clean speech and speech with background music is displayed as continuous red and dotted black, respectively. The histogram of the sum of the spectral peaks and valleys are also shown in the figure. It can be observed that all the features show a good discrimination between the two classes with a small degree of overlap between the classes. It can be observed in Figure 4.5 (c) and (f), that there is a higher degree of overlap for the sum of the spectral contrast computed on the DFT spectrum compared to the HNGD spectrum which supports the observations of the frame wise and utterance wise behaviors of the feature. This shows that the sum of the spectral contrast computed on the HNGD spectrum has a higher level of discrimination between the clean speech and speech with background music regions.

**Table 4.1:** *Bhattacharyya coefficients computed on the histograms given in Figure 4.5*

| Feature → | Sum of Spectral Peaks | Sum of Spectral Valleys | Sum of Spectral Contrast |
|---|---|---|---|
| DFT | **0.4841** | **0.4985** | **0.6142** |
| HNGD | **0.5031** | **0.4877** | **0.5528** |

In order to quantify the separability between the two distributions, Bhattacharyya coefficient was computed on the histograms and displayed in Table4.1. The coefficient values display a higher separability for the HNGD case compared to the DFT case for the sum of the spectral contrast which validates the observations earlier.

## 4.4 Description of Feature Extraction and Classification of Clean Speech vs Speech with Background Music

The overall clean speech versus speech with background music classification system can be seen in Figure 4.6. Given an audio signal, the different features as shown in the figure are extracted. First, the MFCCs are extracted for a window size of 20 ms with a shift of 10 ms. The spectral contrast, spectral peak, and spectral valley are computed for a window size of 5 ms with every sample shift. Next, the spectral contrast, spectral peak, and the spectral valley are each summed over all their respective bands. The next step involves taking a longer and fixed window size to compute the statistics of the features so that they may be combined uniformly. The variance of MFCCs over 1 s is taken.

Similarly, the sum of the spectral contrast shows good discrimination after mean smoothing which

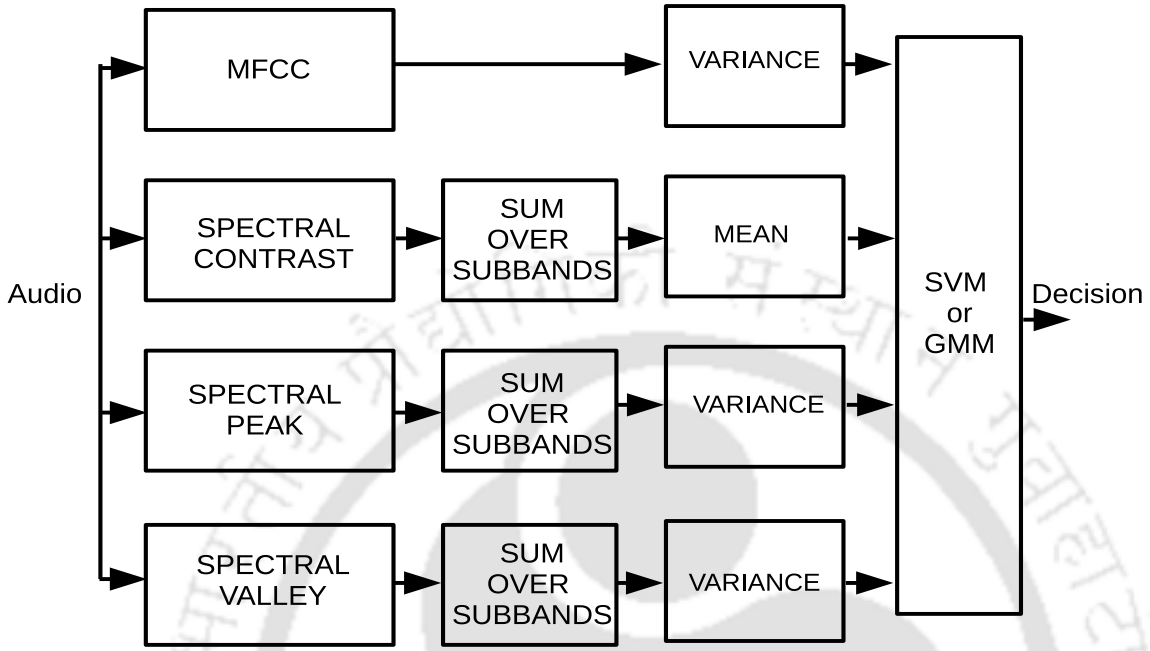


**Figure 4.6:** *Representation of how features are obtained from an Audio input and fed to the classifiers for training and testing. The MFCC features are computed for a frame size of 20 ms with a shift of 10 ms while the spectral based features are computed for a frame size of 5 ms with a shift of every sample. 6 sub-bands are considered so the summing is done over these 6 sub-bands. The statistics, mean and variance are computed over a window of 1 s with a shift of 1 s.*

tells that taking the mean of the spectral contrast sum is a good option to obtain the feature. The same 1 s window size is taken for computing the mean. The sum of the spectral peak and valley shows good discrimination after variance smoothing and hence the variance of the spectral peak and valley over 1 s is computed to obtain the statistic of this feature. The features obtained are hence either passed individually or combined by concatenation through the classifiers for training and testing. The classifiers used in this work consists of the GMM and the SVM. The parameters of the classifiers are decided by training with different parameter values varied by a grid search. The parameters which perform the best on the testing set are selected. All the experiments using SVM were carried out using the libSVM [84] with a radial-basis function (RBF) kernel.

## 4.5 Results and Discussion

### 4.5.1 Database

The clean speech versus speech with background music classification task is tested on data which consists of the Scheirer & Slaney (S&S) database [12] (having 60 files of speech and 60 files of speech

**Table 4.2:** *Classification accuracy (%) using different features on S&S database*

| Spectrum → | DFT | | | HNGD | | |
|---|---|---|---|---|---|---|
| Feat (classification) ↓ | Speech | Music | Overall | Speech | Music | Overall |
| MFCC (GMM) | 84.77 | 84.33 | 84.55 | - | - | - |
| SP (GMM) | 79.44 | 91.00 | 85.22 | 80.33 | 93.22 | 86.77 |
| SV (GMM) | 79.66 | 90.00 | 84.83 | 80.77 | 93.33 | 87.05 |
| SC (GMM) | 81.66 | 83.44 | 82.55 | 84.22 | 86.66 | 85.44 |
| SP+MFCC(GMM) | 87.22 | 86.33 | 86.77 | 89.22 | 87.22 | 88.22 |
| SV+MFCC(GMM) | 86.88 | 86.22 | 86.55 | 88.44 | 87.33 | 87.88 |
| SC+MFCC(GMM) | 88.55 | 85.44 | 87.00 | 90.33 | 85.88 | 88.11 |
| MFCC+SP+SV+SC(GMM) | **89.77** | **87.88** | **88.83** | **91.55** | **89.55** | **90.55** |
| MFCC (SVM) | 85.33 | 86.33 | 85.83 | - | - | - |
| SP (SVM) | 87.22 | 87.22 | 87.22 | 85.22 | 89.11 | 87.16 |
| SV (SVM) | 87.55 | 85.11 | 86.33 | 85.22 | 89.00 | 87.11 |
| SC (SVM) | 84.44 | 80.22 | 82.33 | 87.44 | 83.55 | 85.50 |
| SP+MFCC(SVM) | 89.11 | 87.33 | 88.22 | 90.44 | 88.44 | 89.44 |
| SV+MFCC(SVM) | 88.22 | 86.66 | 87.44 | 90.33 | 88.00 | 89.16 |
| SC+MFCC(SVM) | 89.44 | 87.11 | 88.27 | 90.33 | 87.33 | 88.83 |
| MFCC+SP+SV+SC(SVM) | **91.33** | **87.77** | **89.55** | **91.66** | **88.77** | **90.22** |

with background music, each of length 15 s). The audio data in this database has a sampling frequency of 22050 Hz and has been down-sampled to 8000 Hz. The broadcast news (BN) database recorded from Indian broadcast news channels (having 500 files of speech and 500 files of speech with background music, each of length 5 s) is also considered for the task. The sampling frequency of this data is 8000 Hz. The speech with background music segments in the broadcast news case are the ones taken from voice-over speech and news headlines.

### 4.5.2 Results using GMM and SVM

The evaluated results for the S&S database are shown in Table 4.2. The results using MFCCs are quite promising considering the complexity of the two classes. First, the results using the GMM

**Table 4.3:** *Classification accuracy (%) using different features on BN database*

| Spectrum → | DFT | | | HNGD | | |
|---|---|---|---|---|---|---|
| Feat (classification) ↓ | Speech | Music | Overall | Speech | Music | Overall |
| MFCC (GMM) | 78.28 | 86.52 | 82.40 | - | - | - |
| SP (GMM) | 83.36 | 91.20 | 87.28 | 84.60 | 92.76 | 88.68 |
| SV (GMM) | 83.24 | 90.88 | 87.06 | 84.44 | 92.84 | 88.64 |
| SC (GMM) | 76.96 | 80.32 | 78.64 | 83.28 | 81.48 | 82.38 |
| SP+MFCC(GMM) | 87.56 | 88.00 | 87.78 | 87.00 | 88.48 | 87.74 |
| SV+MFCC(GMM) | 87.08 | 87.88 | 87.48 | 86.56 | 88.40 | 87.48 |
| SC+MFCC(GMM) | 84.60 | 87.60 | 86.10 | 84.72 | 88.08 | 86.40 |
| MFCC+SP+SV+SC(GMM) | **89.80** | **88.52** | **89.16** | **89.28** | **89.96** | **89.62** |
| MFCC (SVM) | 91.40 | 89.00 | 90.20 | - | - | - |
| SP (SVM) | 87.56 | 88.16 | 87.86 | 89.20 | 89.96 | 89.58 |
| SV (SVM) | 88.04 | 87.80 | 87.92 | 89.20 | 89.44 | 89.32 |
| SC (SVM) | 79.40 | 77.48 | 78.44 | 84.80 | 79.72 | 82.26 |
| SP+MFCC(SVM) | 91.80 | 91.24 | 91.52 | 92.08 | 92.00 | 92.04 |
| SV+MFCC(SVM) | 91.68 | 90.76 | 91.22 | 91.88 | 91.76 | 91.82 |
| SC+MFCC(SVM) | 92.92 | 90.56 | 91.74 | 93.60 | 90.84 | 92.22 |
| MFCC+SP+SV+SC(SVM) | **93.32** | **91.32** | **92.32** | **93.88** | **92.36** | **93.12** |

classifier will be discussed. The results on the individual features are shown initially where it is observed that the sum of the spectral peaks gives the best performance followed by the sum of the spectral valley and then the sum of the spectral contrast on the DFT spectrum. However on combining each of these features with MFCCs, the sum of the spectral contrast feature performs better which shows that this feature even though it may not perform well individually has a better complementary nature with MFCCs thereby improving the performance compared to the sum of the spectral peaks and the sum of the spectral valleys when using GMM. The results on the HNGD spectrum are better than the case of DFT as expected due to the additive and high resolution property of the HNGD spectrum which gives a better discrimination between the two classes. On combining the features in the HNGD case, the sum of the spectral contrast with MFCCs gives almost a comparable performance

with the case when the sum of the spectral peaks are combined with MFCCs. The best performances are achieved when all the features are combined together which are the results shown in bold for the GMM case. A similar trend is observed when using SVM as a classifier.

Next, the evaluation is done on the BN database and shown in Table 4.3. The similar trends are observed for this database also. The only difference being that the sum of spectral contrast combination with MFCCs on both DFT and HNGD, performs lesser compared to the combination of the sum of spectral peaks and valleys when using GMM as a classifier. However, the difference is minimum. When using SVM the sum of the spectral contrast combination with MFCCs on both DFT and HNGD has almost the same performance compared to the other two combinations. The combination of all the features in this case also gave the best results and can be seen in bold both using GMM and SVM in the table.

An interesting trend which has been mentioned earlier is observed in the results of Table 4.2 and 4.3. The sum of the spectral contrast feature performs lesser compared to the sum of the spectral peaks and valleys. However, by combining each of these features, their performances are almost comparable and in some cases, the sum of the spectral contrast combination is better than the sum of the spectral peaks or valleys combination with MFCCs. The reason for this may be because the spectral peaks which are captured for computing the sum of the spectral peak feature may have already been captured by the MFCCs since the MFCCs represent approximately the spectral envelope which includes peaks of the spectrum as well. As a result, there is common information captured by both of them hence resulting in the combination to not be complementary enough. The spectral valley information may also have been captured by the MFCCs since the spectral envelope represented by the MFCCs may have some information of the valleys of the spectrum. A certain band for computing MFCCs may also cover a valley of the spectrum. On the other hand, the MFCCs do not capture the difference between the peaks and valleys in each sub-band. In the sense, the MFCCs sum up the values in each band thereby averaging out the spectral characteristics of that band. The spectral contrast gives a relative difference of the peaks and valleys in each band. These two features hence complement each other in a better way and it is reflected in the results of the Table 4.2 and 4.3.

### 4.5.3 Mismatched Training and Testing Data

An experiment was also conducted to see how training on one database and testing with the other affects the results. The training was done on the S & S database and tested on the Broadcast News

**Table 4.4:** *Classification accuracy (%) on the BN database using the models trained on the S & S database.*

| Spectrum → | DFT | | | HNGD | | |
|---|---|---|---|---|---|---|
| Feat (classification) ↓ | Speech | Music | Overall | Speech | Music | Overall |
| MFCC (GMM) | 69.64 | 93.20 | 81.42 | - | - | - |
| SP (GMM) | 91.00 | 85.48 | 88.24 | 94.28 | 82.88 | 88.58 |
| SV (GMM) | 90.20 | 86.24 | 88.22 | 94.24 | 82.44 | 88.34 |
| SC (GMM) | 82.24 | 73.60 | 77.92 | 91.24 | 71.20 | 81.22 |
| SP+MFCC(GMM) | 88.24 | 91.44 | 89.84 | 91.04 | 90.96 | 91.00 |
| SV+MFCC(GMM) | 87.12 | 91.64 | 89.38 | 90.96 | 91.00 | 90.98 |
| SC+MFCC(GMM) | 84.20 | 92.60 | 88.40 | 86.08 | 91.68 | 88.88 |
| MFCC+SP+SV+SC(GMM) | **91.84** | **89.96** | **90.90** | **94.52** | **88.52** | **91.52** |
| MFCC (SVM) | 72.60 | 93.24 | 82.92 | - | - | - |
| SP (SVM) | 94.56 | 80.64 | 87.60 | 95.88 | 76.92 | 86.40 |
| SV (SVM) | 93.84 | 80.80 | 87.32 | 95.64 | 76.76 | 86.20 |
| SC (SVM) | 84.60 | 69.48 | 77.04 | 92.72 | 67.76 | 80.24 |
| SP+MFCC(SVM) | 91.08 | 89.48 | 90.28 | 93.60 | 88.40 | 91.00 |
| SV+MFCC(SVM) | 89.68 | 90.44 | 90.06 | 93.24 | 88.24 | 90.74 |
| SC+MFCC(SVM) | 88.12 | 92.56 | 90.34 | 89.52 | 90.80 | 90.16 |
| MFCC+SP+SV+SC(SVM) | **92.08** | **90.12** | **91.10** | **95.04** | **88.08** | **91.56** |

database. The results of this are shown in Table 4.4.The similar trends as Table 4.2 and 4.3 is observed where the sum of the spectral contrast feature performs poorly compared to the sum of the spectral peaks and the spectral valleys. The combination of the sum of spectral contrast with MFCCs is almost comparable to the case of combining the sum of the spectral peaks or valleys with MFCCs. The HNGD case performs better compared to the DFT case. The combination of all the features gave the best performance using both GMM and SVM.

### 4.5.4 Results without Summing the Features

It was mentioned earlier that the spectral contrast was summed to reduce the dimensional of the feature hence requiring lesser computations. An experiment was conducted to see the effect of

**Table 4.5:** *Performance in terms of classification accuracy (%) using the different individual features on the S&S database and the BN database. In the table, the abbreviations, MFCC indicates the mel frequency cepstral coefficients, SC indicates spectral contrast, SV indicates spectral valley and SP indicates spectral peak. The classifier used is GMM.*

| Database → | S&S database | | | | | | BN database | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Spectrum → | DFT | | | HNGD | | | DFT | | | HNGD | | |
| Features ↓ | Speech | SBM | Overall | Speech | SBM | Overall | Speech | SBM | Overall | Speech | SBM | Overall |
| MFCC | 84.77 | 84.33 | 84.55 | - | - | - | 78.28 | 86.52 | 82.40 | - | - | - |
| SP | 84.33 | 91.88 | 88.11 | 78.33 | 94.55 | 86.44 | 79.37 | 89.28 | 84.32 | 77.05 | 89.88 | 83.47 |
| SV | 79.66 | 90.66 | 85.16 | 78.44 | 94.55 | 86.50 | 77.94 | 89.34 | 83.64 | 75.97 | 89.65 | 82.81 |
| SC | 65.77 | 78.44 | 72.11 | 69.00 | 79.33 | 74.16 | 46.85 | 81.08 | 63.97 | 59.05 | 81.82 | 70.44 |
| MFCC+SC+SV+SP | **85.11** | **90.11** | **87.61** | **83.77** | **92.88** | **88.33** | **81.62** | **87.05** | **84.34** | **80.97** | **87.65** | **84.39** |

not summing the spectral contrast on both the DFT and HNGD spectrum. The spectral peaks and valleys are also evaluated. The results of using the spectral contrast as a 6-dimensional feature (from the 6 sub-bands) are given in the second last row of Table 4.5. The experiments for this case have been conducted using GMM on both the DFT and HNGD. By comparing the spectral contrast result in Table 4.5 with the sum of the spectral contrast result in the sixth row of Table 4.2 for the S&S database, it is observed that the sum of the spectral contrast performs better on both the DFT and HNGD cases. For the BN database also the sum of the spectral contrast performs better than the 6-dimensional spectral contrast. The result of the sum of the spectral contrast on the BN database can be found in the sixth row of Table 4.3. The spectral peaks and valleys have almost comparable performances for the sum of the spectral contrast and the 6-dimensional spectral contrast. The results when combining all the features is also better on both the DFT and the HNGD, after summing the features compared to the case when directly the large dimensional features are used. These summed results are shown bold in Table 4.5 as well as in Table 4.2 and 4.3. This shows that in addition to the lower computational cost for the added features, the results are also better when using the sum of the spectral contrast as a single dimensional feature.

