FEATURES FOR
AUDIO CLIP CLASSIFICATION

A THESIS

submitted by

ANVITA BAJPAI

for the award of the degree

of

MASTER OF SCIENCE

(by Research)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY, MADRAS.

MAY 2004

THESIS CERTIFICATE

This is to certify that the thesis entitled **Features for Audio Clip Classification** submitted by **Anvita Bajpai** to the Indian Institute of Technology Madras for the award of the degree of Master of Science (by Research) is a bonafide record of research work carried out by her under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chennai 600 036 — Prof. B. Yegnanarayana
Date: — Dept. of Computer Science and Engg.

ACKNOWLEDGEMENTS

I express my sincere gratitude to Prof. B. Yegnanarayana for providing an excellent research environment and stimulating research goals. Prof. is not only a nice teacher and researcher, but also a nice person. I thank him for giving me moral support which helped me to carry out my work.

I thank Dr. C. Sekhar, my GTC members, head of the department (Prof. Raman) and other faculty members of the department for their valuable advice which directly or indirectly helped me in refining my research work. Today when I am writing my acknowledgments, credit goes to my first teacher also, who taught me how to write *'A'*. I thank all my teachers for making me what I am today. In particular, my thanks goes to Prof. T. V. Prabhakar, Dr. T. Archana, Prof. Sumit Ganguly, (*all at IITK*), Dr. Gursaran, Dr. Satish Kumar, Dr. Gunjan Aggarwal, (*all at DEI, Agra*), for making me understand different aspects of science, technology and life. Feel happy for having teacher-friend like Kendra Crossen Burroughs, who makes me realize a few values of life.

I believe that in order to do s'thing great, one has to be good technically as well as mentally, and this is where friends play an important role. One needs to have a few people in life, whom he can count on at any point of time. I'm lucky for having a few very nice friends in my life. My special thanks goes to Vinod for keeping my spirit up during my stay in IITM. I can't thank Rajesh for the smiles he'd given to me, Kapil for chocolates (*still his chocolates are waiting for the 'ant' to come and eat them*), Aparna for loving heart, Jyotirmoy for small talks/tips. Gaurav's cute-small cares, inspite of mine getting angry with him at times, still bring smiles on my face. Acknowledgments to Amit for his philosophy (*I must admit that I enjoy reading Amit, GauravV's mails.. they have a style*) and NEVER forgetting my b'day (*and best part is that I always forget his :*), Amboo for taking care of my ODD habits (*that less-sugar tea .. I can only smile now*), Thanks to Dwarika, Sidd, Manu, Avinash, Vikas, Prashant, Sanjeev, Pari (*Parishikta, like her name*), Sheetal, Ekta, Mirnal, Shamita (*her ladoo ..*), Shweta (*sweet girl*), Swati, Sarada, Simmi (S^3*, the tikari in Sarav, IITM .. here generally* 2 *or more of us used to attack others for teasing, and then there was no end to the stories we used to visualize. Can't forget those blushing faces I enjoyed during those bullas .. hey don't think I'm THE bad, at times I also used to be the bakara (bakari rather)*).

Thanks to Rupali for her concern "if I had food?" and Maggie I used to make in her room after my wash. Can't forget Satyadev's helps/tips, outing with Meenu, sudden movie plans with Shamita and Pari, walks with Himani, Vandana's home ....

Guess it's difficult to write all names/events here, but my heart remembers all of them and thanks them for making me feel special in some way or other, for giving me strength which was needed for devoting myself to my work. I thank my lab members for helping me at times. Prasanna deserves thanks for his guardianly attitude for me and supporting me most of the times by saying *"ek hi ladaki hai hamari lab main, and one should care for her"*. Surya's *bulla* and never inviting me for tea with a teasing comment, brings smile on my face. Nice time that I had with seniors, also helped me in settling down in IITM. I thank the ugliness around too, for having made me realize the beauty of *the beauty around*.

Life rotates around home, family. Thanks to Dr. Prabhakar and Archana ma'am for providing me home away from home, and Mr. Murty and aunty, Vinod for welcoming me to their home during my stay in Madras. Going back to place where I was born, feel lucky for having a grandmother to teach me the importance of being optimistic. Gratitudes to my relatives/cousins also for giving me smiles, making me feel *laadli beti*, for all nice times I had since my childhood. Thanks to my youngsters for enjoying whatever little I did for them and making me feel precious by depending on me for small-small things, doing small things for me, having small fights together, sharing small fears together, for every small experience and feeling of responsibility (coz of mine being elder). .... Words can't thank my parents for making me a batter person, for teaching me the importance of being honest.

I thank to God for making me *"myself"* and keep on coming to me in the form of nice people around.

- Anvita Bajpai

ABSTRACT

Keywords: *Audio indexing; Audio classification; Multimedia indexing; Source characteristics; Subsegmental features; Segmental features; Suprasegmental features; Linear prediction residual; Neural network models; Autoassociative neural network; Multilayer perceptron; Feedforward neural network; Hilbert envelope.*

The kind of data that is used today for various applications is mostly in the form of audio, video and multimedia. This data is large in volume, and increasing at an accelerated pace. Therefore, manual handling of it is impractical for real-time applications. Audio plays an important role in handling the multimedia data as it is easier to process when compared to the video data, and also the audio data contains perceptually significant information. Audio indexing is the task of dealing with analysis, storage and retrieval of audio data. Classification of audio data into different categories is one important step in building an audio indexing system. The information about the audio category is contained in the excitation source (subsegmental), system/physiological (segmental) and behavioral (suprasegmental) characteristics of the audio data. In this work, the presence of audio-specific information in the excitation source part of the audio signal is demonstrated. The linear prediction (LP) residual is used to represent the excitation source. The residual of a signal may be less affected by channel degradations as compared to the spectral information. This work emphasizes the importance of information present in the LP residual of audio signals, which if combined with the spectral information, can give a better performing audio indexing system. Since it may be difficult to extract information from the residual using known signal processing algorithms, neural network (NN) models are proposed to capture the information at the subsegmental level in the LP residual. In this thesis, the autoassociative neural network (AANN) models are used to capture the audio-specific information from the LP residual of signals. The multilayer perceptron (MLP) models are used to classify the audio data using the audio-specific information captured by AANN models. Since source (subsegmental) information is complementary in nature to the system (segmental) information, methods to combine the two levels of information present in the audio signal are discussed. For humans, the information about the audio category is perceived by listening to a longer segment of audio signal. These behavioral characteristics of the audio data are perceived in the LP residual of audio signal also. The Hilbert envelope of the residual signal is used to demonstrate the presence of audio-specific suprasegmental information in the LP residual.

TABLE OF CONTENTS

Thesis Certificate **i**

Acknowledgements **ii**

Abstract **iv**

List of Tables **ix**

List of Figures **x**

Abbreviations **0**

1 INTRODUCTION **1**
1.1 Audio Indexing 1
1.1.1 Audio Indexing Problem Definition and its Importance 1
1.1.2 Audio Indexing System 3
1.1.2.1 How it is Different from a Text-based Indexing System? 3
1.1.3 Categories of Audio Indexing 5
1.1.4 Examples of Audio Indexing Tasks 6
1.2 Audio Clip Classification - Proposed Problem 7
1.3 Issues in Audio Clip Classification 8
1.3.1 Audio-specific Feature Extraction 10
1.3.2 Generating Models for Audio Clip Classification 10
1.3.3 Developing Matching Algorithm for Audio Clip Classification . . 11
1.3.4 Developing Decision Logic for Audio Clip Classification 12
1.4 Objectives of the Present Study 12
1.5 Organization of the Thesis 16

2 REVIEW OF APPROACHES TO AUDIO INDEXING **17**
2.1 Types of Audio Indexing Tasks 17
2.1.1 Spoken Document Retrieval 17
2.1.2 Other Audio Indexing Tasks 18
2.2 Feature Extraction for Audio Indexing 20
2.2.1 Characteristics of a Feature 20
2.2.2 Features for Speaker Recognition 21