### 4.5.5 Results on Speech with Background Noise of BN database

The features have shown to perform well for the classification of clean speech and speech with background music segments. In the broadcast news, there may also be a presence of speech with

**Table 4.6:** *Classification accuracy (%) using different features on BN database for the clean speech vs speech with background noise classification*

| Spectrum → | DFT | | | HNGD | | |
|---|---|---|---|---|---|---|
| Feat (classification) ↓ | Speech | Music | Overall | Speech | Music | Overall |
| MFCC (GMM) | 78.45 | 87.48 | 82.97 | - | - | - |
| SP (GMM) | 82.00 | 88.57 | 85.28 | 81.60 | 89.02 | 85.31 |
| SV (GMM) | 80.68 | 87.54 | 84.11 | 80.40 | 88.97 | 84.68 |
| SC (GMM) | 73.82 | 74.22 | 74.02 | 78.40 | 76.80 | 77.60 |
| SP+MFCC(GMM) | 84.80 | 88.57 | 86.68 | 84.74 | 88.68 | 86.71 |
| SV+MFCC(GMM) | 84.45 | 88.57 | 86.51 | 84.22 | 88.34 | 86.28 |
| SC+MFCC(GMM) | 81.71 | 89.25 | 85.48 | 82.62 | 89.48 | 86.05 |
| MFCC+SP+SV+SC(GMM) | **88.22** | **89.25** | **88.74** | **87.60** | **89.77** | **88.68** |
| MFCC (SVM) | 86.05 | 86.80 | 86.42 | - | - | - |
| SP (SVM) | 78.91 | 89.20 | 84.05 | 86.28 | 85.82 | 86.05 |
| SV (SVM) | 81.94 | 86.28 | 84.11 | 85.88 | 85.02 | 85.45 |
| SC (SVM) | 79.20 | 69.94 | 74.57 | 75.71 | 78.11 | 76.91 |
| SP+MFCC(SVM) | 89.82 | 88.97 | 89.40 | 89.71 | 89.48 | 89.60 |
| SV+MFCC(SVM) | 89.14 | 88.74 | 88.94 | 89.31 | 89.20 | 89.25 |
| SC+MFCC(SVM) | 91.25 | 89.82 | 90.54 | 91.37 | 89.88 | 90.62 |
| MFCC+SP+SV+SC(SVM) | **91.48** | **90.22** | **90.85** | **91.25** | **90.00** | **90.62** |

background noise in cases where the coverage is in outdoors. An example can be for the case of a reporter's speech. Such outdoor scenarios may be under the influence of different types of noise. It will be interesting to see if the features work for the classification of clean speech and speech with background noise cases. 500 files of clean speech and 500 files of speech with background noise have been taken and the algorithm using the defined features is evaluated for the classification of clean speech versus speech with background noise segments. The results are evaluated and shown in

Table 4.6. It can be observed from the table that good performances are achieved for this case also and the trend of the results are similar to the classification task of clean speech versus speech with background music. This shows that the features are robust against any kind of information present in the background, be it music or noise. The combination of all the features gave the best performance and are shown as bold fonts in Table 4.6. The HNGD and the DFT seem to be giving comparable performances for this type of classification.

## 4.6 Summary

This work explores the vocal tract features of speech in the form of the sum of the spectral contrast on DFT and HNGD spectrum for clean speech versus speech with background music segments. These features have been compared with MFCCs and have found to be better. The MFCCs represent the average characteristics while the sum of the spectral contrast represents the relative characteristics of the spectrum related to the vocal tract system of speech. The relative and average characteristics have also been combined by concatenating the MFCCs with the spectral contrast. The combination case gave the best performances across the two databases (S&S and BN database). Among the DFT and HNGD spectrum, the HNGD performs better due to its additive and high resolution property.

# 5

# Speech Enhancement and Phone Recognition of Speech with Background Music

Contents

### Objective

*This work explores the significance of source information for speech enhancement resulting in better phone recognition of speech with background music segments. Standard procedure for speech enhancement in noisy conditions involves sequential processing in terms of the temporal, spectral and perceptual methods. This work follows the same sequential processing but with the additional modification of studying the effect of the source, particularly in the temporal and perceptual based enhancement techniques for enhancing speech with background music segments. The source information is studied in terms of the epoch locations and epoch strength obtained after passing the sum of the mean and standard deviation of the component envelopes computed across frequencies obtained using the single frequency filter (SFF) through a zero frequency filter (ZFF). This method of obtaining epoch locations and epoch strength will be termed as SFF-ZFF in this work. The enhanced segments are passed through a phone recognizer built using Gaussian Mixture Model-Hidden Markov Model (GMM-HMM), Subspace Gaussian Mixture Model-Hidden Markov Model (SGMM-HMM) and Deep Neural Network-Hidden Markov Model (DNN-HMM) system, where the models are trained on clean speech. The enhanced audio files show a better phone error rate (PER) than the speech with background music audio files, which means that performing enhancement in terms of the source information is significant for the speech with background music regions.*

## 5.1 Introduction

The problem of automatic transcription of broadcast audio when there is a mismatch between the training and testing samples is a challenging one. Assuming that there are models trained on clean speech when tested with the speech which has been added with background music (for example, the news headlines and voice over speech in Indian broadcast news), the accuracy drops drastically due to the acoustic mismatch between the training and testing data. The separate models for the speech with background music can be trained, however, there may be insufficient data for training since the speech with background music segments in broadcast audio occur less frequently when compared to the clean speech case. Even though their occurrence may be lesser than the clean case, their transcription may be necessary for multimedia related tasks since the news headlines and voice over speech contain sufficient information in terms of the transcripts about certain events.

In this work, an alternate approach is followed wherein the speech with background music is enhanced prior to testing on the models which have been trained on the clean speech. There have been a lot of attempts for speech enhancement of speech degraded by noise. There are spectral based approaches which rely on the fact that the perception of human speech is not sensitive to short-time phase [34, 35] and some of the spectral based methods have been explored in [15, 35–40]. There are also subspace based approaches proposed in [57–59] where the components of speech and noise are decomposed into mutually orthogonal subspaces and it is assumed that the speech and noise can be represented in terms of a low-rank linear model. Methods which incorporate the use of the probability distributions have been explored in [44, 45]. All these approaches are based on either modeling the background noise present or are dependent on the nature of the noise. These approaches seem to be simple and effective, however, there are certain drawbacks. For example, if the noise is from unknown sources, then modeling it will be difficult. There are also non-stationary noises and modeling such noises becomes cumbersome.

There have been approaches which are based on directly exploiting the speech signal characteristics particularly the high signal to noise ratio (SNR) regions without taking into account the background noise present. These approaches have also been explored in [16–18, 71] which exploits the high SNR regions to enhance the speech signal. Since these methods do not depend on the degradation present, they will be explored for the case when music is present as a degradation. The enhancement strategy incorporated in this work involves the sequential method of processing, consisting of the temporal [17,71], spectral [17,71] and perceptual [18] enhancement on the noisy speech [17], reverberant speech [71] and the foreground speech [18]. These sequential steps of processing have been followed in this work for the enhancement of speech with background music segments, where each of the steps (mostly temporal and perceptual enhancement) have been modified based on the presence of a different kind of degradation, which is music in the background.

The use of speech-specific features for speech/music classification has been explored in [96]. The features in that work were developed based on the fact that speech is modeled in terms of the source, vocal tract system, and suprasegmental information. The music (instrumental) modeling is different from that of speech and based on the kind of music, the model may be different. Hence, defining features in terms of music may be difficult. However, defining features in terms of speech which is characterized in terms of source, vocal tract system, and suprasegmental information, may cause these

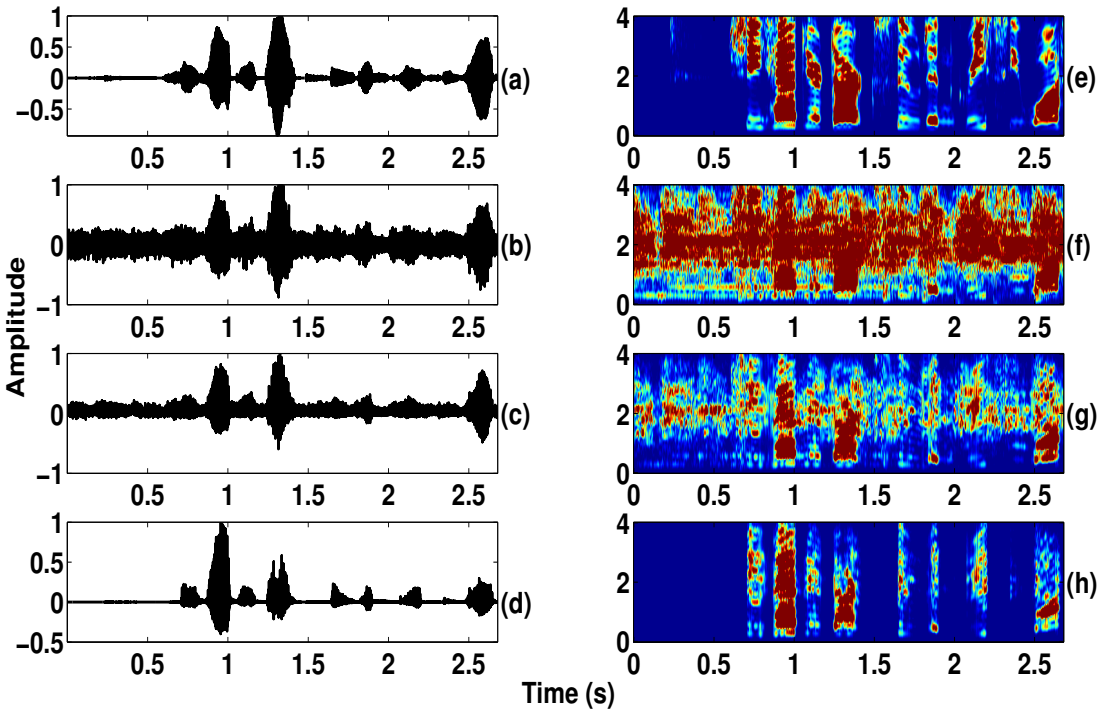


**Figure 5.1:** *Illustration of the significance of source information (a) Clean Speech (b) Speech added with rock music (SMR=0dB) (c) Enhanced Speech with source from clean speech and vocal tract system from speech with background music (not considering strength of excitation (SoE)) (d) Enhanced Speech with source from clean speech and vocal tract system from speech with background music (considering strength of excitation (SoE)) (e)-(h) Spectrogram of (a)-(d), respectively.*

features to deviate significantly in music thus achieving some sort of discrimination between the two segments. These features have been exploited for obtaining the gross weight function for temporal enhancement of speech with background music. The gross weight function plays the same role as the speech/non-speech detection in temporal enhancement [17].

The main contribution of the work is using the epoch locations obtained from single frequency filter-zero frequency filter (SFF-ZFF) combination, for generating the fine weight function which is part of temporal enhancement [17]. These epoch locations will also be used for generating the impulse train in the perceptual based enhancement [18]. The SFF-ZFF combination is proposed in this work to locate the instants of significant excitation (epoch locations) in speech with background music. The epoch locations are obtained by passing the sum of the mean and standard deviation of the component envelopes computed across frequencies obtained using the single frequency filter (SFF) [97,98], through a zero frequency filter (ZFF) [19]. The reason for using this method is that if the speech with

background music segment was passed directly through the ZFF (which is similar to the zero band filtering (ZBF) used for temporal enhancement in [18]), spurious epochs will be introduced. On the other hand, the music components get de-weighted after summing the values of the mean and standard deviation of the component envelopes computed across frequencies obtained using the single frequency filter (SFF). The summed output, when passed through the ZFF, gives epoch locations which are better compared to using ZFF alone, as will be illustrated in the later sections. This method of obtaining the epoch locations will be termed as SFF-ZFF in this work. As mentioned, the epoch locations derived from this method is used for generating the fine weight function. The gross weight function and the fine weight function are multiplied to get the total weight function. The total weight function is then used to weight the LP residual. The modifications made over the earlier temporal enhancement [17, 71], is basically in terms of the source information, where the fine weight function is derived based on the epoch locations obtained from SFF-ZFF. The perceptual based enhancement [18] has also been modified based on the source wherein the excitation source input has been replaced with the impulse train derived from the SFF-ZFF. The approximate vocal tract response is estimated from the Mel Cepstral Coefficients (MCCs) by using the Mel Log Spectral Approximation (MLSA) filter as in [18]. The reason why the excitation is replaced with the SFF-ZFF is due to the fact that the excitation source plays a major role in enhancement.

The significance of the source information can be understood by considering Figure 5.1. Figure 5.1 (a) shows a clean speech signal taken from the TIMIT database and Figure 5.1 (b) shows the same TIMIT utterance and with the addition of rock music at a signal to music ratio (SMR) of 0 dB. The source information from the clean speech is extracted in terms of the epoch locations from zero frequency filtered signal (ZFFS) and the vocal tract system information is extracted from the speech with background music, in terms of the MCCs obtained using MLSA filter. The speech signal is synthesized by using the excitation source signal in the form of an impulse train (consisting of impulses at obtained epoch locations from ZFFS of clean speech and uniform epoch strength which means the strength of excitation (SoE) is not considered) and the vocal tract system consisting of the MCCs obtained using MLSA filter from the speech with background music. The synthesized speech is plotted in Figure 5.1 (c). It can be seen that the music components are still present in the signal and the music in the background is still audible perceptually. Next, the speech is synthesized by using the excitation source signal in the form of an impulse train (consisting of impulses at obtained epoch

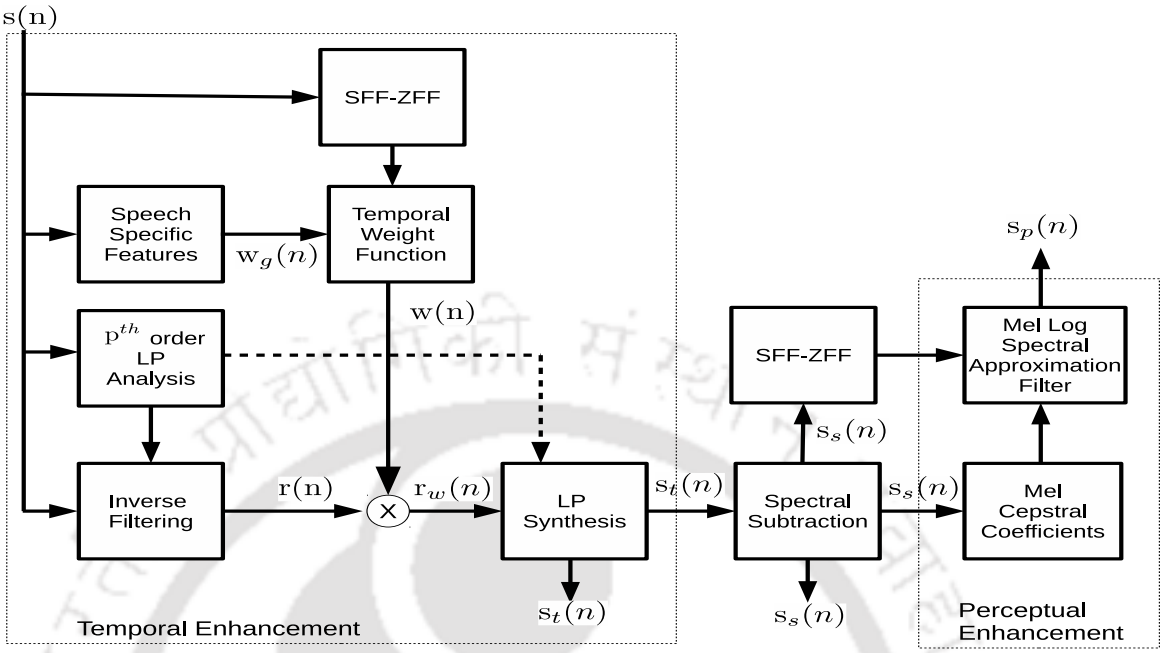


**Figure 5.2:** *Overall block diagram of the enhancement of speech with background music, where, $s(n)$ is the input speech with background music, $w_g(n)$ is the gross weight function derived using speech-specific features, $w(n)$ is the final temporal weight function, $r(n)$ is the LP residual signal, $r_w(n)$ is the temporally weighted LP residual, $s_t(n)$ is the temporally enhanced speech, $s_s(n)$ is the temporally and spectrally enhanced speech and $s_p(n)$ is the temporally, spectrally and perceptually enhanced output.*

locations from ZFFS of clean speech, and epoch strength also from the ZFFS, which means that the strength of excitation (SoE) is considered) and the same MCCs from the speech with background music for the vocal tract system. The synthesized signal is shown in Figure 5.1 (d). It can be observed that the music component has been reduced drastically. Perceptually, the effect of the music has also been reduced. This shows that the source information in terms of both the epoch locations and epoch strength is significant for enhancement and ideally, obtaining a source information which has epoch locations and epoch strength, similar to the one extracted from clean speech, would provide good quality enhancement of the speech with background music signal.