2.2.3 Features for Language Identification 21
2.2.4 Features for Music Classification 21
2.2.5 Features for General Case of Audio Indexing 22
2.2.6 Features for Audio Clip Classification 22
2.3 Techniques for Audio Indexing 23
2.3.1 Model-based Techniques 24
2.3.2 Metric-based Techniques 24
2.3.3 Rule-based Techniques 26
2.4 Discussion 26
2.4.1 Missing Components in Existing Approaches and their Importance 26
2.4.2 Proposed Methods to Overcome the Limitations of Existing Approaches 27
2.5 Summary 28

3 BASELINE SYSTEM USING SEGMENTAL FEATURES 29
3.1 Spectral (Segmental) Features for Audio Clip Classification 30
3.1.1 Volume-based Features 31
3.1.1.1 Volume Standard Deviation (VSTD) 31
3.1.1.2 Volume Dynamic Range (VDR) 33
3.1.1.3 Volume Undulation (VU) 33
3.1.1.4 4Hz Modulation Energy (4ME) 34
3.1.2 Zero Crossing Rate (ZCR)-based Features 35
3.1.2.1 ZCR Standard Deviation (ZSTD) 35
3.1.2.2 Nonsilence Ratio (NSR) 36
3.1.3 Pitch-based Features 37
3.1.3.1 Pitch Standard Deviation (PSTD) 40
3.1.3.2 Similar Pitch Ratio (SPR) 40
3.1.3.3 Nonpitch Ratio (NPR) 41
3.1.4 Spectrum-based Features 41
3.1.4.1 Frequency Centroid (FC) 42
3.1.4.2 Bandwidth (BW) 42
3.1.4.3 Energy Ratio in Subbands (ERSBs) 43
3.2 Hidden Markov Models for Audio Clip Classification 44
3.3 Database Used for Study 45
3.4 Audio Clip Classification Results using Baseline System 46
3.5 Summary 47

4 EXTRACTION OF AUDIO-SPECIFIC SUBSEGMENTAL INFORMATION PRESENT IN THE LP RESIDUAL USING AANN MODELS 48

4.1 Sound Production Mechanism 49
4.2 Source Characteristics for Audio Clip classification 50
4.3 Significance of the LP Residual for Audio Clip Classification 51
4.3.1 The Linear Prediction Residual 51
4.3.2 Presence of Audio-Specific Source Characteristics in the LP Residual 52
4.4 The AANN Models for Capturing Audio-Specific Information in the LP Residual 54
4.5 Knowledge of Audio Components for Defining Audio Category 57
4.6 Audio Clip Classification Studies 59
4.6.1 Building AANN Models for Audio Components 59
4.6.2 AANN Models for Capturing Subsegmental Features 61
4.6.3 Work-flow of the System 64
4.7 Summary 65

5 AUDIO CLIP CLASSIFICATION USING SUBSEGMENTAL AND SEGMENTAL FEATURES 66
5.1 The MLP for Audio Clip Classification using Subsegmental Features . . 67
5.1.1 Structure of the MLP 67
5.1.2 Training the MLP 68
5.1.3 Testing the MLP 69
5.2 Classification Results using Subsegmental Features 70
5.3 Combining Evidences from the Two Systems 72
5.3.1 Multiple Classifiers and Combining Evidence 73
5.3.2 Abstract Level Combination of Two Systems 75
5.3.3 Rank and Measurement Level Combination of the Two Systems 76
5.4 Summary 78

6 SUPRASEGMENTAL FEATURES FOR AUDIO CLIP CLASSIFICATION 80
6.1 Suprasegmental Features in the LP Residual Signal 81
6.2 The Hilbert Envelope of the LP Residual Signal 82
6.2.1 Computation of the Hilbert Envelope from Residual Signal . . . 82
6.2.2 Presence of Suprasegmental Features in the Hilbert Envelope . . 83
6.3 Suprasegmental Features in the Hilbert Envelope 84
6.4 Summary 88

7 SUMMARY AND CONCLUSIONS 89
7.1 Major Contributions of the Present Work 90
7.2 Scope for Future Work 91

Appendix A **92**

Appendix B **95**

Appendix C **98**

Appendix D **102**

Bibliography **105**

List of Publications **113**

LIST OF TABLES

3.1 Audio clip classification performance using system based on the spectral features. 47

5.1 Audio clip classification performance using system based on the LP residual. 70
5.2 Audio clip classification performance for combined (at abstract level) system for DB2. 76
5.3 Audio clip classification performance for combined (at rank and measurement level) system for DB2. 77
5.4 An example of class-wise normalized output confidence values of the two systems and its sum. 78

LIST OF FIGURES

1.1 Block diagram for pattern classification task. 9
1.2 Block diagram for training phase. 11
1.3 Block diagram for testing phase. 11
1.4 Block diagram for combining evidences. 13

2.1 Schematic diagram showing the task of SDR. 18

3.1 (a) Benchmark cricket clip, (b) Volume standard deviation, (c) Volume dynamic range, (d) Volume undulation, (e) 4Hz modulation energy. . . 32
3.2 (a) Benchmark cricket clip, (b) Zero crossing rate standard deviation, (c) Nonsilence ratio, (d) Pitch standard deviation, (e) Similar pitch ratio, (f) Nonpitch ratio. 37
3.3 (a) The speech waveform, (b) Residual, (c) Hilbert envelope. 39
3.4 (a) Benchmark cricket clip, (b) Frequency centroid, (c) Bandwidth, (d) Energy ratio in subband 1, (e) Energy ratio in subband 2, (f) Energy ratio in subband 3. 43

4.1 The LP residual for segments of audio clips belonging to five different categories. 52
4.2 (a) Spectrum of news audio signal. (b) Spectrum of residual signal. . . 54
4.3 Structure of the AANN model used for capturing audio-specific information. 57
4.4 Training error curves for six AANN component models used for the study. 60
4.5 Average confidence score values (over a window of 10 samples, with every sample shift) with respect to six components for a segment of news test clip. 62
4.6 Average confidence score values (over a window of 10 samples, with every sample shift) with respect to six components for the segments of test clips of the five audio categories. 63
4.7 Work-flow of the audio clip classification system using subsegmental information present in the audio signal. 64

5.1 Structure of the MLP used for audio clip classification task. 68
5.2 Venn diagram showing the complementary nature of the two systems. . 73

6.1 The LP residual for the segments of audio clips belonging to five audio categories. 81
6.2 The waveform, LP residual, Hilbert envelope, modified Hilbert envelope of the residual signal for a noisy audio segment. 84
6.3 The Hilbert envelope of the residual signal for the segments of audio clips belonging to five audio categories. 85
6.4 Autocorrelation sequences of the Hilbert envelope of the residual signal of five audio categories. 86
6.5 Statistics of autocorrelation sequences of the Hilbert envelope of five audio categories. 88

C.1 Basic feedforward neural network. 99

ABBREVIATIONS

LP	- Linear Prediction
LPC	- Linear Prediction Coefficients
LPCC	- Linear Prediction Cepstral Coefficients
MFCC	- Mel-frequency Cepstral Coefficients
NN	- Neural Network
ANN	- Artificial Neural Network
AANN	- Autoassociative Neural Network
MLP	- Multilayer Perceptron
SOM	- Self Organizing Maps
MLFFNN	- Multilayer Feedforward Neural Network
FFNN	- Feedforward Neural Network
CCV	- Component Confidence Vector
NFL	- Nearest Feature Line Method
SDR	- Spoken Document Retrieval
TV	- Television
VQ	- Vector Quantization
HMM	- Hidden Markov Model
GMM	- Gaussian Mixture Model
CHMM	- Continuous Hidden Markov Model
SVM	- Support Vector Machine
PDF	- Probability Density Function
ML	- Maximum Likelihood
k-NN	- k-Nearest Neighbor
RMS	- Root Mean Square
MSE	- Mean Squared Error

CHAPTER 1

INTRODUCTION

1.1 AUDIO INDEXING

1.1.1 Audio Indexing Problem Definition and its Importance

In this era of information technology, the data that we use is mostly in the form of audio, video and multimedia. The volume of this nontext data in use today is about 10^5 times to that of the text data [1]. Therefore, manually retrieving information from multimedia data is a laborious task. For example, an advertisement agency may want to find or verify the TV channel and time during which their advertisement has not been fully broadcast in last one month. In order to generate such a report, a whole lot of programs recorded from different TV channels for the past month has to be processed. A wide variety of multimedia data, like music, discourses, lectures, conferences, news broadcasts, sports, TV programs and telephone conversations are being recorded and stored in large volumes worldwide. They are also made available online for public access [2], as there are no longer any constraints on data storage. The data, once recorded and stored, is opaque in nature (in the form of bits) to the user, and conveys no significant information in order to organize and use it. The volume of data is large, and is increasing continuously. Therefore it is difficult to organize the data manually and we need to have an automatic method to index the data for further search and retrieval [3–5].

Audio plays an important role in classifying multimedia data [6], as it is easier to process when compared to the video data, and also the audio data contains significant information. For these reasons, commercial products of audio retrieval are emerging [7, 8]. Audio indexing is the task of dealing with analysis, storage and retrieval of multimedia data using its audio part. Classification of data into different categories is one important step for building an audio indexing system.

Most of the existing audio indexing techniques are based on the meta-data (data which describes the actual data of interest) such as the name of the composer or band, the song title, the type of music, the year, and sometimes based on lyrics. This approach requires the meta-data to be present or given manually before indexing. Thus it is not practical due to the following reasons. The volume of data is too large to be indexed manually. The user should have precise knowledge of the meta-data in order to be capable of locating and thereby retrieving the required audio clip. Some approximate knowledge of even multiple meta-data attributes may not be of any help. Inclusion of subjective attributes in the meta-data also may not help as the description of a piece of audio (e.g., type of music) by one person may not match the description by another (there are 350 different varieties of music categorized at [9]).

1.1.2 Audio Indexing System

In an audio indexing system, the query can be in the form of an audio segment instead of text. The result of the query can also be in the audio form. The goal of an audio indexing system is to provide the user with capabilities to locate, manipulate, browse, retrieve and index audio data in an efficient manner. The system should be able to present relevant results based on the user queries. It should also have capabilities for skimming through the results (playback audio at a faster rate without changing the pitch), so that the user finds the desired audio file quickly.

1.1.2.1 How it is Different from a Text-based Indexing System?

Audio contains much more information as compared to a text. For example, the written text "I want the apple" is only a sequence of four words with equal weight, but once spoken, depending on the duration and emphasis on each word, the same sentence may convey different meanings. The dimension of information in audio is not just one - it has speech, speaker, language, prosody, and a lot more information. In real life, we hear so many kinds of sounds and depending on concentration, sometimes we notice a sound and sometimes we do not. For example, background music, noise created by fans, speech of other persons (cocktail party effect) can go unnoticed. Sometimes we concentrate on some property of sound such as kind of music, musical instruments, rhythm, lyrics, singer, and depending on interest we notice some or all the properties. We perceive some pattern out of audio data and use that knowledge to organize/identify the audio content. Organization of audio data by machine is different from indexing text because of multidimensional information contained in audio data. Audio indexing system should be able to exploit higher dimensional information present in audio.

From an implementation point of view, when one talks about using audio content for browsing or other kind of search and retrieval, the first thing that is needed by any search engine or any other information retrieval system is the audio database. Design of database, normalization etc., are other issues which could be solved depending on the application, end user's requirement and type of query. However, the main issue here is how to fill the fields of database by using the information contained in opaque audio file.

Depending on the kind of audio and application different techniques (speech/speaker/ music recognition/verification/enhancement etc.) are required. For example, the possible queries for a song are: Title/album, type of music, instrument used, singer, lyrics etc. For song title, logical naming of files is required. To query the song by type of music or musical instrument, one needs to apply music analysis techniques to extract this information. To respond a query by name of the singer, one needs to segment the speech regions in the song (music/speech discrimination) and then to apply speaker recognition/verification techniques.

Since we have bulk of audio data to be used in practical applications, before taking the audio data and fitting it in any database, the first step that is required is to do a broad classification and get an index of the audio. The signature of the audio could be obtained by further applying the relevant speech/audio technique on it, and could finally be inserted into the audio database, which could further be used by search engines or other information retrieval modules. Now from the information retrieval point of view a query may either be in text or audio format. For audio query, again knowledge of the audio is needed in order to understand the actual requirement of the end user, and retrieve the information from the database.

The other difference from text indexing system is that in the case of audio indexing system, the user needs devices (hardware support in machine) to take input and give output in the audio format. The result of query could be displayed as in text indexing system or could be read out for the user, out of which he can select one or more, and listen to them one by one. User should be given the facility to forward or rewind a given result of the query in an audio format.

1.1.3 Categories of Audio Indexing

For efficient retrieval of audio, some sort of audio similarity measure for organizing similar sounds into one category is desirable. Neural network-based clustering of audio files [10] and self organizing maps (SOM) [11] have been used to organize similar sounds. Likewise in Muscle Fish [7], Wold *et al.,* have used a vector of attributes like loudness, pitch, bandwidth and harmonicity [8] to organize the sounds into similar classes. Such type of classification comes under unsupervised categorization of audio files. Supervised categorization of audio is instrumental when the user is searching for a particular known/familiar audio clip. In doing so, a portion of the clip of the original file may be presented, or the tune of the music clip may be hummed. In such cases, the distances between the histograms of a discriminately trained vector quantizer may be used [12] to categorize different training classes or one can go for query-by-humming technique developed by Ghais *et al.* [13].

1.1.4 Examples of Audio Indexing Tasks

Audio indexing encompasses speaker identification, language identification, music classification (jazz/pop/classical), music/speech discrimination (music/speech/silence), audio classification (category of audio file presented - commentary/advertisement/interview etc.) among others.

In speaker identification, we are interested in classifying an unknown speech utterance, as belonging to one from a set of M reference speakers [14]. The primary use being to retrieve speech by a specific speaker (say Gandhiji's speech). The other applications of speaker identification are indexing messages in an answering machine. Language identification is also an important issue when we are looking for a discourse in a particular language (e.g., it is known that Sathya Sai Baba normally speaks only in Telugu). This will help in eliminating irrelevant audio files that are not of that particular language, and thus narrow down the search.

As the volume of music recorded, stored and made available is large, locating a particular audio track of say, a particular instrument is a difficult task. Algorithms to identify a particular instrument based on its special characteristics would improve music retrieval performance. This type of classification falls under musical instrument identification. More often than not, a user is interested in a particular class of music like pop, jazz, rock, heavy metal, classical, instrumental etc. In such cases, classification of music [15, 16] into appropriate classes is an important aspect of audio indexing. For speech recognition or audio content retrieval, one needs to analyze the audio clip only in the regions of interest say, music or speech. For this, one needs to distinguish between such regions, which is done under music/speech discrimination [17].

Another important aspect of audio indexing is to classify a given audio clip into music, commentary, advertisement, cartoon or some other class. This is termed as audio clip classification [18]. It is normally a closed-set problem, where the given audio clip is identified as belonging to one among the known audio classes. This task is discussed in detail in the following section.

1.2 AUDIO CLIP CLASSIFICATION - PROPOSED PROBLEM

This study considers identification of TV programs using audio to classify the given audio clip as one of the following predefined five categories - cartoon, advertisement (or commercials), cricket, football and news[1].