The rest of the work is organized as follows. The speech enhancement details is given in Section 5.2. The experimental details related to phone recognition are given in Section 5.3. Finally the conclusion is given in Section 5.4.

## 5.2 Speech Enhancement

The enhancement of speech with background music will be carried out in a number of sequential stages and the different stages can be observed in Figure 5.2.

### 5.2.1 Temporal Enhancement

The temporal enhancement involves deriving total weight function based on the analysis at the gross and fine levels [17] which in turn is used to modify the LP residual of the speech with background music. At the gross level, features such as the sum of first ten largest peaks, hilbert envelope of LP residual and the modulation spectrum which represent some aspects of speech production have been explored for the noisy speech in [17]. These features are used to derive the gross weight function which is similar to the VAD technique. In this work, since the speech with background music enhancement is carried out, the speech-specific features which have been shown to perform well for speech/music classification [96] have been used for deriving the gross weight function. These features include the normalized autocorrelation peak strength (NAPS) of zero frequency filtered signal (ZFFS), the hilbert envelope of LP residual the log mel spectrum energy and the modulation spectrum. The hilbert envelope of LP residual and the modulation spectrum which have been shown to perform well for speech/music classification, have also been used for deriving the gross weight function for the noisy speech enhancement in [17]. The overall behavior over a speech with background music of the features for deriving the gross weight function is shown in Figure 5.3. It can be observed that the feature values are high in the speech regions and lower in the music regions. Some errors also occur around the 0.1 s region due to the failure of some of the features to differentiate between speech and music which could be due to the similar nature of the music as that of speech. The log mel spectrum energy has almost the same nature as the sum of the first ten largest peaks of the DFT spectrum although the log mel spectrum energy has a higher value (Figure 5.3(d)). The four features in Figure 5.3 are summed up together, normalized and non-linear mapped using the mapping function [17]

$$w_g(n) = \frac{1}{1 + e^{-\lambda(s_i(n)-T)}} \tag{5.1}$$

where $\lambda$ is the slope parameter set to 20 [17], and $w_g(n)$ is the non-linearly mapped values of the normalized sum $s_i(n)$ and $T$ is the average value of $s_i(n)$. The $w_g(n)$ is termed as the gross weight function.

The fine weight function involves emphasizing the instants of significant excitation regions which are mostly the high signal to music ratio regions. These instants are also known as the glottal closure instants (GCI). A robust method is necessary to obtain these GCI locations so that they can act as anchor points for the enhancement of speech with background music regions. The existing zero

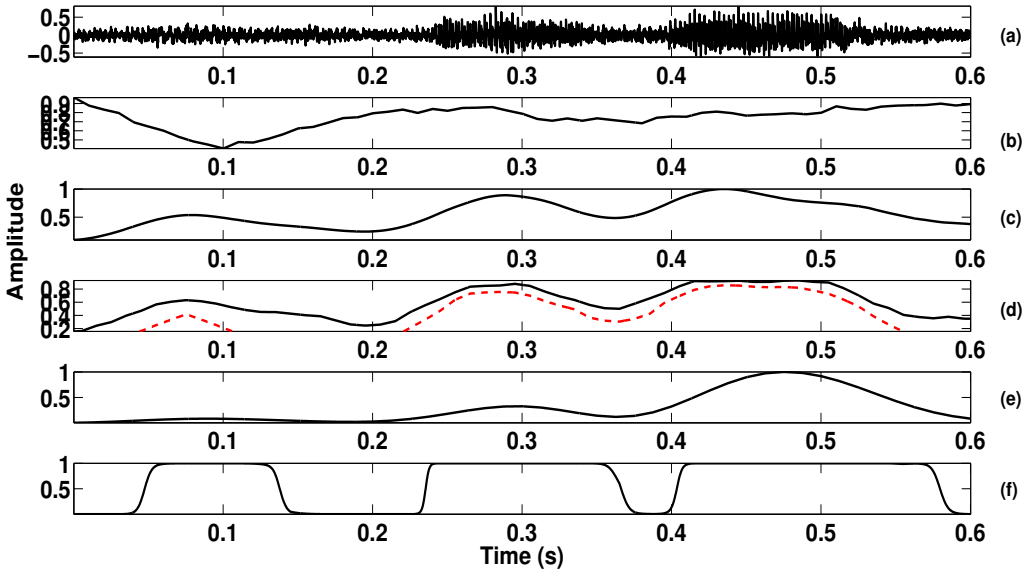


**Figure 5.3:** *Illustration of Gross Weight Function Derivation (a) Speech added with rock music (SNR=0 dB) (b) NAPS of ZFF (c) HE of LP Residual (d) Log Mel Spectrum Energy (continuous black) and sum of first ten largest peaks of DFT (dotted red) (e) Modulation Spectrum Energy (f) Gross Weight Function*

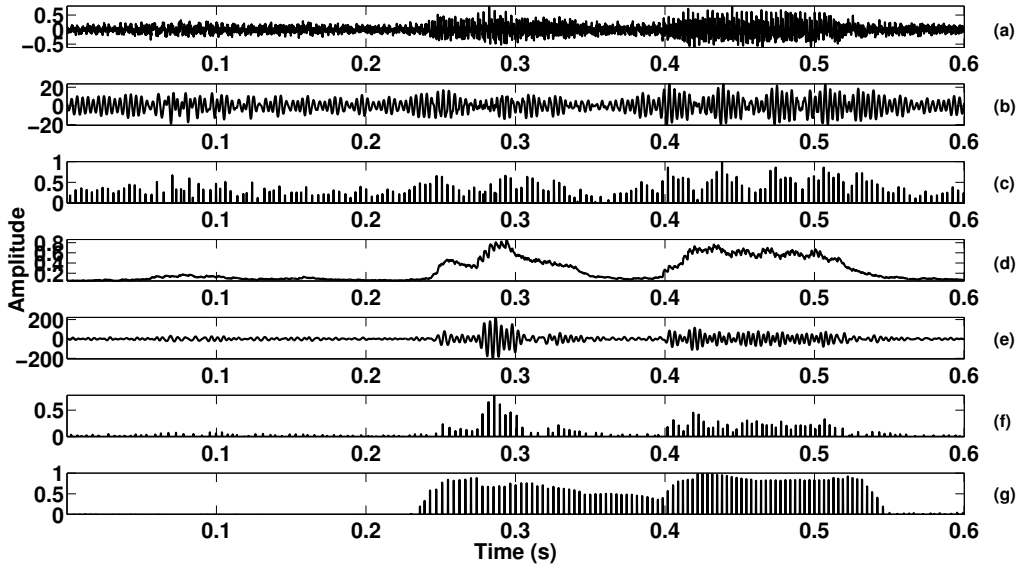


**Figure 5.4:** *Illustration of SFF-ZFF combination (a) Speech added with rock music (SNR=0 dB) (b) ZFF of speech with background music (c) Strength of Excitation (SoE) from ZFFS (d) Sum of the mean and variance of the amplitude envelopes from all frequencies obtained from SFF (e) SFF-ZFF of speech with background music (f) Strength of Excitation using SFF-ZFF (g) ZFF of clean speech*

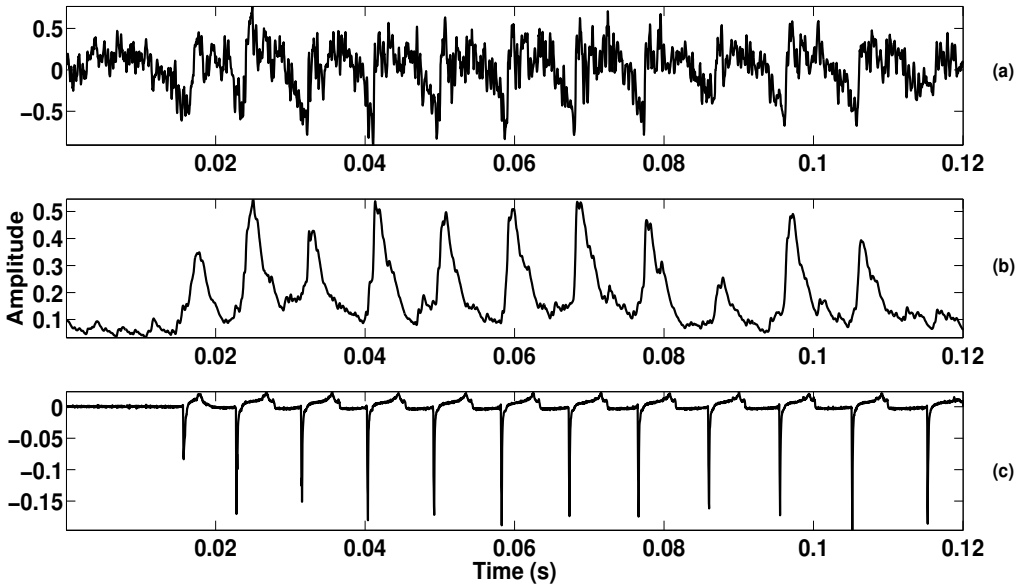


**Figure 5.5:** *Illustration of Sum of the mean and variance of the amplitude envelopes from all frequencies obtained from SFF (a) Speech added with rock music (SNR=0 dB) (b) Sum of the mean and variance of the amplitude envelopes from all frequencies obtained from SFF (c) differential electroglottograph (DEGG) of speech in (a), but without the addition of rock music.*

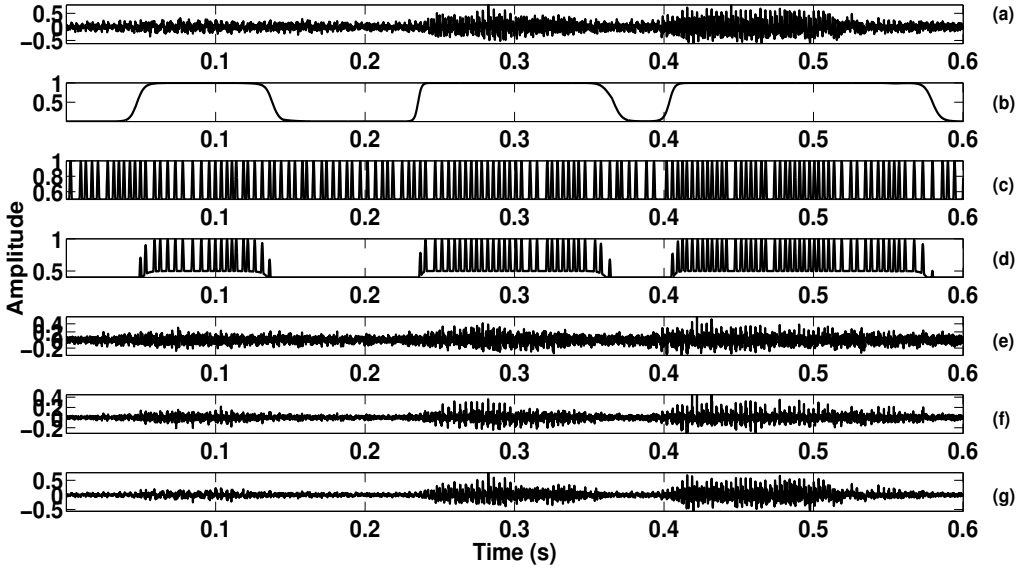


**Figure 5.6:** *Illustration of Fine Weight Function Derivation (a) Speech added with rock music (SNR=0 dB) (b) Gross Weight Function (c) Fine weight function using the Epoch locations obtained from SFF-ZFF (d) Overall Weight Function (e) LP Residual from speech with background music (f) Weighted LP Residual (g) Temporally Enhanced Speech*

frequency filtering (ZFF) [19] technique can be used to obtain the epoch locations. But the robustness of this method for speech with background music is lower considering the fact that some of the music signals have their frequencies around the fundamental frequency of speech resulting in spurious epoch locations as shown in Figure 5.4(c) which are the negative to positive zero crossings of the ZFF signal in Figure 5.4(b). Alternatively, to reduce the spurious epochs, a single frequency filtering (SFF) method proposed in [97] for speech/non-speech detection has been explored. This method derives the envelopes at a certain frequency range. The mean and standard deviation of the envelopes is computed over the frequency range and then summed. Speech has a specific source nature where the spectral energy is concentrated around the fundamental frequency and its harmonics. Additionally, the formant nature is also specific for speech. Music, on the other hand, may not have such spectral energy concentration and formant structure which means that if the mean and standard deviation is computed over a frequency range, the value of the mean and the standard deviation may be high for speech and lower for music. Also, the computation of the mean and the standard deviation will preserve the characteristics of speech such that the fundamental frequency and the formant structure information may still be present while the characteristics of the music will be de-weighted.

The resulting summed value of the mean and the standard deviation can be observed in Figure 5.5 (b) for a segment of speech added with rock music at a speech to music ratio (SMR) of 0 dB. It can be seen that the summed value contains peaks which correspond to the epoch locations and this can be verified by comparing with the differential electroglottograph (DEGG) signal of the same segment of speech but without the addition of rock music, shown in Figure 5.5 (c). These peaks are approximately aligned with the DEGG peaks within a glottal cycle. The presence of these peaks means that some information corresponding to the epoch locations is present in the summed signal and the epoch locations can be obtained by passing the summed signal through the zero frequency filter. The summed mean and standard deviation of the amplitude envelopes across frequencies obtained from the SFF shown in Figure 5.4(d) is passed through the ZFF to obtain the signal shown in Figure 5.4(e). Finally, the epoch locations which are obtained from the negative to positive zero crossings of the signal in Figure 5.4(e) is shown in Figure 5.4(f). Note that the spurious epochs have been reduced in Figure 5.4(f) compared to Figure 5.4(c) which shows the significance of using the SFF-ZFF.

The envelope of the signal at each frequency is obtained as follows [97]:

- Difference the speech signal $s(n)$ having sampling frequency $f_s$, $x(n) = s(n) - s(n-1)$

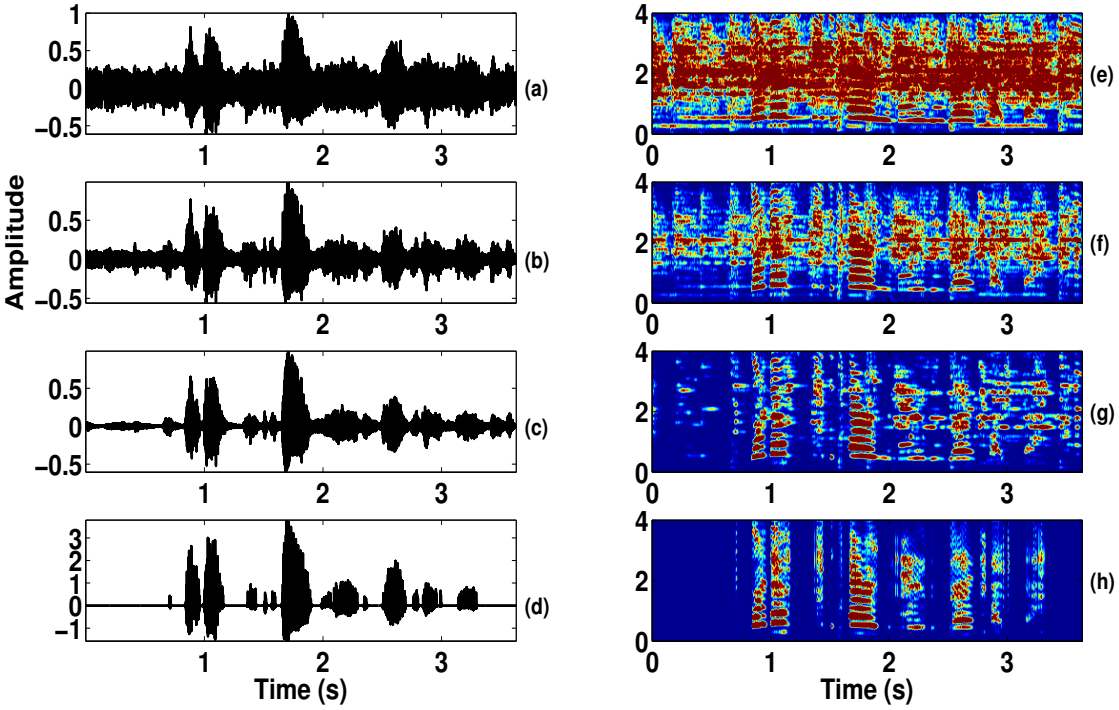


**Figure 5.7:** *Illustration of different stages of Enhancement (a) Speech added with rock music (SNR=0 dB) (b) Temporally Enhanced Speech (c) Temporally and Spectrally Enhanced Speech (d) Temporally, Spectrally and Perceptually Enhanced Speech (e) Spectrogram of (a) (f) Spectrogram of (b) (g) Spectrogram of (c) (h) Spectrogram of (d)*

- Multiply $x(n)$ by a complex sinusoid $e^{j\omega_k n}$, $x_k(n) = x(n)e^{j\omega_k n}$
- Multiplying the signal $x(n)$ by a complex sinusoid results in $X_k(\omega) = X(\omega - \omega_k)$, where $X_k(\omega)$ and $X(\omega)$ are spectra of $x_k(n)$ and $x(n)$, respectively.
- Pass the signal $x_k(n)$ through a single pole filter having a pole on the real axis at a distance of $r$ from the origin. The transfer function of the filter is given as

$$H(z) = \frac{1}{1 + rz^{-1}} \tag{5.2}$$

- The output of the filter is represented as,

$$y_k(n) = -ry_k(n-1) + x_k(n) \tag{5.3}$$

- The envelope of $y_k(n)$ is given as, $e_k(n) = \sqrt{y_{kr}^2(n) + y_{ki}^2(n)}$, where $y_{kr}^2(n)$ and $y_{ki}^2(n)$ are the real and imaginary components of $y_k(n)$.

Note that the location of the pole is at $z = -r$ which corresponds to half of the sampling frequency ($fs/2$). The filtering is done at $fs/2$ and the envelope corresponds to the envelope of the signal at a frequency of

$$f_m = \frac{fs}{2} - f_k, where f_k = \frac{\omega_k f_s}{2\pi} \tag{5.4}$$

This is the single frequency filtering approach of obtaining the envelope of the component at a frequency $f_m$.

The mean ($\mu(n)$) and standard deviation ($\sigma(n)$) of $e_k(n)$ is computed over a certain frequency range and the sum of the mean and standard deviation is computed ($\mu(n) + \sigma(n)$) to obtain the plot shown in Figure 5.4(d). This signal has a high value in the speech regions and lower value in the music regions. This signal is then passed through the ZFF to obtain the signal in Figure 5.4(e). The epoch locations can be obtained from the SFF-ZFF by considering the negative to positive zero crossings of the SFF-ZFF output. These epoch locations are used for deriving the fine weight function as in [17], where the epoch locations are convolved with a Hamming window which has a temporal duration of 3 ms, corresponding to the closed phase interval of the glottal cycle. Let the epoch locations be considered as a shifted train of impulses. The fine weight function $w_f(n)$ can be written as

$$w_f(n) = \left( \sum_{k=1}^{N_k} \delta(n - i_k) \right) * h_w(n) \tag{5.5}$$

where $N_k$ is the total number of epochs located, $i_k$ is estimated location of epoch. $w_f(n)$ is given a threshold value of $T$ to keep the distortion low because of overemphasized epoch locations in LP residual and is expressed as,

$$w_f(n) = \begin{cases} T, & \text{if} \quad w_f(n) < T \\ w_f(n), & otherwise \end{cases} \tag{5.6}$$

where $T$ is set as 0.5 in this work. It is to be noted that the temporal processing is not sensitive to T [17].

The final weight function $w(n)$ in Figure 5.6(d) is obtained by multiplying the gross weight function shown in Figure 5.6(b) with the fine weight function $w_f(n)$ in Figure 5.6(c) and is expressed as

$$w(n) = w_g(n) \times w_f(n) \tag{5.7}$$

The total weight function is multiplied by the LP residual signal shown in Figure 5.6(e) to obtain the weighted LP residual shown in Figure 5.6(f). The temporally enhanced speech signal can be obtained by synthesizing as follows:

$$S_t(z) = \frac{R_w(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{5.8}$$

where, $S_t(z)$ is the temporally enhanced speech and $R_w(z)$ is the weighted LP residual and $a_k$ are the LP filter coefficients. The plot of the speech which has been temporally enhanced for a short segment is shown in Figure 5.6(g). Similarly, the temporally enhanced speech on a single utterance is shown in Figure 5.7(b). It can be seen that the degradation has been reduced in the temporally enhanced speech as seen in the spectrogram of Figure 5.7(f).

### 5.2.2 Spectral Enhancement

The temporally enhanced speech is passed through a spectral enhancement module. In conventional spectral processing methods, the estimation of the short-term magnitude of the degradation and the degraded speech are performed first. The magnitude spectrum of the degraded speech is applied using a spectral gain function, to obtain the enhanced speech spectrum. The enhanced speech magnitude spectrum and the degraded speech phase spectrum are then combined to give an estimate of clean speech. Overlap-add (OLA) method is used for the time domain re-synthesis. In this work, the minimum mean square error of log-spectral amplitude estimator (MMSE-LSA) estimator is used which has a spectral gain function given by [99], as follows.

$$H(k) = \frac{\zeta_k}{1 + \zeta_k} exp(\frac{1}{2} \int_{v_k}^{\infty} \frac{e^{-x}}{x} dx) \tag{5.9}$$

where,

$$v_k = \frac{\zeta_k}{1 + \zeta_k} \gamma_k$$

$\zeta_k$ and $\gamma_k$ are a priori Speech to music ratio (SMR) and a posteriori SMR, respectively.

The plot of the temporally and spectrally enhanced speech for a single utterance is shown in Figure 5.7(c) where the degradation has been further reduced as seen in the spectrogram of Figure 5.7(g).

### 5.2.3 Perceptual Enhancement

The spectrally subtracted speech obtained above is passed through another enhancement module which is the perceptual based enhancement [18]. This type of enhancement is based on the cepstral analysis and synthesis on the mel frequency scale. This type of enhancement was motivated by the fact that the spectrum obtained from the Mel Cepstral Coefficients (MCCs) resembles the human auditory spectral resolution which gives higher resolution at lower frequencies and lower resolution at higher frequencies [100]. The MCCs are the Fourier cosine coefficients of the spectral envelope derived from the mel log spectrum. The approximate vocal tract response is obtained from the MCCs by using the MLSA filter [101] which adopts the adaptive algorithm. The MLSA filter's true spectrum for $m^{th}$ order MCCs $c(m)$ is given as,

$$H^{\alpha}(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}} \tag{5.10}$$

represents an all-pass function, which represents the mel-warped frequency characteristics and $\alpha$ is a coefficient corresponding to the mel-scale ($\alpha = 0.35$ for 10 kHz sampling rate)

$$\beta_{\alpha}(\Omega) = tan^{-1}\frac{1 - \alpha^2 sin(\Omega)}{(1 + \alpha^2)cos(\Omega) - 2\alpha} \tag{5.11}$$

where, $\alpha$ depends on the sampling frequency and $\beta_{\alpha}(\Omega)$ is the phase of all-pass function, the smooth spectral envelope $G_{\alpha}(\widehat{\Omega})$ of mel log spectrum is expressed as a polynomial function of order $M$ given by

$$G_{\alpha}(\widehat{\Omega}) = \sum_{m=0}^{M} C_{\alpha}(m)cos(m\widehat{\Omega}) \tag{5.12}$$

where, $\widehat{\Omega}$ is mel frequency scale given by $\beta_{\alpha}(\Omega)$ and $C_{\alpha}(m)$ are the cepstral coefficients of order $M$.

Initially, the 34-dimensional MCCs are computed on a frame which is windowed with a Hamming window of size 20 ms with a frame shift of 10 ms. The MLSA filter is used to compute the smooth spectral envelope from the MCCs by giving the best mean square approximation of the log spectrum envelope on the linear frequency scale. The smooth spectral envelope along with the excitation signal are then used to synthesize the speech signal. Normally the excitation signal consists of $F_0$ information along with the voiced/unvoiced decision. In this work, the excitation signal consists of the impulse

train derived from the SFF-ZFF. The SFF-ZFF gives the epoch locations which are high SMR regions. The strengths of the epochs are also obtained. These epoch locations along with their strengths have been used as an impulse train to represent the excitation signal. The impulse train is used directly as the source in order to suppress the other components of the signal which may, in particular, contain music. The temporally, spectrally and perceptually enhanced speech for a single utterance is shown in Figure 5.7 (d) along with its spectrogram in Figure 5.7 (h). It is to be noted that the degradation has been reduced to a minimum in this case as seen in the figure.

**Table 5.1:** *Phone Error Rate (PER) for synthesized speech tested on models trained with clean speech. In the table, 'SBM' indicates speech added with rock music in the background, 'CS' indicates clean speech, 'VTS' indicates vocal tract system, 'SoE' indicates strength of excitation, 'Source(CS with SoE), VTS (SBM)' indicates the speech synthesized by using the source from CS (along with SoE) and the VTS from SBM, 'Source(CS without SoE), VTS (SBM)' is the same as 'Source(CS with SoE), VTS (SBM)' but without considering the SoE for the source. 'Source (SBM with SoE), VTS (CS)' indicates the speech synthesized by using the source from SBM (along with SoE) and the VTS from CS.*

| | PER (%) | | | | |
|---|---|---|---|---|---|
| Model↓ | CS | SBM | Source (CS with SoE), VTS (SBM) | Source (CS without SoE), VTS (SBM) | Source (SBM with SoE), VTS (CS) |
| GMM | 21.1 | 84.6 | 70.0 | 81.4 | 31.3 |
| SGMM | 19.4 | 82.1 | 69.1 | 79.2 | 28.9 |
| DNN | 22.6 | 81.1 | 69.0 | 79.5 | 30.3 |

## 5.3 Experimental Evaluation

The temporally, spectrally and perceptually enhanced speech is tested based on the phone recognition accuracy. A phone recognizer system is built using KALDI toolkit [102]. The acoustic modeling is based on Gaussian Mixture Model (GMM), Subspace GMM (SGMM) and Deep Neural Network (DNN) models to from a GMM-HMM, SGMM-HMM, and DNN-HMM hybrid models respectively. The GMM based system is a widely used system and can be found in most of the previous literature on phone recognition. The details of the SGMM and DNN based acoustic models can be found in [103]. The models are trained on clean speech and tested on either the speech with background music or the enhanced speech. TIMIT database, as well as the broadcast news database, are used for training and testing the systems. Music is added to the TIMIT test samples at a speech to music ratio (SMR) of 0 dB. A total of 14 hours of data was used for the broadcast news in which 80% was used for training and the remaining 20% was used for testing. The 14 hours of data has been collected from Indian news channels which is recorded over Tata Sky. The data has been manually transcribed by a human

transcriber at word level and the dictionary has been obtained from the CMU sphinx website. The sampling rate was set to 16 kHz for all the audio samples.

### 5.3.1 GMM-HMM

The GMM-HMM system was implemented using KALDI toolkit. Training of cross-word tri-phone acoustic model is done. The state tying which is based on the decision tree is used. Each tri-phone is modeled by using a 3-state HMM with 16 diagonal-covariance Gaussian components per state. Silence and short pause are modeled using a 3-state HMM with 32 Gaussian components per state. Mel frequency Cepstral coefficients (MFCCs) are used as features each of 13 dimensions computed on a 20 ms Hamming windowed segment. The first and second order temporal derivatives are also used to form a total feature dimension of 39. The phone error rate (PER) is used as a measure of recognition performance.

### 5.3.2 SGMM-HMM

The SGMM-HMM system was also implemented in Kaldi and the number of gaussians used for training the universal background model (UBM) is 400. The number of leaves and Gaussians in the SGMM is chosen to be 9000 and 7000, respectively. Also, S=D=39 has been chosen.

### 5.3.3 DNN-HMM

The HMM-DNN system was implemented in Kaldi. The MFCC features are spliced over 4 frames with LDA+MLLT+fMLLR (40-dimensional features) are used as the input. This means that the input is spliced over 4 frames to the left and right of the central frame or 9 frames in total. The number of hidden layers is varied between 2 to 5. However, there is not much change in the PER as the amount of training data is moderate. It is given in Kaldi documentation that the 4 hidden layers are effective when 100 hours of speech data is available. The learning rate is initially selected to be 0.015 and then reduced to 0.002 in 20 epochs. Additional 10 epochs are employed after reducing the learning rate. Kaldi employs a preconditioned form of stochastic gradient descent (SGD). In this approach, instead of using a scalar learning rate, a matrix-valued learning rate is employed. This is motivated by the basic idea to reduce the learning rate in dimensions where the derivatives have a high variance. This approach, in turn, is to control instability and stop the parameters moving too fast in any one direction. The minibatch size for neural net training was selected as 128. The number of senones in CD-GMM-HMM system training is fixed to 2000 with 16 Gaussian mixture per state. The total

number of parameters trained happens to be 1.5 million. The same wordnet and the dictionary are employed in the decoding using the SGMM and the DNN based systems.

### 5.3.4 Results

Table 5.1 shows the results using the GMM-HMM, SGMM-HMM and DNN-HMM systems on clean speech as well as the speech with background music. The results when using the SGMM-HMM showed the best performance. The results when the clean speech is added with background music (in this case rock music) at a signal to music ratio of 0 dB is showed in the third column of the table and it can be observed that the phone error rate (PER) rises drastically when there is music addition. It was showed earlier the significance of the source information for the speech enhancement when the degradation is music. For a single file, the source was extracted from clean speech and the vocal tract system was extracted from the speech with background music. These two were used to synthesize the speech and the synthesized speech had most of the music components suppressed which showed the significance of the source information and it shows that if the source information obtained from the speech with background music is modified to the extent of having its behavior as the source information of the clean speech, the enhancement can be achieved and the music may be suppressed. The recognition result for all the files synthesized as just mentioned is given in the fourth column of Table 5.1. Note that there is a significant improvement in the PER when the source is extracted from the clean speech. This result serves as the baseline and indicates the best performance that can be achieved when modifying the source. This result is obtained for the case when the strength of excitation (SoE) is taken into account. As mentioned earlier, considering SoE means the strength of excitation is considered for generating the impulse train for the source signal and no SoE means the impulse train for the source signal is generated with a uniform strength of excitation. The result when the SoE is not taken into account is also displayed in the table. Note that not so good improvements are obtained when the SoE is not considered justifying the observations earlier in the introduction section regarding the SoE for a single speech file. The results when the source is extracted from speech with background music and the vocal tract system from the clean speech is shown in the last column of Table 5.1.

The results of the enhancement performed on the speech added with different music types are shown in the form of bar plots in Figure 5.8, 5.9 and 5.10 using GMM-HMM, SGMM-HMM, and DNN-HMM, respectively. The results on the enhanced files are marked as 'Enh' as shown in the

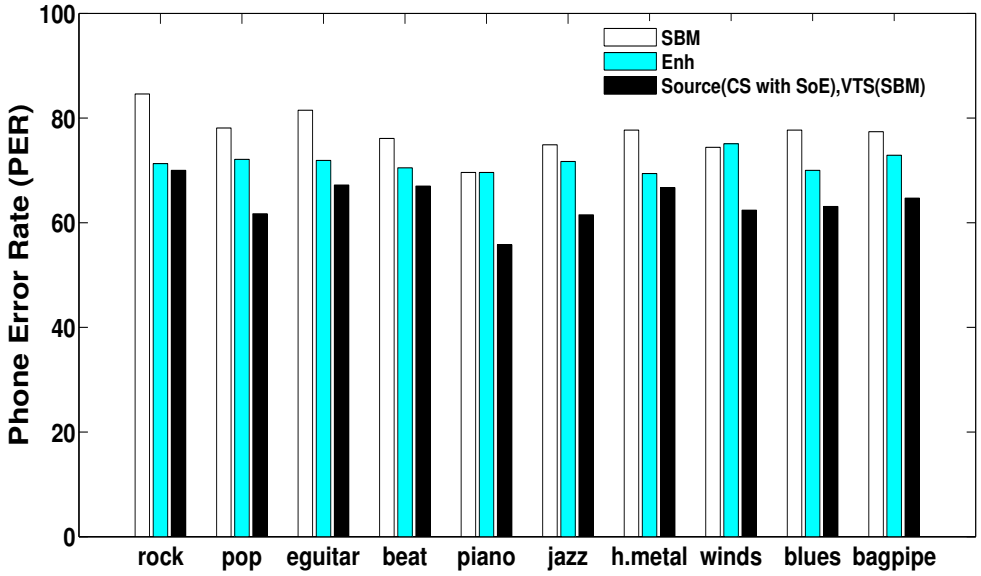


**Figure 5.8:** *Bar Plot showing the Phone Error Rate (PER) for synthesized speech tested on models trained with clean speech using GMM-HMM. In the figure, 'CS' indicates clean speech, 'VTS' indicates vocal tract system, 'SBM' indicates speech added with respective background music shown as labels on the x-axis, 'Enh' indicates the temporally, spectrally and perceptually enhanced speech and 'Source (CS with SoE), VTS (SBM)' indicates the speech synthesized by using the source (impulse train having impulses located at epoch locations extracted from the ZFFS of clean speech) and the vocal tract system (MCCs extracted from the speech with background music) along with consideration of the strength of excitation (SoE) for generating the impulse train for the source signal.*

bar plot. For comparison, the results where the source is extracted from clean speech and the vocal tract system extracted from the speech with background music is also shown in the barplot denoted as 'Source (CS with SoE), VTS (SBM)'. It is to be noted in the Figure 5.8, 5.9 and 5.10, that the performance of the combined temporally, spectrally and perceptually enhanced speech segments is good for the case of rock, electric guitar, beat music, heavy metal, and bagpipe music. This performance can be understood from the bar plot by noting the difference of the bar representing 'enh' and the bar representing 'Source (CS with SoE), VTS (SBM)'. The smaller the difference, the better is the performance, which means that the enhanced speech performance is close to the ideal case of having the source extracted from clean speech for synthesizing the speech signal. The fact that rock, electric guitar, beat music, heavy metal and bagpipe music gave good performances, means that the algorithm works well for the music degradation which is heavy in nature as in heavy metal or rock while it performs poorer for the case of soft music degradation as in piano or jazz.