[1]Depending on interest one may look at the problem from different angles. For example, one may be interested in knowing which TV channel is being watched, or what program is being broadcast. Such queries cannot be processed by just watching a small clip of the program (like in the case of sports, there are dedicated sports channels, and there are TV channels (like DD1) which also broadcast sports sometimes, and there may be sports clip between a news broadcast. Therefore, by using a small small clip it is difficult to say if it belongs to a sports channel or sports program or a part of the news broadcast). By considering longer enough clips and other specific properties of a particular TV channel/program, or may be some video clues, one may be able to process this kind of query. Other example is song that comes in movies, music programs, music channels, and advertisements and cartoon also have some music and speech. So answering any query related to these programs

The motivation for choosing these categories is that it covers most of the popular TV programs that are broadcast. Moreover, at the same time the task is quite challenging, as classifying cricket and football, or advertisement and cartoon is sometimes difficult for human beings also. The other important factor is that there is a significant variation in the audio of a category from different TV programs/channels. Therefore a single category can have N subcategories, and the problem indeed becomes a M class classification problem instead of 5 class classification, where M is total number of such subcategories. The task of classification becomes challenging because of intra-class variability. This study does not label these subcategories (like BBC news and DD1 news), and gives only one label (like news), i.e. the problem considered here is a 5 class classification problem.

In this study the clip has been defined as follows: An audio clip has audio belonging to a single audio category considered for study. The duration of the clip considered is about 10 sec. This was based on the premise that 10 seconds should be enough to capture the inherent temporal behavior of the audio clip.

1.3 ISSUES IN AUDIO CLIP CLASSIFICATION

Content-based classification of audio is essentially a pattern classification problem [19] in which there are two basic issues: *Feature extraction* and *classification* based on selected features. The classifier operates in two phases: Training and testing phase.

without knowledge of whole domain is quite impossible and at the same time considering whole domain for the study also does not seem feasible/logical. So for different user/application's need type/requirement of query may vary, and so the view.

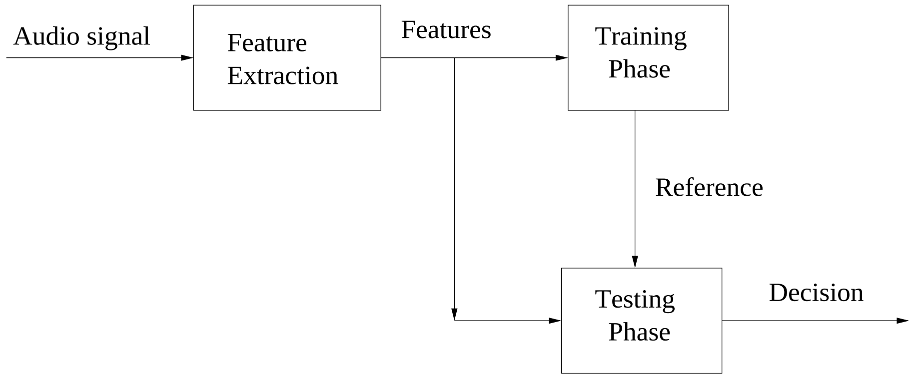


Fig. 1.1: Block diagram for pattern classification task.

Training is the process of familiarizing the system with the audio characteristics of a given category, and testing is the actual classification task, where a test audio clip is assigned a label. Audio clip classification like any other pattern recognition system can be represented as shown in Fig. 1.1. Further, if there are more than one system for the task, one needs to combine the evidences from different systems. Hence for the automatic audio clip classification system to work, there are four major tasks to be accomplished. They are given below:

(a) Extracting appropriate audio-specific features from the given audio data.

(b) Generating a model for each type of audio.

(c) Developing a matching algorithm.

(d) Developing a decision logic and combining evidences from two or more systems, if any.

We shall briefly discuss each of the above tasks.

1.3.1 Audio-specific Feature Extraction

This deals with using an effective representation of audio data that is required to capture all significant properties of audio chosen for the task, robust under various circumstances, and general enough to describe various audio sound classes. This is the key step in developing a system for pattern classification [20–26]. Feature vectors are pattern vectors extracted from the audio signal which capture the characteristics of an audio category. Most audio clip classification systems are based on features derived using spectral analysis of the signal, which carry significant properties of audio data but miss out on significant properties too. One can also get the message while listening to the signal after removing the spectral part. Hence features derived using residual part of the signal can be complementary in nature to the spectral features-based system.

1.3.2 Generating Models for Audio Clip Classification

Once proper set of feature vectors are obtained, the next task in audio clip classification is to develop models (prototypes) for each type of audio [19]. The development of audio models comes under training phase. The block diagram of the training phase is shown in Fig. 1.2.

Feature vectors representing the audio characteristics of a category are extracted in the feature extraction phase, and they are used for building the reference models. Performance of an audio clip classification system depends primarily on effectiveness

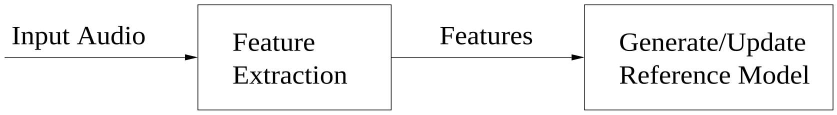


Fig. 1.2: Block diagram for training phase.

of the model in capturing the audio category-specific information.

1.3.3 Developing Matching Algorithm for Audio Clip Classification

Matching techniques generally utilize the uniqueness in the feature vector distribution or temporal behavior of feature vector sequence of audio, for discriminating the audio categories. An effective matching technique should try to reduce the variability within an audio category, and increase variability among different audio categories. This matching technique and decision logic (discussed in next section) comes under the testing phase of a pattern recognition system. The block diagram for testing phase is given in Fig. 1.3.

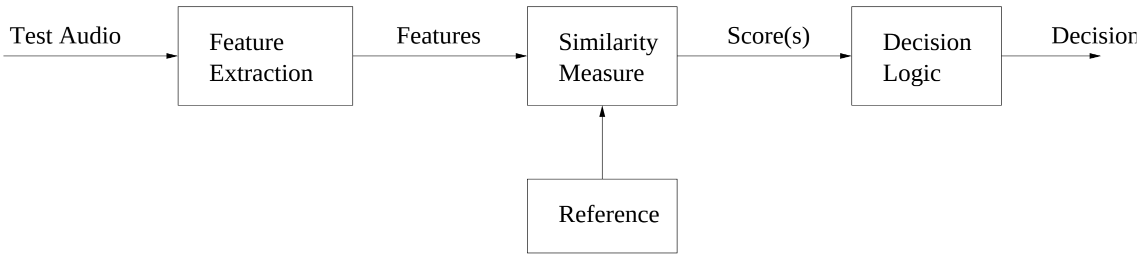


Fig. 1.3: Block diagram for testing phase.

1.3.4 Developing Decision Logic for Audio Clip Classification

The final stage in the audio clip classification system is the decision logic stage, where the decision of labeling the test audio clip with an audio category name is taken based on the result of the matching stage. Matching generally gives a score which will be a measure of how well the test feature vector matches with the reference feature vector and a decision can be taken based on these scores. However, the decision logic can also be complex involving combining of evidences from several matching scores. Combining evidences can vary from simple addition of the scores to sophisticated combination methods using neural network models [27–30]. In certain cases one single system may not be able to capture information present at various levels of audio signal. Different feature vectors can emphasize different types of information present in the audio signal and hence may need different types of models to capture this information to take decision. Since the information captured by different models can be complementary in nature, they can be combined to get a better performing system. The block diagram for combining evidences is shown in Fig. 1.4.

1.4 OBJECTIVES OF THE PRESENT STUDY

Audio clip classification is the problem of labeling a given audio clip to one of the predefined categories like news, advertisement etc. Such a task is an important step for building an audio indexing system. An audio signal can be interpreted as a result of time varying excitation given to a system. Therefore information about the audio

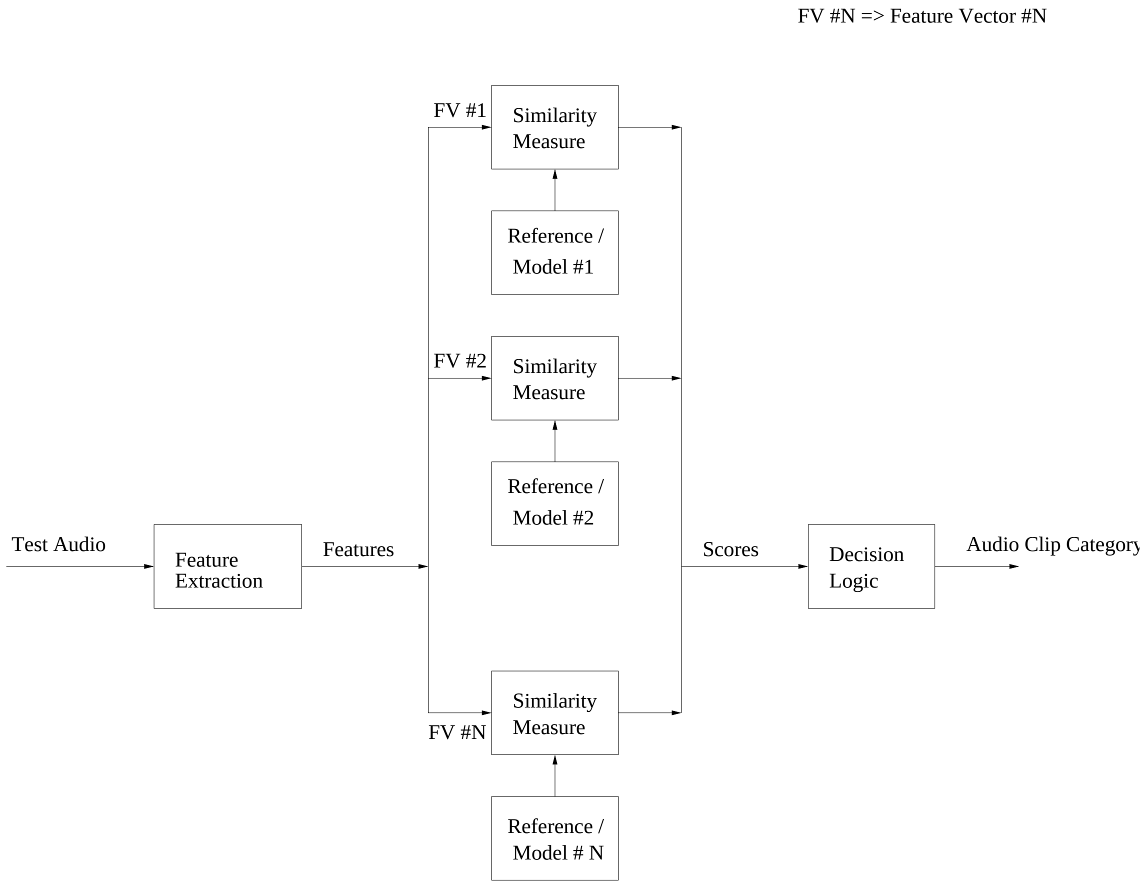


Fig. 1.4: Block diagram for combining evidences.

category is contained in the excitation source, system (physiological) and behavioral characteristics of audio data. The features related to excitation source characteristics are extracted from a very short segment of audio signal (typically $1-5$ msec) and hence are called as subsegmental features. The features related to physiological or system characteristics (such as vocal-tract characteristics in case of speech) are extracted from a short segment of audio signal (typically $10-50$ msec) and hence are called segmental features. The features related to the behavioral characteristics are extracted from a relatively longer segment of audio signal (typically $100 - 300$ msec), and hence are called suprasegmental features[2].

One objective of this study is to explore features present at various levels of the audio signal, in addition to the features that are currently used, in order to improve the performance of an audio indexing system. In particular, features that are not used explicitly or implicitly in the existing audio indexing systems are being investigated. Many interesting and perceptually important features are present in the residual signal that is obtained after removing the predictable part at the short segment level. Perceptually, there are some discriminating features present in the residual which could help in various audio indexing tasks [31]. The challenge lies in developing algorithms to capture these perceptually significant features from the residual, as it is difficult to extract information using known signal processing algorithms. Thus the main objective of this study is to explore the features present in the linear prediction (LP) residual [32] for the audio clip classification task. The reason for considering the residual data in this study is that the residual part of the signal contains perceptually significant audio-specific information, and it may also be subject to less degradation as compared to the system part [33]. In this thesis we study the following -

(a) Subsegmental information present in the LP residual signal is studied. The residual data contains higher order relation among samples. As known signal processing and statistical techniques are not suitable to capture this relation,

[2]Here the duration of the signal used to study the information present at various levels is derived based on speech knowledge. This could vary for different types of audio signals.

an autoassociative neural network (AANN) model is proposed to capture these higher order relations among samples of the residual signal [34]. The AANN has been studied to capture information from the residual signal for tasks such as speaker recognition [35–40]. Each audio category is composed of one or more audio components (like music, speech), which individually or together are specific to a particular audio category. The components present in an audio clip are recognized using the AANN models. The multilayer feedforward neural network (MLFFNN) or multilayer perceptron (MLP) is proposed for deciding audio category using this component knowledge captured by the AANN. The MLP has been shown to be suitable for pattern classification tasks because of their ability to form complex decision surface by using discriminating learning algorithms [41–49].

(b) The LP residual is obtained after removing the spectral part of the signal [32]. Therefore information captured using the LP residual is complementary in nature to the information captured using spectral features. Hence the complementary nature of subsegmental information (obtained using the LP residual signal) to the segmental information (obtained using spectral part of the signal) is studied, which on combining give a better performing audio clip classification system [40, 50].

(c) The presence of audio-specific suprasegmental information in the LP residual is also studied. This information can also be useful for audio clip classification.

1.5 ORGANIZATION OF THE THESIS

The thesis is organized as follows:

Chapter 2 gives a brief review of various audio indexing techniques. The features and matching techniques used for various audio indexing tasks are discussed. The drawbacks of existing techniques and its significance are also discussed.

A baseline system for audio clip classification using spectral (segmental) features is discussed in chapter 3. However, it is shown that perceptually significant information of audio is present in the form of sequence of events, which is obtained after removing the predictable part of the audio signal [31]. Chapter 4 discusses extraction of this information present in the LP residual using the AANN models for audio clip classification task.

Chapter 5 discusses the ability of the MLP to use the information captured by the AANN for audio clip classification task. This chapter also discusses the complementary nature of this information to the spectral (segmental) information, and presents an approach to combine evidences from the two systems capturing information at subsegmental and segmental levels of audio to build a better audio clip classification system.

The long term behavioral characteristics of one audio category are different from that of other, and hence can be useful for audio clip classification task. Chapter 6 discusses the presence of suprasegmental information in the LP residual that can be useful for audio clip classification. Finally, chapter 7 summarizes the work done with major contributions, conclusions and future scope of the present work.

CHAPTER 2

REVIEW OF APPROACHES TO AUDIO INDEXING

The majority of the data used today in various applications is in nontext format [1]. As the volume of the data is too large to be manually indexed, we need to have an automatic method to organize it. Audio indexing is the task of dealing with the storage and retrieval of audio data. Audio clip classification is an important step in building an audio indexing system. Since dimension of information in audio is more than one, it is difficult to study all the dimensions together at the same time. Therefore, depending on the application or the audio data in hand the task of audio indexing is addressed differently. This chapter gives an overview of various audio indexing tasks. It also analyzes various approaches to index the audio data and concludes with a discussion on the limitations of conventional approaches to audio indexing, followed by a discussion on the new approach proposed to overcome these limitations.