Table 5.2 shows the results of the experiment performed to prove whether the PER improvements

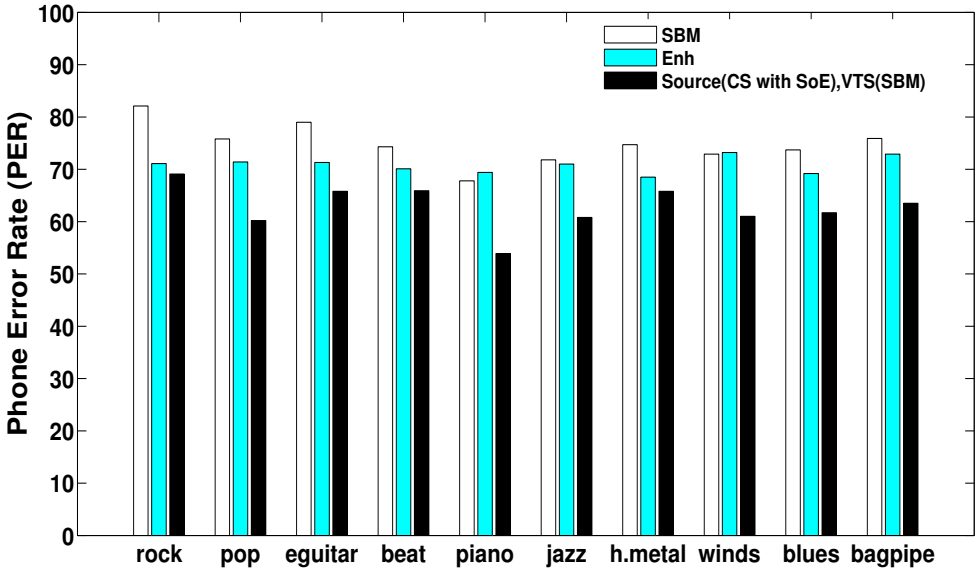


**Figure 5.9:** *Bar Plot showing the Phone Error Rate (PER) for synthesized speech tested on models trained with clean speech using SGMM-HMM. In the figure, 'CS' indicates clean speech, 'VTS' indicates vocal tract system, 'SBM' indicates speech added with respective background music shown as labels on the x-axis, 'Enh' indicates the temporally, spectrally and perceptually enhanced speech and 'Source (CS with SoE), VTS (SBM)' indicates the speech synthesized by using the source (impulse train having impulses located at epoch locations extracted from the ZFFS of clean speech) and the vocal tract system (MCCs extracted from the speech with background music) along with consideration of the strength of excitation (SoE) for generating the impulse train for the source signal.*

are obtained from the enhancement algorithms or whether it is only due to the silence model since the unvoiced portions of the speech with background music become silent frames after enhancement. The silence model is removed and the recognition experiments are conducted. The results show that even without the silence models, the improvements are obtained which indicates that the enhancement strategies do help in improving the recognition accuracy and the improvements are not only due to the introduction of silent frames in the unvoiced portions of the enhanced speech.

The experiments on broadcast audio are also performed and are shown in Table 5.3. Note that the improvements in the PER are obtained for this case as well. The improvements for the broadcast audio are not as much as the case when music is added (Figure 5.8, 5.9 and 5.10). This is because the speech to music ratio (SMR) in broadcast audio may have several varying levels depending on the scenario, as opposed to the case when music is added to speech, which is at a fixed SMR of 0 dB level. In addition, the music in broadcast audio may be of different types and may vary from time to time as opposed to the case in Figure 5.8, 5.9 and 5.10, when only a single fixed kind of music is added to

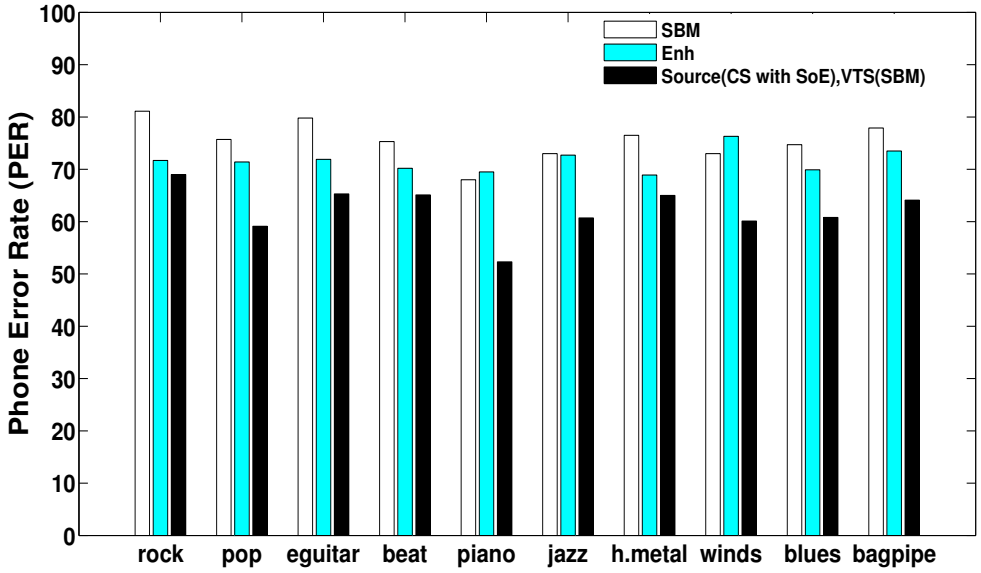


**Figure 5.10:** *Bar Plot showing the Phone Error Rate (PER) for synthesized speech tested on models trained with clean speech using DNN-HMM. In the figure, 'CS' indicates clean speech, 'VTS' indicates vocal tract system, 'SBM' indicates speech added with respective background music shown as labels on the x-axis, 'Enh' indicates the temporally, spectrally and perceptually enhanced speech and 'Source (CS with SoE), VTS (SBM)' indicates the speech synthesized by using the source (impulse train having impulses located at epoch locations extracted from the ZFFS of clean speech) and the vocal tract system (MCCs extracted from the speech with background music) along with consideration of the strength of excitation (SoE) for generating the impulse train for the source signal.*

**Table 5.2:** *Phone Error Rate (PER) for synthesized speech tested on models trained with clean speech (no silence model).In the table, 'CS' indicates clean speech, 'SBM' indicates speech added with rock music in the background, 'Enh' indicates the temporally, spectrally and perceptually enhanced speech and 'Source (CS with SoE), VTS (SBM)' indicates the speech synthesized by using the source (impulse train having impulses located at epoch locations extracted from the ZFFS of clean speech) and the vocal tract system (MCCs extracted from the speech with background music) along with consideration of the strength of excitation (SoE) for generating the impulse train for the source signal.*

| PER (%) | | | | | |
|---|---|---|---|---|---|
| **Music ↓** | **Model ↓** | **SBM** | **Enh** | **'Source (CS with SoE), VTS (SBM)'** | **Clean Speech** |
| **Rock** | **GMM** | 82.1 | 77.5 | 73.4 | 26.2 |
| | **SGMM** | 78.7 | 72.5 | 71.1 | 21.7 |
| | **DNN** | 79.0 | 77.3 | 71.8 | 24.9 |

speech.

**Table 5.3:** *Phone Error Rate (PER) for speech with background music, enhanced speech and clean speech, tested on models trained with clean speech taken from Broadcast Audio (BA). In the table, 'SBM' indicates speech added with background music, 'Enh' indicates the temporally, spectrally and perceptually enhanced speech*

| PER (%) | | | |
|---|---|---|---|
| **Model ↓** | **SBM** | **Enh** | **Clean Speech** |
| **GMM** | 81.92 | 75.13 | 31.17 |
| **SGMM** | 80.66 | 74.39 | 30.41 |
| **DNN** | 82.28 | 76.62 | 29.25 |

## 5.4 Summary

This work studied the impact of source information for speech enhancement resulting in improved phone recognition of speech with background music regions. The SFF-ZFF representing the source information was used in the temporal and perceptual enhancement steps. This method gave improved performance in terms of the PER for the phone recognition of speech with background music particularly when the degradation consisted of heavy music types while the performance reduces for the soft music types. The algorithm was also tested on the broadcast news and improved PER was observed in this case too.

# 6

# Significance of Preprocessing Methods for Phone Recognition in Broadcast Audio

Contents

### Objective

*This work explores the phone transcription of broadcast audio, in particular, the anchor speakers' speech segments, which is done in two steps. The first step is the preprocessing step which involves the different stages of classification and enhancement to obtain clean speech segments. This step is necessary considering the complex nature of the broadcast audio even for anchor speakers' speech, which may contain voice-over, as well as new headlines and these segments, in turn, may contain background music. The second step is the phone recognition step which basically expects clean speech as input to do the speech to phone transcription. The experiments in this work are performed in a sequential manner wherein the audio to the different modules are passed and the output of one module acts as the input to another. The final phone recognition accuracy is determined based on both the ideal and practical performances of the preprocessing steps. It is observed that the preprocessing steps contribute to the improvements of the overall phone recognition performance.*

## 6.1 Introduction

The automatic transcription of broadcast audio has been attempted in several ways in the literature [4, 6, 8, 10, 63]. Due to the complexity of the broadcast audio, preprocessing steps are necessary. The preprocessing steps aim at obtaining homogeneous segments. These homogeneous segments then provide for a better performance of the speech recognition or the speaker adaptation system. The complexity of the preprocessing steps is higher as the number of broadcast audio scenarios considered for the transcription is higher. In this work, the complexity of the preprocessing steps is reduced by considering the phone transcription of certain scenarios in the broadcast audio. These scenarios correspond mostly to the anchor speakers' speech. By transcribing the anchor speakers' speech segments, already a lot of information to a particular event would have been captured with respect to speech for multimedia related applications like audio summary [1]. The anchor speaker is mostly present in the studio and consists mostly of clean speech. However, there may also be voice-over speech in between the anchor speakers' speech as well as the news headlines in the beginning and end of the news show. These segments generally contain speech with background music. Hence by considering only the anchor speakers' speech, these different scenarios have to still be considered and some preprocessing step is necessary. The preprocessing steps employed for the anchor speakers' speech segments are simpler.

This simpler level of complexity allows for the exploration of other robust features for the preprocessing stage as opposed to the mel frequency cepstral coefficients (MFCCs) which are the general features used for the preprocessing stages of the general broadcast audio transcription in previous work.

The preprocessing stages considered consists of the speech/music classification [11–14] to remove any unwanted music segments. The features employed for this step consists of speech-specific features [96] which are more robust for this task compared to the general time and frequency domain features. These features localize the speech regions better and deviate significantly in the music regions thus providing a sense of discrimination between speech and music. The clean speech, as well as the speech with background music segments, will be classified in the class of speech due to the speech-specific nature of the features. Passing the segments containing both clean speech and speech with background music segments through the phone recognizer trained on clean speech may result in a lot of errors as will be demonstrated in the experiments later in this work. Hence another form of classification is necessary to classify between clean speech and speech with background music segments. The concept of speech-specific feature is exploited in this module as well. The feature for this task is defined in such a way that it behaves differently in clean speech and speech with background music considering the presence of music in the latter. Even though the production of speech is similar in the two classes, the addition of music after the speech has been produced in the speech with background music regions causes the deviation in the value of the speech-specific features in this case. The speech with background music regions are passed through an enhancement module to obtain the clean speech segments so that the acoustic mismatch between the training and testing speech is reduced. The clean speech, as well as the enhanced speech, are passed through the phone recognizer to obtain the phone transcription. Note that the models of the phone recognizer are trained on the clean speech since these segments are abundant in the broadcast audio and also to have a simpler speech recognition system.

Based on the motivation of the use of the different modules required for the transcription of broadcast audio described above, the next few sections will focus on the detailed description of each of these modules and how they impact the overall phone recognition accuracy. The rest of the work is described as follows. The preprocessing modules are described in Section 6.2. The phone recognition module is described in Section 6.3. Finally, the results are given in Section 6.4.

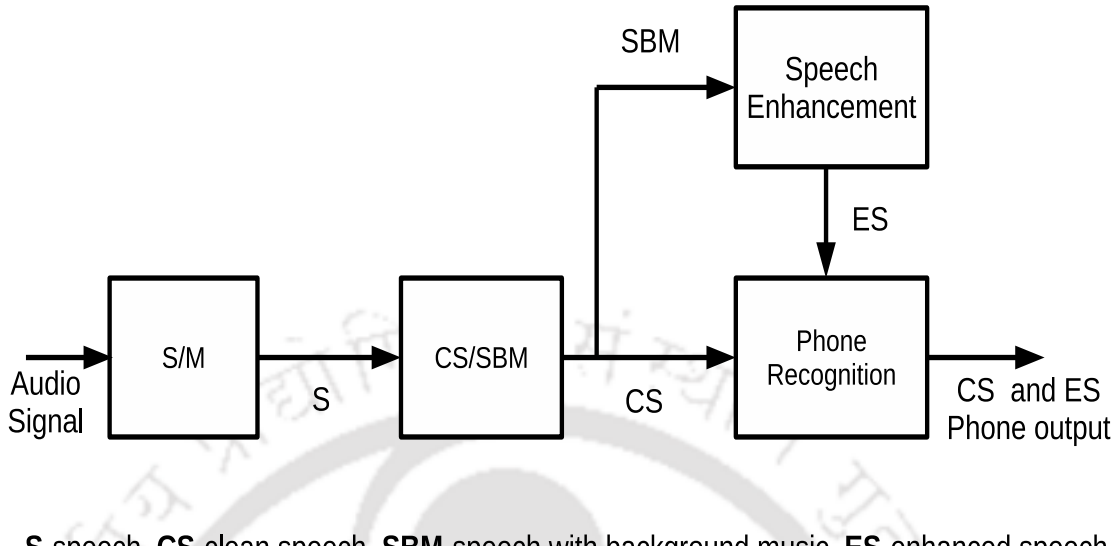


**Figure 6.1:** *Overall block diagram for the transcription of broadcast audio*

## 6.2 Modules necessary for Preprocessing

The transcription of broadcast audio in this work is done in a sequential manner and the method is proposed as shown in the Figure 6.1. The incoming audio stream is passed through the different blocks of preprocessing consisting of the speech/music classification module, the clean speech/speech with background music module and the speech enhancement module. The final module is the phone recognition module. The details of all these modules are given in the following sections.

### 6.2.1 Speech/Music Classification

This is the first preprocessing module for the automatic transcription proposed in this work. The incoming audio stream is given as input to this module. The speech/music classification is employed here in order to remove any unwanted music components in the broadcast audio which may introduce errors in the phone recognition system. The speech/music classification is a popular task and has been used in plenty of applications. This task generally involves the use of two stages which are the feature extraction stage and the classification stage. The features are generally defined based on the time domain [11–13] and frequency domain characteristics [14] while the standard classifiers like K-Nearest Neighbor (KNN) [12], Gaussian Mixture Model (GMM) [12], and Support Vector Machine (SVM) [82, 83] are utilized for the classification. In this work, the SVM classifier will be used for

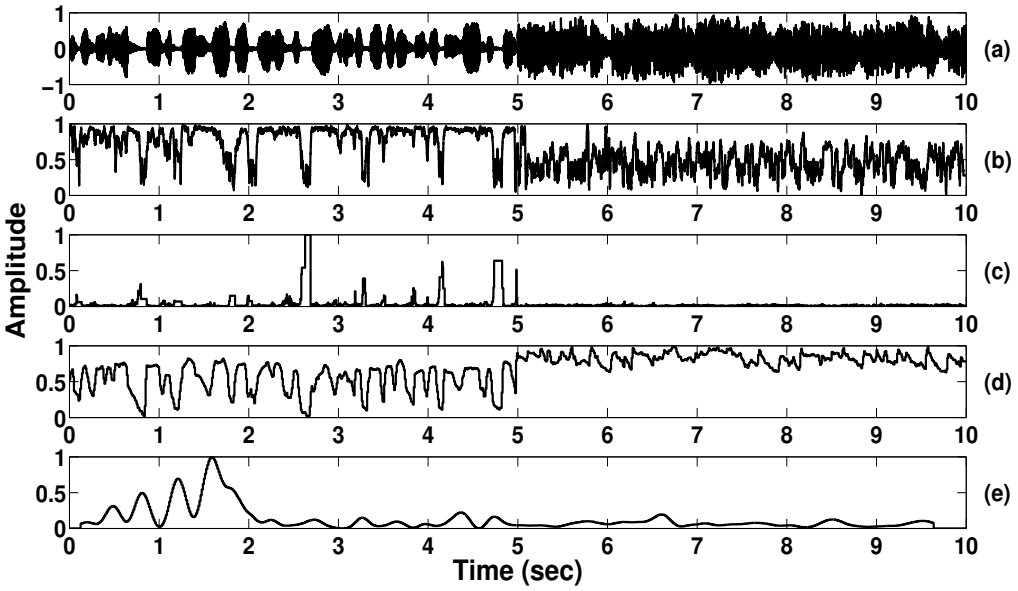


**Figure 6.2:** *Speech-Specific Features (a) Audio signal with first 5 s of speech and next 5 s of instrumental music (b) Normalized autocorrelation peak strength of zero frequency filtered signal (ZFFS) (c) Peak to side lobe ratio of hilbert envelope of linear prediction residual (d) Log mel spectrum energy (e) Modulation spectrum energy.*

the classification since it has been shown to perform well in previous work and also because SVM is suitable for binary classification. The features which are explored in this work are based on the speech-specific nature and they will be called as speech-specific features [96]. These features were developed based on the following idea. The speech production system has been studied extensively in terms of the source, vocal tract system and the suprasegmental characteristics since the mechanism of speech production is similar among humans. The music signal, on the other hand, can be produced from different sources and a specific model for music is difficult to define. Hence defining the features in terms of speech which has been understood well in terms of the production characteristics is a better idea hoping that the speech-specific features deviate significantly in the music segments thus achieving some sort of discrimination between speech and music.

The speech-specific features are the features defined in terms of the source, vocal tract system, and the suprasegmental information. The normalized autocorrelation peak strength (NAPS) of zero frequency filtered signal (ZFFS) of speech and the peak-to-sidelobe ratio (PSR) of hilbert envelope (HE) of linear prediction (LP) residual of speech are the two features which represent the source of speech production. The NAPS of ZFFS represents the quasi-periodic nature of the source signal while the PSR of HE of LP residual represents the impulsive nature of the source signal. These

characteristics of the source signal may be different for music since the production aspect of music is different from speech. A plot of the NAPS of ZFFS and the PSR of HE of LP residual for a segment of audio which contains first 5 s of speech and next 5 s of music is shown in Figure 6.2 (c) and (d). It can be observed that the nature of the feature for speech and music is different as expected and thereby giving a sort of discrimination between speech and music. Another speech-specific feature explored is the log mel spectrum energy which represents the vocal tract system of speech production. This feature was developed based on the motivation that speech contains vowel-like sounds which have most of their energy concentrated in the lower frequency region ($\leq 2kHz$) of the spectrum while the music does not have this kind of nature. A plot of the log mel spectrum energy for the audio signal is also given in Figure 6.2 (e). The discrimination between speech and music can also be observed using this feature. The modulation spectrum energy feature is the feature explored and is based on the suprasegmental characteristic of speech production. More specifically, this feature represents the syllabic rate of speech. Speech has a characteristic energy peak around the 4 Hz syllabic rate while music does not have this particular nature. This mechanism is exploited to define this feature and it is plotted in Figure 6.2 (f) which shows that the modulation spectrum energy feature does have some discrimination between speech and music. These features have been shown to perform well when used together in concatenating form [96].