2.1 TYPES OF AUDIO INDEXING TASKS

2.1.1 Spoken Document Retrieval

In spoken document retrieval (SDR), the goal is to retrieve documents from a large collection of spoken documents like news bulletins. The task is to retrieve documents that have similar contents of keywords specified in the query by user. Conventional

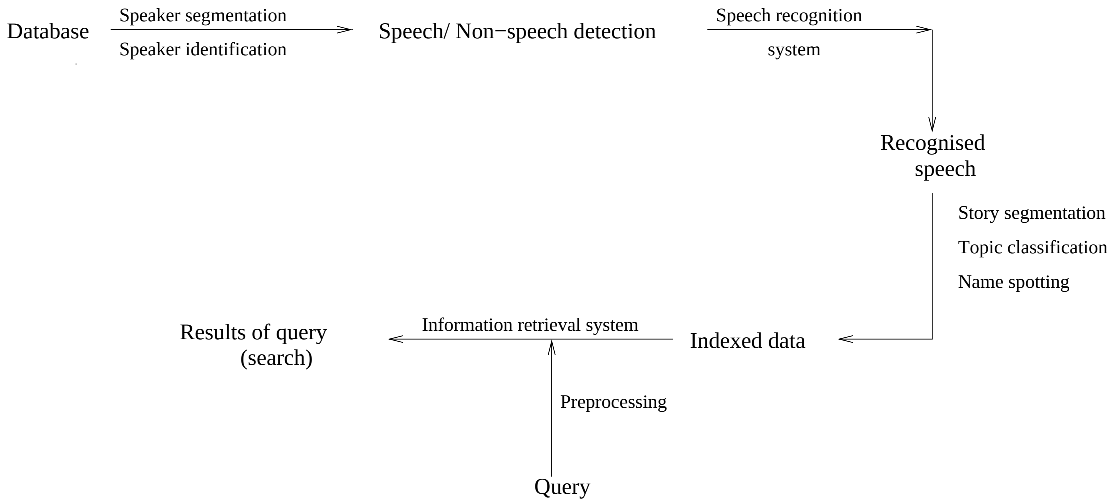


Fig. 2.1: Schematic diagram showing the task of SDR.

approach is to convert the news audio to text before indexing and retrieval. In combination with the text-based information retrieval techniques, the task of SDR can be represented schematically as shown in Fig. 2.1. Makhoul *et al.,* presented speech and language technologies for SDR [51].

2.1.2 Other Audio Indexing Tasks

SDR deals with a special case of audio in use. The drawback of this approach is that one is no longer using any speech-specific knowledge once speech is converted to text. Also, this approach will not work for nontext audio like music. Existing research on content-based audio data management is very limited. A large number of audio types (sounds) are in use, and in general what one hears is a combination of audio sounds.

So depending on the requirements of the application various subdomains have been addressed.

Zhang *et al.,* [52] mention four directions in the research of content-based audio management:

(a) First direction is content-based classification of audio for segmentation purpose. For example, classification of audio in speech and music [17]. Saraceno *et al.,* classify it in four audio groups - silence, speech, music and noise [6]. Wold *et al.,* classify it in ten groups of sound units like laughter, music, speech, sound of bells [8]. Audio is divided into speech, music, silence and environment sounds in [53]. Liu *et al.,* classify audio into five classes of television (TV) programs [18]. Pfeiffer *et al.,* detect violence in the audio of movies to rate it [54]. Matityaho *et al.,* classify music audio into classical and pop [55]. Lambrou *et al.,* classify music audio into rock, piano and jazz [15]. Wyse *et al.,* classify audio into music, speech and others for parsing of news stories [56]. In [57], audio recordings are classified into speech, laughter and nonspeech sounds for the purpose of segmenting discussions in meetings.

(b) The second direction to address the problem is content-based retrieval. One specific technique in content-based retrieval is query-by-humming [13]. Work on generic audio retrieval is presented in [8, 12] also.

(c) The third direction is audio analysis for video indexing. Liu *et al.,* presented work on five kind of video scenes [58]. Audio characterization is performed on moving picture experts group (MPEG) sub-band level data for the purpose of video segmentation and indexing [59].

(d) The fourth direction is the integration of audio and visual information for video segmentation and indexing [60, 61].

2.2 FEATURE EXTRACTION FOR AUDIO INDEXING

2.2.1 Characteristics of a Feature

Content-based organization of audio data is a pattern recognition problem, in which the first basic issue is to find an effective representation (feature) of the audio data. Feature extraction is important as the raw signal is recorded in the form of bits and sampling frequency, which do not convey the information present in audio. Hence, it cannot be used for the task. Moreover, by converting it to any other form (like text or vector quantized codebook), the audio-specific information present in the signal is lost. Feature extraction helps to reduce the volume of data to be processed. These features must occur frequently, should be easily measurable, should not vary with time, and must not be distorted by channel variation and background noise [62]. These features may be derived from the data or may be based on the knowledge of the characteristics of the data. However, in all cases of audio indexing, one cannot derive a feature that has universal discrimination capability, to discriminate between speakers, languages, music, musical instruments etc. Each feature has its own limitations on the ability to capture the information in the given data. This necessitates the need for a separate set of features depending on the task in hand.

2.2.2 Features for Speaker Recognition

The set of features commonly used for speaker recognition are formants of retroflex vowels and nasal, voice-pitch frequency, amplitude spectrum of vowels and nasal, slope of the glottal source spectrum, word duration and pitch frequency [14, 62]. Prosodic features like mean F_0 [63], mean and variance of the pitch periods in voiced segments and energy contours of speech [64] have been used. The first four statistics of the pitch, that is, mean, variance, skew and kurtosis have also been used [65].

2.2.3 Features for Language Identification

In the case of language identification the prosodic features that have been explored are speech rate, syllable timing etc., [66]. Pitch contour, differential pitch, size (distance between syllables and syllable duration), the differential size (differenced distance between syllables and syllable duration), shape of amplitude contour, differential amplitude, rhythm (low frequency Fourier transform of the amplitude envelope) and phrase location, were found to be useful for language identification [66].

2.2.4 Features for Music Classification

Loudness, pitch, bandwidth and harmonicity have been used to classify music [67]. Music/speech discrimination has been done using the average length of time in which peaks exist in a narrow frequency range [56]. To retrieve music, query-by-humming has been reported by Ghais *et al.,* [13].

2.2.5 Features for General Case of Audio Indexing

In the conventional approach to indexing spoken documents, where speech is first converted to text prior to indexing, cepstral coefficients have been the predominant choice as features for training hidden Markov models (HMMs) [68]. For the general case of audio indexing, features like energy contours, zero crossing rate and spectral centroid have been used [10]. For the cases where we have an exact audio segment to search, a quick searching of the broadcast audio to locate and detect sound segments containing a certain reference templates based on an active search algorithm and histogram modeling of zero crossing features has been proposed by Smith *et al.,* [69].

2.2.6 Features for Audio Clip Classification

For audio clip classification, an important recent work is done by Wold *et al.,* [8] represented by their system called "Muscle Fish". In this system, various perceptual features are used to represent a sound. A normalized Euclidean (Mahalanobis) distance and the nearest neighbor rule are used to classify and query the sounds into one of the sound classes in the database. Foote [12] chose to use 12 mel-frequency cepstral coefficient (MFCC) and energy as the audio feature. Moreover, distance between the histogram of a discriminately trained vector quantized (VQ) codes is used to classify sounds. In a work by Pfeiffer *et al.,* [54], a filter bank consisting of 256 phase compensated gamma phase filters proposed by Cook [70] is used to extract audio features.

In [52], Zhang *et al.,* uses spectral features like energy function, average zero crossing rate, fundamental frequency and spectral peak tracks with a rule-based approach to classify audio into seven categories (harmonic and nonharmonic environmental sound, pure music, pure speech, song, speech and environmental sound with music background). In [71], Subramanya *et al.,* describe transform-based indexing of audio data for use in multimedia database. Indicators like loudness, frequency, pitch, onset, offset have been used for violence detection in audio of a movie [54]. Lu *et al.,* use spectral features derived from zero crossing rate, energy, spectrum to classify audio stream into silence, speech, music and environmental sounds [72].

2.3 TECHNIQUES FOR AUDIO INDEXING

After choosing task-specific features, the next important task to be performed is to formulate a distance measure and rule/model for the task. Techniques used for various audio indexing tasks can be divided into following three categories:

(a) Model-based techniques

(b) Metric-based techniques

(c) Rule-based techniques

It has been known that model-based and metric-based techniques outperform the simpler rule-based technique. In the case of audio segmentation, model-based segmenters achieve very high level of segment boundary precision, the metric-based segmenter performs better in terms of segment boundary recall [73]. Each of these techniques has its own merits and demerits. Depending on the task in hand, one can choose a hierarchical combination of these techniques.

2.3.1 Model-based Techniques

Different models, such as, Gaussian mixture models, are constructed for a fixed set of acoustic classes, such as anchor speaker, language, music etc., using a training corpus. The incoming audio stream can be classified by maximum likelihood estimation. In the case of segmentation, segment boundaries are assumed where a change in the acoustic class occurs. In the model-based segmentation [74, 75], a set of models for different acoustic classes is defined and trained for the given task. The incoming audio stream is classified using the models, usually imposing additional minimum class length constraints which is used to define the window-size for experimental study. Boundaries between the classes are used as segment boundaries. Model-based techniques assumes knowledge about the type of audio that is to be used for the task. HMMs have been used for various audio classification tasks [18, 60]. In [76], support vector machines (SVM) with a binary tree recognition strategy are used to tackle the audio classification problem.

2.3.2 Metric-based Techniques

In the case of audio segmentation, the audio stream is segmented at maxima of the distances between neighboring windows placed in evenly spaced time intervals. In this technique, two neighboring windows of relatively small size are moved over the audio signal. The similarity between the contents of the two windows is computed using a distance function. Local maxima exceeding a threshold indicate segment boundaries. For classification, distance between feature vectors of audio clips is measured, and the clips having least distance between their feature vectors are grouped in the same class. In audio verification/identification case, the distance between reference(s) and test audio clip is measured. If this distance is less than a pre-defined threshold, the clip is said to be verified/identified.

Various metric-based algorithms differ in the kind of the distance function they employ, the size of the two windows, the time increments of the shifting of the two windows, and the way the resulting similarity values are evaluated and thresholded. The use of a symmetrized Kullback-Leibler distance metric is proposed in [77]. A distance measure introduced by Gish [78] is proposed in [73]. This metric is based on a likelihood ratio hypothesis test. The null hypothesis assumes that the data in both neighboring windows or audio clips has been produced by the same stochastic source. The alternate hypothesis assumes that the data has been produced by different sources. For the actual computation of the likelihoods in [73], one multivariate full covariance Gaussian model is estimated on each of the windows of audio clips and on the union of the two windows. The threshold, which is used to decide whether a segment boundary is introduced or not, can be determined by using information theoretical measures, such as minimum description length (MDL) [79], Bayesian information criterion [73]. This approach has the advantage that no hard thresholds need to be defined, and hence the algorithm is robust against changes of the acoustic conditions.

In [80], Li proposed a method called nearest feature line (NFL) for audio classification. In the NFL method, information provided by multiple prototypes has been explored. This contrasts the nearest neighborhood classification in which the test clip is compared to each prototype individually.

2.3.3 Rule-based Techniques

In this method audio-specific features are derived from energy, zero crossing rate, etc., of the audio signal. Further, the audio is categorized/segmented/recognized according to a set of rules defined using the features. For example, the silence in the input audio stream is detected either by a decoder or directly by measuring and thresholding the audio energy. The audio segments are generated by cutting the input at silence locations. Approaches based on energy features have been widely used [81, 82] and are particularly easy to implement. In this, the silence periods in the input signal are detected, and segment boundaries are hypothesized in such silence periods if some additional constraints, like minimum length of the silence period, are satisfied. A rule-based approach to audio classification using spectral features is presented in [52, 72].

2.4 DISCUSSION

2.4.1 Missing Components in Existing Approaches and their Importance

(a) The features used for the task of audio clip classification are mostly based on spectrum, amplitude, zero crossing and periodicity. These features describe the system (medium) characteristics of the signal, whereas it is known that the message part of the signal is contained mostly in the excitation source. The characteristics of the excitation source can be observed better by removing the spectral information from the signal. Doing so, the residual signal is obtained. Extracting features from this residual is a challenging task.

(b) In the case of system based on spectral features performance degrades drastically in case of noisy data and channel distortion.

(c) Large amount of training data is needed to train models in the model-based approach. In the case of rule-based approach justifying the threshold is very difficult.

2.4.2 Proposed Methods to Overcome the Limitations of Existing Approaches

To overcome these problems, it is necessary to use features that capture audio-specific information instead of the distribution of the feature vectors. Present work overcomes some of these problems. Excitation source characteristics of the audio are used to capture the audio properties instead of spectral features. Since excitation source characteristics are likely to be more robust to noise, the effect of environment on the system's performance can be reduced. In the proposed approach, the audio-specific information is captured directly by the autoassociative neural network models, in terms of source characteristics using the LP residual signal of the given audio. The AANN models have been used to capture speech and speaker-specific information [38, 83, 84]. Further, since the source information is obtained after removing spectral part of the signal, the complementary nature of both kinds of information is studied to get an improved system.

2.5 SUMMARY

In this chapter the review of research work done in the area of audio indexing is discussed. Apart from the techniques to capture the audio-specific information, the limitations of existing approaches have also been discussed. The need for a new approach to overcome some of these limitations is highlighted.

CHAPTER 3

BASELINE SYSTEM USING SEGMENTAL FEATURES

In the previous chapter, existing approaches to various audio indexing tasks were reviewed. In this chapter, a baseline system using spectral features for the audio clip classification task is discussed. The audio is perceived and labeled by humans effortlessly as we focus our attention around some events. But if the length of clip played is very small, it becomes difficult, if not impossible even for us to determine the category of the given audio clip. This implies that temporal changes occurring in an audio signal do play an important role in perception. In the baseline system, this temporal behavior of audio is exploited for the classification task.

This chapter is organized as follows: Extraction of spectral features to represent the audio data for the classification task is discussed in Section 3.1. Section 3.2 discusses the development of models to capture the essence of various classes. The data used for this study is discussed in Section 3.3. Classification results for the baseline system are given in Section 3.4. Section 3.5 summarizes the chapter.

3.1 SPECTRAL (SEGMENTAL) FEATURES FOR AUDIO CLIP CLASSIFICATION

Feature extraction is the most crucial phase for any classification or indexing task because it is these features which represent the original data in all the later stages of classification. Any kind of inaccuracy incurred at this stage may result in unjustifiable aberrations in the final classification.

Audio data is collected at 22050 Hz sampling frequency, and is stored as 16 bit integers. The length of the audio clip is 10 sec. For the purpose of feature extraction, a frame as an audio segment of around 10 ms duration is defined. The duration is kept small so that the feature value does not vary significantly within a frame. The features extracted for a frame are volume, zero crossing rate, pitch, frequency centroid, bandwidth and energy ratio in various frequency subbands to the total energy. From these low-durational features, gross features for a window consisting of 100 frames are extracted. A set of 14 features is used to represent the audio data. The volume-based features are standard deviation of volume, dynamic range of volume, volume undulation [85] and 4Hz modulation energy [67]. The features based on zero crossing rate are standard deviation of ZCR and nonsilence ratio (the ratio of number of nonsilence frames to the total number of frames in the window). Pitch-based features include pitch standard deviation, similar pitch ratio (the ratio of number of neighboring frames having similar pitch to the total number of frames), nonpitch ratio (the ratio of frames with zero pitch i.e., frames for which pitch is not defined, to the total number of frames). Spectrum-based features used are frequency centroid, bandwidth and ratio of energy in various frequency subbands to the total energy.

Fig. 3.1 (a) shows waveform of an audio clip belonging to cricket category. This clip is used as benchmark to analyze the efficacy of feature extraction process. This clip contains background noise throughout its length. There is a sound of ball hitting the bat and ground in the first half of the clip which appears as few spikes in the waveform. The later half of the clip contains speech (commentary) along with the background noise.

3.1.1 Volume-based Features

Volume gives a measure of loudness or amplitude of the given audio signal. Volume, being a low-durational feature, is extracted for each frame. Volume is approximated by rms (root mean square) value of the frame at hand i.e,

$$volume = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} s^2(n)} \tag{3.1}$$

where N is the total number of samples in the frame and $s(n)$ is the magnitude of the n^{th} sample of the frame. By taking volume as frame level feature, several window level features are computed. These features and their extraction are described in the following subsections.