In this work, these features will also be concatenated together and given as input to the SVM classifier. The labels representing speech or music will thus be obtained. These labels obtained will be useful for the future stages of preprocessing. The labels containing speech regions from this module are noted for use in the next level of preprocessing while the music is discarded. The speech-specific nature of the features causes all kinds of speech segments to be classified as speech, as the features localize around the speech regions while discarding the music regions. The anchor speakers' speech may also contain segments such as speech with background music, as in voice-over cases. These segments will be classified as speech in the speech/music classification system and hence the segments obtained from this system will be either clean speech or speech with background music. Passing these segments directly through a phone recognizer may introduce errors. Hence another level of classification is necessary to segment these two regions since the phone recognizer will be built using clean speech. After classifying clean speech and speech with background music, the clean speech can directly be passed through the phone recognizer. The speech with background music obtained can then be passed

through an enhancement module before passing through the phone recognizer trained on clean speech.

### 6.2.2 Clean Speech/Speech with Background Music Classification

The clean speech/speech with background music classification module takes the input from the speech output labels of the speech/music classification system. This exploration of this kind of system in the previous literature is very limited although there are some audio classification tasks where these classes appear [31,88]. This task is also similar to the one earlier and consists of the feature extraction stage followed by the classification stage. For the classification scheme in this work, SVM has been used to maintain uniformity of the classification of the preprocessing stages. The main thing is the use of different kinds of speech-specific features and these features should have different characteristics to the one defined earlier for speech/music classification. The speech-specific features defined for this task is mostly in terms of the vocal tract system. The average and relative spectral characteristics of the vocal tract system are explored for this work. The mel frequency cepstral coefficients (MFCCs) represent the average spectral characteristics while the sum of the spectral contrast on the Hilbert envelope of the Numerator of Group Delay (HNGD) spectrum ( a better representation of the vocal tract system), represents the relative spectral characteristics. These two features complement each other in their ability to classify speech and speech with background music thus expecting a good classification accuracy between the two classes.

An example of the behavior of the spectral contrast based features is shown in Figure 6.3. The sum of the spectral contrast has been computed on the HNGD spectrum. Other parameters are also computed while computing the sum of the spectral contrast and these consists of the sum of the spectral peaks and the sum of the spectral valleys. The plot of these features has also been shown in the figure. It can be observed that the sum of the spectral contrast feature has the ability to discriminate between the clean speech and speech with background music. The details of the spectral contrast computation can be found in [90] and the details about the HNGD can be found in [91]. The two labels obtained from this module will correspond to the one containing clean speech and the other containing speech with background music. These labels are useful for the next level of preprocessing. The clean speech labels identify the clean speech regions and these regions are passed directly through the phone recognizer trained on clean speech. The labels which identify the speech with background music are useful for the next preprocessing module which is the speech enhancement module. These labels extract the speech with background music regions which are then processed so as to obtain the

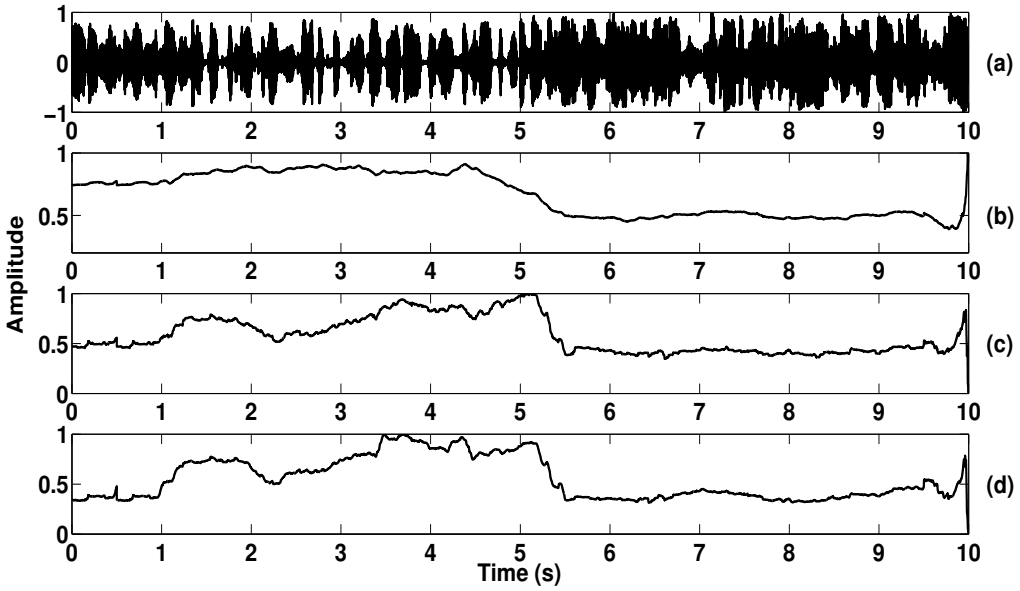


**Figure 6.3:** *Speech-Specific Features (a) Audio signal with first 5 s of speech and next 5 s of speech with background music (b) Smoothed Sum of Spectral Contrast of HNGD (c) Smoothed Sum of Spectral Peaks of HNGD (d) Smoothed Sum of the Spectral Valleys of HNGD.*

enhanced speech, in which the desirable output is obtaining the enhanced speech having the same characteristics as that of clean speech. These enhanced files can then be passed through the phone recognizer, to obtain a better transcription accuracy than passing the speech with background music directly which introduces acoustic mismatch. The details of the enhancement module are given in the next section.

### 6.2.3 Enhancement of Speech with Background Music Regions

The speech with background music regions are passed through the enhancement module which is developed by taking into account the speech-specific feature which is the source feature. There have been several attempts to perform enhancement on the speech with background noise regions. These include the temporal based enhancement [17, 71], the spectral based enhancement [17, 71] and the perceptual based enhancement [18]. These steps are generally performed in a combined sequential manner and have shown to give significantly good results on the method of foreground speech enhancement in [18]. The speech degraded with noise is passed sequentially through the temporal, spectral and perceptual based modules for obtaining an enhanced speech. In this work, the same sequential processing of the speech with background music segment is done, with the additional development

of modifying the source information, particularly in the temporal and perceptual based enhancement modules.

A characteristic of the source is exploited which is in terms of the epoch locations. These epoch locations can be easily identified on the clean speech by using some of the standard methods like zero frequency filtering (ZFF) [19]. However, the ZFF degrades significantly when the speech with background music is passed through it giving epoch locations along with plenty of spurious epochs. A single frequency filter-zero frequency filter (SFF-ZFF) is developed for this work in which the epoch locations are obtained from this method. The sum of the mean and standard deviation of the component envelopes computed across frequencies obtained using the SFF de-weights the music components. The reason is that the spectral energy of speech is concentrated around the fundamental frequency and its harmonics, which means that if the mean of the component envelopes is computed, its value will be high for speech and low for music. Additionally, the speech also has a specific formant nature which means that if the standard deviation is computed, the value will be high for speech and low for music. The low values of mean and standard deviation for music are because this signal has different characteristics in terms of the fundamental frequency and formant structure, compared to speech. This SFF has been shown to contain epoch information in [98]. This epoch information can be extracted by passing the SFF output through the ZFF which is why it is named as SFF-ZFF in this work.

The temporal based enhancement involves obtaining the weight function for modifying the linear prediction (LP) residual [17] of the speech with background music. This weight function is a combination of the gross weight function and the fine weight function. In this work, the gross weight function has been obtained by using features developed in [96]. The fine weight function requires robust epoch locations and the epoch locations obtained from the SFF-ZFF has been used. The gross and the fine weight function are then multiplied and used to weight the LP residual. The weighted LP residual is then used to synthesize the enhanced speech signal. The temporally enhanced speech is given in Figure 6.4(b). The temporally enhanced speech is passed through a spectral based enhancement module which involves modifying the magnitude spectra of the degraded speech using a spectral gain function. The minimum mean square error of log-spectral amplitude (MMSE-LSA) estimator is used in this work having a spectral gain function given in [99]. The temporally and spectrally enhanced speech is given in Figure 6.4(c).

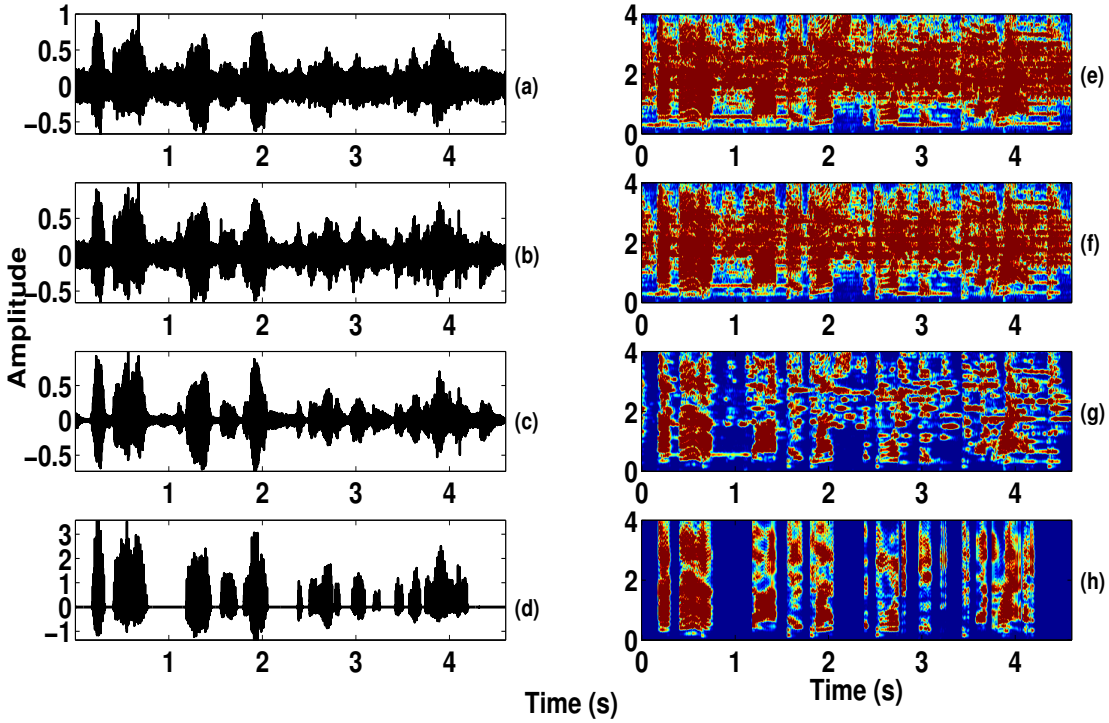


**Figure 6.4:** *Illustration of different stages of Enhancement (a) Speech added with rock music (SNR=0 dB) (b) Temporally Enhanced Speech (c) Temporally and Spectrally Enhanced Speech (d) Temporally, Spectrally and Perceptually Enhanced Speech (e) Spectrogram of (a) (f) Spectrogram of (b) (g) Spectrogram of (c) (h) Spectrogram of (d).*

The temporally and spectrally enhanced speech is then passed through a perceptual based enhancement module. The perceptual based enhancement module consists of a mel log spectral approximation (MLSA) filter, wherein the smooth spectral envelope is computed using this filter from the mel cepstral coefficients (MCCs). The excitation is also required along with the smooth spectral envelope to obtain the synthesized speech. This excitation is usually in terms of the $F_0$ information with the voiced/unvoiced decision. In this work, the SFF-ZFF is used for generating the epoch locations along with their strengths, which are then used to generate the impulse train which acts as the excitation signal for the perceptual based enhancement module. Note that the strengths of the epochs are considered for the impulse train generation. The temporally, spectrally and perceptually enhanced speech is given in Figure 6.4(d).

## 6.3 Phone Recognition

The phone recognition system in this work is built using the three systems which are the gaussian mixture model-hidden markov model (GMM-HMM), the subspace gaussian mixture model-hidden markov model (SGMM-HMM) and the deep neural network-hidden markov model (DNN-HMM) based systems. The input to the systems is in terms of the mel frequency cepstral coefficients (MFCCs). These systems have been implemented in Kaldi toolkit [102] and the parameters of the various systems and the details of the systems can be found in [103]. The models of the systems are trained on clean speech and the testing is done on the clean speech, speech with background music and the enhanced speech.

## 6.4 Results and Discussion

The database used for the experiments include the TIMIT database as well as the broadcast news database. For the TIMIT database, initially, the clean speech is added with rock music (taken from the GTZAN database [96]) at a speech to music ratio (SMR) of 0 dB. Next, the clean speech, the speech added with background music and music of same length are concatenated to form an audio signal. The broadcast news consists of the audio samples recorded from the Indian broadcast news channels where the samples which contain clean speech, speech with background music and music are selected. A total of 14 hours of data was used for the broadcast news in which 80% was used for training and the remaining 20% was used for testing. The sampling rate was taken at 16 kHz for all the audio samples. The overall process of the phone recognition with or without the use of the preprocessing stages for a single audio file can be seen in Figure 6.5 and Figure 6.6. It can be observed from the two figures that using the preprocessing stages the overall accuracy of the phones transcribed for the audio file is improved, especially in the enhanced speech with background music segments.

Initially, the phone recognition is performed separately on the different scenarios like clean speech, speech with background music and enhanced speech for the TIMIT database and the broadcast audio database. The results can be seen in Table 6.1 and Table 6.2. The experiments are then performed based on the ideal and practical conditions. The ideal conditions are when some of the preprocessing tasks like the classification task are assumed to be 100 % accurate and the overall performance of the phone recognition is analyzed. On the other hand, the practical conditions are when the classification tasks are performed using the given features and classifiers and the performance of the phone recog-

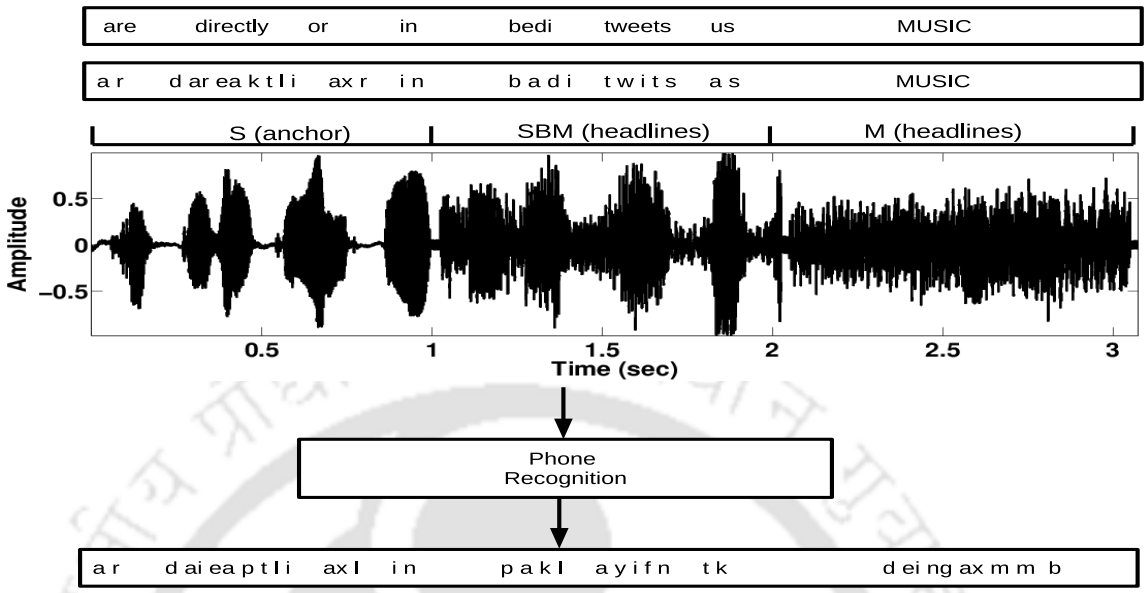


**Figure 6.5:** *Figure illustrating phone recognition of audio without any preprocessing. The word level and the phone level ground truth is shown at the top.*

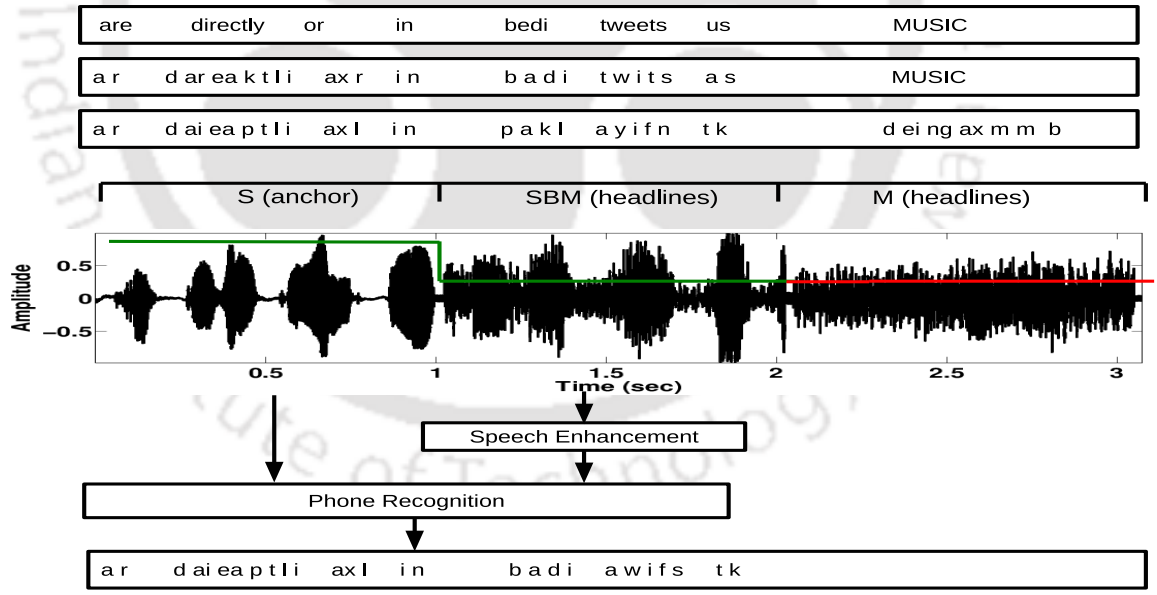


**Figure 6.6:** *Figure illustrating phone recognition of audio with preprocessing. The word level ground truth, the phone level ground truth and the output of the phone recognizer without using preprocessing is shown at the top (one after the other)*

nition is recorded. This is done to understand the contribution of each of the steps for the overall performance. Firstly the audio files which have been created using the TIMIT database are analyzed.

**Table 6.1:** *Phone Error Rate (PER) for clean speech, speech with background music (SBM) and enhanced (Enh) speech segments of the TIMIT database. The speech with background music is the same clean speech but added with rock music at a speech to music ratio of 0 dB.*

| PER (%) | | | |
|---|---|---|---|
| **Model ↓** | **Clean Speech** | **SBM** | **Enh** |
| **GMM** | 21.1 | **84.6** | **71.3** |
| **SGMM** | 19.4 | **82.1** | **71.1** |
| **DNN** | 22.6 | **81.1** | **71.7** |

**Table 6.2:** *Phone Error Rate (PER) for clean speech, speech with background music (SBM) and enhanced (Enh) speech segments of the broadcast audio database.*

| PER (%) | | | |
|---|---|---|---|
| **Model ↓** | **Clean Speech** | **SBM** | **Enh** |
| **GMM** | 31.17 | **81.92** | **75.13** |
| **SGMM** | 30.41 | **80.66** | **74.39** |
| **DNN** | 29.25 | **82.28** | **76.62** |

The audio sample which consists of clean speech, speech with background music and music is given as input directly to the phone recognition system to obtain the performance of degraded speech. The results can be found in Table 6.3 labeled as 'Audio'. It can be observed that the performance of the phone recognizer falls drastically with a high phone error rate (PER) as compared to the clean speech case shown in Table 6.1. This is because of the presence of the speech with background music and music, which introduces acoustic mismatch between the training and testing data since the training of the models has been done using clean speech.

In the next experiment, the audio constructed using the TIMIT database is assumed to be passed through the speech/music classification system and it is also assumed that the system gives a 100 % accuracy. This means that only the clean speech and speech with background music portions of the audio are passed through the phone recognizer. The performance of the recognizer improves. This is because the acoustic mismatch has been reduced slightly by the removal of the music components. The

**Table 6.3:** *PER for synthesized speech from TIMIT database tested on models trained with clean speech*

| TIMIT | PER (%) | | |
|---|---|---|---|
| Model↓ | Audio | S/M Classification (Ideal) | S/M Classification(Practical) |
| GMM | 69.36 | 51.74 | 55.54 |
| SGMM | 68.16 | 49.88 | 56.62 |
| DNN | 67.50 | 50.11 | 57.33 |
| Model↓ | S/M Classification (Ideal) | S/SBM Classification (Ideal) | S/SBM Classification(Practical) |
| GMM | 51.74 | 46.91 | 49.21 |
| SGMM | 49.88 | 45.73 | 48.73 |
| DNN | 50.11 | 45.88 | 48.97 |
| Model↓ | S/M Classification (Practical) | S/SBM Classification (Ideal) | S/SBM Classification(Practical) |
| GMM | 55.54 | 52.33 | 53.12 |
| SGMM | 56.62 | 51.44 | 52.28 |
| DNN | 57.33 | 52.09 | 52.77 |

practical conditions are then considered wherein first the audio from TIMIT database is passed through the speech/music classification system using the features and classifiers mentioned earlier. The labels obtained from this system are then used to retain the speech regions and remove the music regions. Since the practical performance of the speech/music classification is lesser than 100 % as opposed to the ideal case, some music components will be present in the output of the speech/music classification system. This results in slightly degrading the performance of the phone recognizer compared to the case of using an ideal speech/music classification system. However, the overall performance of the phone recognizer is better than the case of degraded speech as shown in Table 6.3.

Next, the labels of the speech/music classification system are used to retain the speech components while removing the music components. The speech with background music portions will be classified as speech due to the speech-specific nature of the features. The resultant audio output of the speech/music classification is expected to contain clean speech and speech with background music, although some music components are also present due to the accuracy of the speech/music classification having a value practically less than the ideal case. The audio obtained from the speech/music classification obtained in the ideal condition is then passed through a clean speech/speech with background music system and it is assumed that this system also has 100 % accuracy ideally. The clean speech is

passed through the phone recognition system and the speech with background music is passed through the enhancement module before passing it through the phone recognizer. The overall performance is given in Table 6.3 labeled as 'S/SBM Classification (Ideal)'. It can be seen that the performance has improved which shows that the enhancement technique has contributed to improving the overall accuracy. The practical case of using the clean speech/speech with background music system involves using speech-specific features and classifiers mentioned earlier. These features have different speech-specific characteristics compared to the features used for speech/music classification system. The output of this module is labeled containing clean speech and speech with background music. The speech with background music portions are enhanced. The results of the phone recognition for the practical case is labeled as 'S/SBM Classification(Practical)' in the Table 6.3. It can be observed that the performance slightly degrades compared to the ideal case of using the S/SBM classification. This is because some of the speech with background music samples may be classified as speech and they may not be enhanced. However, the overall performance of the phone recognition accuracy has improved after S/SBM classification compared to the degraded case which shows the significance of the preprocessing systems for the phone recognition of the audio files.

**Table 6.4:** *PER for broadcast audio samples tested on models trained with clean speech*

| BN | PER (%) | | |
|---|---|---|---|
| Model↓ | Audio | S/M Classification (Ideal) | S/M Classification(Practical) |
| GMM | 72.25 | 55.44 | 58.72 |
| SGMM | 70.35 | 54.36 | 57.64 |
| DNN | 70.51 | 54.93 | 56.99 |
| Model↓ | S/M Classification (Ideal) | S/SBM Classification (Ideal) | S/SBM Classification(Practical) |
| GMM | 55.44 | 52.43 | 53.39 |
| SGMM | 54.36 | 51.33 | 52.15 |
| DNN | 54.93 | 51.16 | 52.37 |
| Model↓ | S/M Classification (Practical) | S/SBM Classification (Ideal) | S/SBM Classification(Practical) |
| GMM | 58.72 | 54.11 | 56.64 |
| SGMM | 57.64 | 53.29 | 54.72 |
| DNN | 56.99 | 52.88 | 54.60 |

The practical speech/music classification output is passed as input to the clean speech/speech with

background music classification for both ideal and practical cases. As expected the results degrade slightly due to the presence of music components and these get propagated to the last stages thereby reducing the accuracy of the phone recognition system. The overall phone recognition accuracy after passing through the practical speech/music classification system, practical clean speech/speech with background music classification and the enhancement system is given in the last row and the last column of Table 6.3 marked as 'S/SBM Classification(Practical)' in the table. The overall phone accuracy is better than the degraded case which shows the significance of the preprocessing systems for the phone recognition of audio containing clean speech, speech with background music and music. The experiments are repeated on segments of broadcast audio. The broadcast audio segments are chosen such that they contain clean speech, speech with background music and music of same length. These segments are tested in the same way as above and shown in Table 6.4. Note that the trend in the results are similar as for the TIMIT database although the performance is slightly lower than the TIMIT case.

Note that the experiments in the previous cases were on the synthetically generated data, wherein the same length was chosen for the clean speech, speech with background music, and music. However in actual broadcast audio, in particular, the anchor speakers' segments, the length of these segments is different. The clean speech is generally of much longer durations compared to the speech with background music and music. In order to illustrate this, a portion from the anchor speakers' segment is taken, which is of a duration of approximately 5 minutes. This portion of the anchor speakers' segment contains 75 % of clean speech, 20 % of speech with background music and 5 % of music approximately. The speech with background music and the music are mainly present in between the anchor speakers' speech as news headlines. The phone error rate (PER) with or without preprocessing is shown in Table 6.5.

In the column of the results for the case where preprocessing is applied, the results are also shown separately for the clean speech and the enhanced speech. The clean speech results are the results obtained by passing the clean speech through the phone recognizer which is obtained by passing the 5 minutes of audio through the speech/music classification and clean speech/speech with background music classification module. The other segments obtained through the classification stages like the music and the speech with background music are neglected and only the PER of clean speech is noted. The enhanced speech results are the results obtained by passing only the enhanced speech

**Table 6.5:** *Phone Error Rate (PER) for anchor speakers' speech taken from Broadcast Audio (BA). Clean speech-75%, Speech with background music-20 %, Music-5 %. Pre: Preprocessing through Speech/Music classification, Clean speech/Speech with Background Music Classification and Speech Enhancement of speech with background music*

| BA | PER (%) | | | |
|---|---|---|---|---|
| | w/o Pre | with Pre | | |
| Model ↓ | | Clean Speech | Enhanced Speech | Overall |
| GMM | 49.43 | 38.05 | 64.28 (71.42) | 44.31 |
| SGMM | 47.15 | 35.82 | 61.90 (66.67) | 42.04 |
| DNN | 47.72 | 36.56 | 63.09 (69.04) | 42.89 |

with background music (SBM) segments through the phone recognizer. The enhanced SBM segments are obtained by passing the SBM segments through the enhancement module. The other segments obtained through the classification stages like the music and the clean speech are discarded and the PER of only the enhanced SBM is noted. The results in the enhanced speech column of Table 6.5, which are indicated in bracket are for the speech with background music segments before enhancement. The overall results shown in the Table 6.5 are for the case when both the clean speech and the enhanced speech with background music segments are passed through the phone recognizer. The music portion is discarded. It can be observed that the results improve after using the preprocessing stages. However the improvements have come mostly from the classification stages and the enhancement stage has not really contributed to the overall task. If the speech with background music regions are not considered, the PER drops drastically as seen in the Table 6.5, for the clean speech case which means that the classification stages which remove music and SBM from the 5 minutes of audio, have contributed to the improvement of the PER. If the enhanced SBM segments are considered the PER increases and this can be seen from the overall results in the last column of the table. This increase of the overall PER is attributed to the high PER of the enhanced SBM segments which means that the enhancement module is not very effective in this work, even though it reduces the PER compared to the SBM segments (shown in bracket of the enhanced speech column of the Table 6.5).

In most of the multimedia related task like audio summarization, the speech with background music (SBM) segments can be neglected. However, in these applications, if the SBM segments are considered,

then better approaches need to be developed to address the issues related to these segments. In this thesis, the preprocessing issues are addressed and effective methods for the classification stages have been developed which remove the music and speech with background music segments. This improves the PER of the clean speech segment obtained after the classification stages compared to the anchor speakers' segment containing the music and speech with background music as well. However, the enhancement stage is not that effective in this thesis and it may need to be further improved to get a better overall PER of the anchor speakers' segments. This may be done only if the SBM segments are to be considered for further processing in task like audio summary. Otherwise, the SBM segments can be neglected. One way of improvement could be having the separate acoustic models for the speech with background music. Another way could be having a better enhancement method to enhance the speech with background music segments. These methods for improving the PER of the SBM segments can be attempted as part of the future work.

## 6.5 Summary

This work combined the different modules of the preprocessing stage for effective phone recognition of the broadcast audio. The first module consists of the speech/music classification which uses speech-specific features in terms of the source, vocal tract system and syllabic rate of speech. The clean speech/speech with background music is the next module which uses speech-specific features having characteristics different from the speech-specific features used for speech/music classification. These features are defined in terms of the average and relative characteristics of the vocal tract system. Another module for processing the speech with background music is the speech enhancement module which employs the source feature for the temporal and perceptual based enhancement. The audio is given as input through these modules sequentially. The obtained output of the preprocessing stages consists of clean speech and enhanced speech. These are passed through the phone recognizer to obtain the phonetic transcription. The accuracy of the overall system is compared with the accuracy obtained by directly passing the audio files through the phone recognizer. It is observed that performing the preprocessing steps before the final phone recognition improves PER of the phone transcription of the anchor speakers' segments of broadcast audio. This shows the significance of using the preprocessing steps proposed in this work. It was observed from the results that the improvement comes mostly from the classification stages, while the enhancement module only slightly contributes to improving

the overall PER.

# 7

# Summary and Conclusions

Contents

*In this chapter, contributions of this thesis towards the classification and phone recognition of broadcast audio using features, defined based on the speech-specific knowledge are summarized. Future research directions made possible by the present work are also outlined.*

## 7.1 Summary

(i) **Speech/Music Classification using Speech-Specific Features:** The speech/music classification has been explored by using the speech-specific features in terms of the source, vocal tract system and syllabic rate of speech production. The normalized autocorrelation peak strength (NAPS) of the zero frequency filtered signal (ZFFS) and the peak-to-sidelobe ratio (PSR) of the hilbert envelope (HE) of linear prediction (LP) residual represent the source. The log mel spectrum energy represents the vocal tract system. The modulation spectrum energy feature represents the syllabic rate of speech. These features were concatenated and passed through the support vector machine (SVM) and gaussian mixture model (GMM) classifier. The speech-specific features were compared with the existing temporal and spectral based features and found to be better, where the testing has been done on the scheirer and slaney (S&S), GTZAN and the broadcast news database. On combining the speech-specific features along with the existing features the best performances are obtained on all the three databases.

(ii) **Clean Speech/Speech with Background Music Classification using HNGD Spectrum:** The clean speech/speech with background music classification system was developed based on other kinds of speech-specific features mostly in terms of the vocal tract characteristics. The average and relative spectral characteristics of the vocal tract system were explored. The mel frequency cepstral coefficients (MFCCs) represent the average spectral characteristics while the sum of the spectral contrast on the hilbert envelope of numerator of group delay (HNGD) spectrum represents the relative spectral characteristics. The HNGD spectrum is a better representation of the vocal tract system and has been shown to perform better that the conventional discrete fourier transform (DFT) spectrum in this work. These features are concatenated to form a feature vector and passed through the SVM and GMM classifier. The features gave the best performance on combining the average and relative spectral characteristics which show their complementary nature for the task. The algorithm was tested on the S&S database and the broadcast news database.

(iii) **Significance of Source Enhancement for Phone Recognition of Speech with Background Music Segments:** The enhancement of speech with background music by using the source information has been developed. The source information is exploited in terms of the epoch locations and strengths. This source information has been incorporated in the temporal and perceptual enhancement modules of the sequential temporal, spectral and perceptual enhancement technique. The epoch locations and strengths are obtained from the single frequency filter-zero frequency filter (SFF-ZFF) combination. These epoch locations are used for generating the fine weight function in temporal enhancement along with the use of speech-specific features for the gross weight function. The epoch locations and strengths are also used for the generation of impulse train which acts as the excitation source for the perceptual based enhancement. The smooth spectral envelope computed using the mel log spectral approximation (MLSA) filter from the mel cepstral coefficients (MCCs) are used as the vocal tract system information for the perceptual based enhancement. The overall enhanced files are tested on the speech with background music segments which have been generated using TIMIT database as well as the broadcast audio. The samples generated using TIMIT are synthesized by adding music obtained from S&S database to the clean speech samples at a speech to music ratio (SMR) of 0 dB. The results show that the enhancement technique contributes to the performance of the overall phone recognition accuracy particularly when the music is heavy in nature as in rock and heavy metal music. The results on the broadcast audio also show good improvements although the performance is slightly lesser than the TIMIT case since the broadcast audio may not have the background music at 0 dB.

(iv) **Phone Recognition of Speech in the Context of Broadcast Audio:** The preprocessing systems are integrated for performing the phone recognition of broadcast audio. The speech/music classification system, the clean speech/speech with background music classification system and the speech enhancement system are combined as the preprocessing stages. The output of the preprocessing stages is passed through the phone recognizer. The overall accuracy of the phone recognizer is improved when the preprocessing stages are used compared to the case of passing the audio containing clean speech, speech with background music and music, directly through the phone recognizer. The audio samples considered are generated using TIMIT database and also taken from the audio data recorded from the Indian broadcast news.

## 7.2 Contributions

The major contributions of the research work reported in this thesis includes,

(i) Speech / Music classification using speech-specific features.

(ii) Clean Speech / Speech with background music classification using average and relative spectral characteristics of the vocal tract system.

(iii) Enhancement of Speech with background music using temporal, spectral and perceptual processing exploiting the effect of the source.

(iv) Phone recognition of clean as well as enhanced speech using a GMM-HMM, SGMM-HMM and DNN-HMM phone recognition system

## 7.3 Directions for future work

Based on the outcome of this thesis work, this section provides some of the possible future directions for research.

(i) The speech/music classification has been defined based on the speech-specific features along with some set of classifiers. Recently deep architectures have been explored for various speech related task. The deep neural networks have been used as powerful classifiers in other tasks. DNN classifiers can also be attempted for the speech/music classification. Another way the DNN can be used is in terms of the feature extractors. The DNN can learn certain aspects of speech as well as music thereby unfolding hidden information in these two classes thus providing an effective classification between speech and music.

(ii) The clean speech/speech with background music classification system has been defined on the basis of the vocal tract system of speech production where the average and relative spectral characteristics are explored. The future work may also include the exploration of source features for this task. It is shown in this thesis that the vocal tract system feature deviates due to the presence of music in the speech with background music segments. The same may happen for the source related feature as well and this may give additional improvement for the task. On similar terms, the suprasegmental aspects of speech production can also be explored for the clean speech/speech with background music classification task.

(iii) The Enhancement of Speech with background music using temporal, spectral and perceptual processing exploiting the effect of the source has been performed in this thesis. The spectral processing can be explored further in future work by exploiting the techniques such as formant enhancement in an efficient manner so that the peaks related to music can be suppressed. For the case of noise, this kind of enhancement has been attempted. The perceptual enhancement can also be looked at in terms of the vocal tract system. In this thesis, the Mel Cepstral Coefficients are used. Other forms of representing the vocal tract system for the perceptual based enhancement can be explored.

(iv) The classification tasks in this thesis have most of their features defined in terms speech and are mostly based on 1 dimensional (1-D) processing. The results showed that combining the speech-specific features with the other existing time and frequency domain features gave the best performance. This shows that the complementary nature of the features contributes to the task. The future work on the classification task may include the use of features based on 2-D signal processing. These features extracted using 2-D signal processing may capture some information for the task which is complementary to the features extracted using 1-D signal processing defined earlier, and may introduce additional improvement.

(v) The classification stages were shown to be effectively improving the overall PER of the anchor speakers' segment in this work, by removing the music and speech with background music segments. The enhancement stage of the speech with background music segments, slightly contributes to improving the overall PER, if the speech with background music segments are considered. The speech with background music segments are to be analyzed properly, if they are to be used for further processing in tasks like audio summarization. The acoustic models for the speech with background music segments can be built separately and these models can be specifically selected when the phone transcription of speech with background music segments are required. This may improve the PER of the speech with background music segments and hence improve the overall PER of the anchor speakers' segment. Another method can be the further improvement of the enhancement stage, by incorporating other aspects of speech production, so that the phone transcription of the enhanced speech with background music can be improved than the one developed in this thesis. These can be attempted as part of the future work.

# Bibliography


[1] S. H. Yella, V. Varma, and K. Prahallad, "Significance of anchor speaker segments for constructing extractive audio summaries of broadcast news," in *Spoken Language Technology Workshop (SLT), 2010 IEEE*. IEEE, 2010, pp. 13–18.

[2] K. Ng and V. Zue, "Phonetic recognition for spoken document retrieval," in *Proceedings of the ICASSP,Seattle, WA, USA*, 1998.

[3] W. Kraaij, J. Van Gent, R. Ekkelenkamp, and D. Van Leeuwen, "Phoneme based spoken document retrieval," in *Proceedings of TREC*, vol. 7, 1998.

[4] C. Hori and S. Furui, "A new approach to automatic speech summarization," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 368–378, 2003.

[5] J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen, R. Schwartz, and A. Srivastava, "Speech and language technologies for audio indexing and retrieval," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338–1353, 2000.

[6] L. Nguyen, S. Matsoukas, J. Davenport, F. Kubala, R. Schwartz, and J. Makhoul, "Progress in transcription of broadcast news using byblos," *Speech Communication*, vol. 38, no. 1-2, p. 213230, September 2002.

[7] S. Wegmann, P. Zhan, and L.Gillick, "Progress in broadcast news transcription at dragon systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1999, pp. 33–36.

[8] J. Gauvain, L. Lamel, and G. Adda, "Transcribing broadcast news for audio and video indexing," *Communications Of the ACM*, vol. 43, no. 2, pp. 64–70, February 2000.

[9] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[10] P. Woodland, "The development of the htk broadcast news transcription system: An overview," *Speech Communication*, vol. 37, no. 1-2, pp. 47–67, May 2002.

[11] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1996, pp. 993 – 996.

[12] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1997, pp. 1331– 1334.

[13] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on Multimedia,*, vol. 7, no. 1, pp. 155–166, 2005.

[14] Y.Lavner and D.Ruinskiy, "A decision-tree-based algorithm for speech/music classification and segmentation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2009, no. 2, January 2009.

[15] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[16] B. Yegnanarayana, C. Avendano, H. Hermansky, and P. S. Murthy, "Speech enhancement using linear prediction residual," *Speech Communication*, vol. 28, no. 1, pp. 25–42, 1999.

[17] P. Krishnamoorthy and S. R. M. Prasanna, "Enhancement of noisy speech by temporal and spectral processing," *Speech Communication*, vol. 53, no. 2, pp. 154–174, 2011.

[18] K. T. Deepak and S. R. M. Prasanna, "Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1204–1218, 2016.

[19] K. S. R. Murthy and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, pp. 1602–1613, Nov 2008.

[20] A. G.Hauptmann and M. J. Witbrock, "Informedia news on demand: Information acquisition and retrieval," *In M. T. Maybury (Ed.), Intelligent Multimedia Information Retrieval*, p. 213239, 1997.

[21] D.Abberley, S. Renals, and G. Cook, "Retrieval of broadcast news documents with the thisl system," in *In Proceedings of the IEEE International Con- ference on Acoustics, Speech, and Signal Processing*, 1998, p. 37813784.

[22] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, no. 1, pp. 5–20, 2000.

[23] H.-m. Wang, "Experiments in syllable-based retrieval of broadcast news speech in mandarin chinese," *Speech Communication*, vol. 32, no. 1, pp. 49–60, 2000.

[24] G. Salton and M. J. McGill, "Introduction to modern information retrieval," *McGraw-Hill*, 1986.

[25] F. Kubala, J. Davenport, H. Jin, D. Liu, T. Leek, S. Matsoukas, D. Miller, L. Nguyen, F. Richardson, R. Schwartz *et al.*, "The 1997 bbn byblos system applied to broadcast news transcription," in *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*. Citeseer, 1998.

[26] T. M.Cover and J. A.Thomas, *Elements of Information Theory*. New York:John Wiley and Sons, 1991.

[27] M. Kos, Z. Kačič, and D. Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features," *Digital Signal Processing*, vol. 23, no. 2, pp. 659–674, 2013.

[28] J. Wolfe, "Speech and music, acoustics and coding, and what music might be for," in *Proc. 7th International Conference on Music Perception and Cognition*, 2002, pp. 10–13.

[29] G. Williams and D. P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH 99)*, Sep. 1999, pp. 687–690.

[30] G. Sell and P. Clark, "Music tonality features for speech/music discrimination," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 2489–2493.

[31] T. Zhang and C. C. J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, 2001.

[32] P. C. Loizou, *Speech enhancement: theory and practice*. CRC press, 2013.

[33] L. R. Rabiner and R. W. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, no. 1, pp. 1–194, 2007.

[34] D. Wang and J. Lim, "The unimportance of phase in speech enhancement," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, no. 4, pp. 679–681, 1982.

[35] S. V. Vaseghi, *Advanced digital signal processing and noise reduction*. John Wiley & Sons, 2008.

[36] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'79.*, vol. 4. IEEE, 1979, pp. 208–211.

[37] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[38] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE international conference on acoustics speech and signal processing*, vol. 4. Citeseer, 2002, pp. 4164–4164.

[39] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, 1980.

[40] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE transactions on speech and audio processing*, vol. 13, no. 5, pp. 845–856, 2005.

[41] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, 1992.

[42] M.-C. You, C.-Y. Mao, J.-S. Wang, and F.-C. Chuang, "A recursive parametric spectral subtraction algorithm for speech enhancement," in *International Conference on Intelligent Computing*. Springer, 2007, pp. 826–835.

[43] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on speech and audio processing*, vol. 7, no. 2, pp. 126–137, 1999.

[44] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[45] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 40, no. 4, pp. 725–735, 1992.

[46] Y. Soon, S. N. Koh, and C. K. Yeo, "Noisy speech enhancement using discrete cosine transform," *Speech communication*, vol. 24, no. 3, pp. 249–257, 1998.

[47] B. Chen and P. C. Loizou, "Speech enhancement using a mmse short time spectral amplitude estimator with laplacian speech modeling." in *ICASSP (1)*, 2005, pp. 1097–1100.

[48] M. Marzinzik and B. Kollmeier, "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 109–118, 2002.

[49] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 92–102, 2012.

[50] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[51] A. Hussain, M. Chetouani, S. Squartini, A. Bastari, and F. Piazza, "Progress in nonlinear speech processing," *Springer Berlin/Heidelberg*, 2007.

[52] H. Taşmaz and E. Erçelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and mmse–stsa estimation in various noise environments," *Digital Signal Processing*, vol. 18, no. 5, pp. 797–812, 2008.

[53] D. L. Donoho, "De-noising by soft-thresholding," *IEEE transactions on information theory*, vol. 41, no. 3, pp. 613–627, 1995.

[54] M. K. Hasan, S. Salahuddin, and M. R. Khan, "Reducing signal-bias from mad estimated noise level for dct speech enhancement," *Signal Processing*, vol. 84, no. 1, pp. 151–162, 2004.

[55] J.-H. Chang, S. Gazor, N. S. Kim, and S. K. Mitra, "Multiple statistical models for soft decision in noisy speech enhancement," *Pattern Recognition*, vol. 40, no. 3, pp. 1123–1134, 2007.

[56] K. Hermus, P. Wambacq *et al.*, "A review of signal subspace speech enhancement and its application to noise robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 1, pp. 1–15, 2006.

[57] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from noise: A regenerative approach," *Speech Communication*, vol. 10, no. 1, pp. 45–57, 1991.

[58] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on speech and audio processing*, vol. 3, no. 4, pp. 251–266, 1995.

[59] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated qsvd," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 439–448, 1995.

[60] F. S. Cooper, "Acoustics in human communication: Evolving ideas about the nature of speech," *The Journal of the Acoustical Society of America*, vol. 68, no. 1, pp. 18–21, 1980.

[61] W. Jin and M. S. Scordilis, "Speech enhancement by residual domain constrained optimization," *Speech communication*, vol. 48, no. 10, pp. 1349–1364, 2006.

[62] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–541.

[63] P. Fousek, L. Lamel, and J.-L. Gauvain, "Transcribing broadcast data using mlp features." in *InterSpeech*, vol. 8, 2008, pp. 1433–1436.

[64] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[65] J. Shirazi and S. Ghaemmaghami, "Improvement to speech-music discrimination using sinusoidal model based features," *Multimedia Tools and Applications*, vol. 50, no. 2, pp. 415–435, 2010.

[66] K. Deepak, B. D. Sarma, and S. R. M. Prasanna, "Foreground speech segmentation using zero frequency filtered signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[67] N. Adiga and S. R. M. Prasanna, "Detection of glottal activity using different attributes of source information," *IEEE Signal Processing Letters*, vol. 22, no. 11, pp. 2107–2111, Nov 2015.

[68] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Transactions on Speech and Audio Processing,*, vol. 13, no. 5, pp. 751–761, 2005.

[69] C.-H. Lee, J.-L. Shih, K.-M. Yu, and H.-S. Lin, "Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features," *IEEE Transactions on Multimedia*, vol. 11, pp. 670–682, June 2009.

[70] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[71] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 253–266, 2009.

[72] S. R. M. Prasanna, B. S. Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 17, no. 4, pp. 556–565, 2009.

[73] J. Makhoul, "Linear prediction: A tutorial review," in *Proc. IEEE*, vol. 63, no. 04, April 1975, pp. 561–580.

[74] T. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 4, pp. 309–319, August 1979.

[75] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* N. Delhi, India: Prentice-Hall India, 1975.

[76] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 1997, pp. 1647–1650.

[77] C. L. Smith, C. P. Browman, R. S. McGowan, and B. Kay, "Extracting dynamic parameters from speech movement data," *J. Acoust. Soc. Amer.*, vol. 93, no. 3, pp. 1580–1588, 1993.

[78] H. Dudley, "Remaking speech," *J. Acoust. Soc. Amer.*, vol. 11, no. 2, p. 169177, October 1939.

[79] R. Drullman, J. M. Festen, and R. Plomp, "Effect of temporally envelope smearing on speech reception," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1053–1064, 1994.

[80] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, pp. 117–132, 1998.

[81] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters,*, vol. 17, no. 3, pp. 273–276, 2010.

[82] K. Sang-Kyun and J.-H. Chang, "Speech/music classification enhancement for 3gpp2 smv codec based on support vector machine," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 92, no. 2, pp. 630–632, 2009.

[83] C. Lim and J.-H. Chang, "Efficient implementation techniques of an svm-based speech/music classifier in smv," *Multimedia Tools and Applications*, pp. 1–26, 2014.

[84] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[85] G. Tzanetakis and P. Cook, "Sound analysis using mpeg compressed audio," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, 2000. ICASSP'00.*, vol. 2, 2000, pp. II761–II764.

[86] J.-L. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech communication*, vol. 37, no. 1, pp. 89–108, 2002.

[87] P. Beyerlein, X. Aubert, R. Haeb-Umbach, M. Harris, D. Klakow, A. Wendemuth, S. Molau, H. Ney, M. Pitz, and A. Sixtus, "Large vocabulary continuous speech recognition of broadcast news–the philips/rwth approach," *Speech Communication*, vol. 37, no. 1, pp. 109–131, 2002.

[88] D. Castán, A. Ortega, A. Miguel, and E. Lleida, "Audio segmentation-by-classification approach based on factor analysis in broadcast news domain," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–13, 2014.

[89] J. Vavrek, E. Vozáriková, M. Pleva, and J. Juhár, "Broadcast news audio classification using svm binary trees," in *Telecommunications and Signal Processing (TSP), 2012 35th International Conference on.* IEEE, 2012, pp. 469–473.

[90] D.-N. Jiang, L. Lu, H.-J. Zhang, J.-H. Tao, and L.-H. Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. 113–116.

[91] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782–795, 2013.

[92] B. Yegnanarayana, "Formant extraction from linear-prediction phase spectra," *The Journal of the Acoustical Society of America*, vol. 63, no. 5, pp. 1638–1640, 1978.

[93] B. Yegnanarayana and H. A. Murthy, "Significance of group delay functions in spectrum estimation," *Signal Processing, IEEE Transactions on*, vol. 40, no. 9, pp. 2281–2289, 1992.

[94] M. Anand Joseph, S. Guruprasad, and B. Yegnanarayana, "Extracting formants from short segments of speech using group delay functions," 2006.

[95] K. S. Srinivas and K. Prahallad, "An fir implementation of zero frequency filtering of speech signals," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2613–2617, 2012.

[96] B. K. Khonglah and S. R. M. Prasanna, "Speech/music classification using speech-specific features," *Digital Signal Processing*, vol. 48, pp. 71–83, 2016.

[97] G. Aneeja and B. Yegnanarayana, "Single frequency filtering approach for discriminating speech and nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, 2015.

[98] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52–63, 2017.

[99] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[100] G. Fant, "Speech sounds and features." 1973.

[101] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation." in *ICSLP*, 1994.

[102] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.

[103] S. Shahnawazuddin, D. Thotappa, A. Dey, S. Imani, S. R. M. Prasanna, and R. Sinha, "Improvements in iitg assamese spoken query system: Background noise suppression and alternate acoustic modeling," *Journal of Signal Processing Systems*, pp. 1–12, 2016.

# List of Publications

- **Publications composing the thesis**

  **J** Banriskhem K. Khonglah and S. R. M. Prasanna, "Speech / Music Classification using Speech Specific Features," Digital Signal Processing, Elsevier, Jan. 2016.

  **J** Banriskhem K. Khonglah and S. R. M. Prasanna, "Clean Speech/Speech with Background Music Classification using HNGD spectrum," **accepted in**, International Journal of Speech Technology, Springer, August 2017.

  **J** Banriskhem K. Khonglah, Abhishek Dey and S. R. M. Prasanna, "Speech Enhancement using Source Information for Phoneme Recognition in Speech with Background Music," **under review in**, Circuits, Systems and Signal Processing, Springer, August 2017.

  **C** B. K. Khonglah, R. Sharma, and S.R.M. Prasanna, "Speech vs music discrimination using Empirical Mode Decomposition," in NCC-2015, Mumbai, India.

  **C** Banriskhem K. Khonglah and S. R. M. Prasanna, "Speech / Music Classification using Vocal Tract Constriction Aspect of Speech," in INDICON-2015, IEEE, New Delhi, India.

  **C** Banriskhem K. Khonglah and S. R. M. Prasanna, "Low Frequency Region of Vocal Tract Information for Speech / Music Classification," in TENCON-2016, IEEE, Signapore.

- **Publications other than the thesis**

  **J** Nagaraj Adiga, Banriskhem K. Khonglah and S. R. M. Prasanna, "Glottal activity detection for HMM based speech synthesis," Digital Signal Processing, Elsevier, December 2017.

  **J** Banriskhem K. Khonglah, Ramesh K. Bhukya and S. R. M. Prasanna, "Processing Degraded Speech for Text Dependent Speaker Verification,", International Journal of Speech Technology, Springer, August 2017.

  **C** Banriskhem K. Khonglah, Biswajit Dev Sarma and S. R. M. Prasanna, "Exploration of Deep Belief Networks for Vowel-Like Regions Detection," in INDICON-2014, IEEE, Pune, India.

  **C** Arghya Pal, B. K. Khonglah, S. Mandal, Himakshi Choudhury, S. R. M. Prasanna, H.L. Rufiner, Vineeth N. Balasubramanian, "Online Bengali handwritten numerals recognition using Deep Autoencoders," in NCC-2016, Guwahati, India.

  **C** Shikha Baghel, Banriskhem K. Khonglah, S. R. M. Prasanna and Prithwijit Guha, "Shouted/Normal Speech Classification using Speech-Specific Features," in TENCON-2016, IEEE, Signapore.

***J indicates Journal and C indicates Conference**