3.1.1.1 Volume Standard Deviation (VSTD)

Volume standard deviation is the standard deviation of volume across all the frames of a window. Standard deviation of a distribution gives some idea about the variability

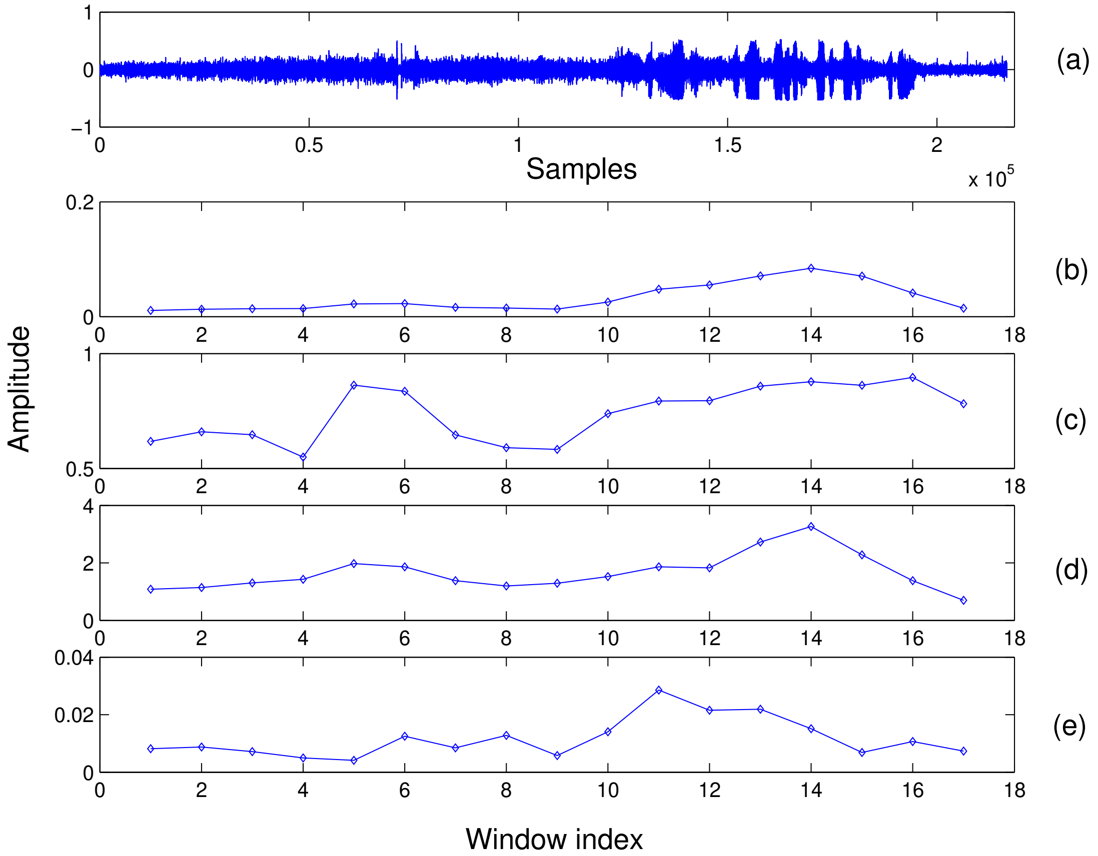


Fig. 3.1: (a) Benchmark cricket clip, (b) Volume standard deviation, (c) Volume dynamic range, (d) Volume undulation, (e) 4Hz modulation energy.

(deviation from mean) of the quantity observed. Formally, volume standard deviation can be defined as,

$$VSTD = \sqrt{\frac{\sum_{f=0}^{F-1}(v(f) - \mu_v)^2}{\sum_{f=0}^{F-1} v^2(f)}} \tag{3.2}$$

where F is the number of frames in the window, $v(f)$ is the volume of f^{th} frame and μ_v is the mean volume for the window. A speech segment consists of interspersed voiced and unvoiced segments. A typical voiced segment has high volume in contrast with an unvoiced one which has low volume. Due to the occurrence of alternating high and low volume regions, one can expect standard deviation of volume to be high for speech segments. Fig. 3.1 (b) shows volume standard deviation for our benchmark cricket clip. Quite clearly, VSTD is high in region containing noisy commentary. It is low for rest of the regions containing only background noise.

3.1.1.2 Volume Dynamic Range (VDR)

Volume dynamic range measures the range in which volume varies across various frames in the clip. A window containing both silence and high volume frames will have high VDR. Formally, it can be written as,

$$VDR = \frac{v_{max} - v_{min}}{v_{max}} \tag{3.3}$$

where v_{max} is the maximum frame volume and v_{min} is the minimum frame volume for the window. The difference is divided by maximum volume to normalize the quantity obtained. Fig. 3.1 (c) shows the VDR plot for the benchmark cricket clip. VDR is high for the region of the clip containing the sound of ball hitting the bat and ground. This is due to the occurrence of short spiky region in between relatively low volume region. As can be seen, VDR for speech segment is also high. The reason behind this is interspersing of voiced and unvoiced segments in the speech region.

3.1.1.3 Volume Undulation (VU)

Both VSTD and VDR measure gross statistical properties of volume which do not depend exactly on frame to frame variation of volume. Even if the sequence of frames in the clip is shuffled, VSTD and VDR values will remain same. Volume undulation is sensitive to the ordering of frames in the clip. This way it contains some information which is not contained in VSTD and VDR. As name suggests, VU captures the change in volume from one frame to the next frame. It is estimated as the accumulated sum of difference between volumes of adjacent frames. Formally, the same can be written as,

$$VU = \sum_{f=1}^{F-1}(v(f) - v(f-1)) \tag{3.4}$$

where F is the number of frames in the window and $v(f)$ is the volume of the f^{th} frame of the window. The volume undulation is not very high for noisy regions because frame to frame variation is captured and not sample to sample variation. Fig. 3.1 (d) shows the volume undulation plot for the benchmark clip. As can be seen, VU is high for speech region due to the interlacing of voiced and unvoiced regions.

3.1.1.4 4Hz Modulation Energy (4ME)

The volume-based feature discussed above are general signal features which lack any particular focus on audio data, and derived from volume contour of the clip directly. Whereas, 4Hz modulation energy is computed from the frequency spectrum of volume. It measures the energy around 4Hz in the frequency spectrum of the volume contour of the clip. 4ME is typically high for speech segments. This is based on the premise that a typical speech contains 4-5 syllables per second. Formally, 4ME can be written as,

$$4ME = \frac{\int_0^\infty W(\omega) | F(\omega) |^2 d\omega}{\int_0^\infty | F(\omega) |^2 d\omega} \tag{3.5}$$

where ω denotes frequency, $W(.)$ is a triangular window function around 4Hz and $F(.)$ is the Fourier transform of volume contour. Fig. 3.1 (e) shows 4ME plot for the benchmark clip. As expected, 4ME is high for speech region. This feature may be used to segment an audio clip between speech and other reasonably loud regions.

3.1.2 Zero Crossing Rate (ZCR)-based Features

Literally, zero crossing rate of a signal means the rate at which the signal crosses the time-axis. So it denotes the frequency at which the signal changes its sign. It can be computed as,

$$ZCR = \frac{f_s}{N} \sum_{n=1}^{N-1} (sign(s(n)) \sim sign(s(n-1))) \tag{3.6}$$

where N is the total number of samples in the frame, $s(n)$ is the n^{th} sample of the frame, f_s is sampling frequency, $sign(s(n))$ is the sign of $s(n)$ and $\sim$ is difference operator.

3.1.2.1 ZCR Standard Deviation (ZSTD)

This denotes standard deviation of zero crossing rate across all the frames of the window. Formally, it can be written as,

$$ZSTD = \sqrt{\frac{\sum_{f=0}^{F-1} (z(f) - \mu_z)^2}{\sum_{f=0}^{F-1} z^2(f)}} \tag{3.7}$$

where F is the number of frames in the window, $z(f)$ is ZCR value of f^{th} frame of the window and μ_z is the mean ZCR for the clip. ZCR value for a typical unvoiced region of speech is very high compared to voiced speech. Therefore, ZSTD should be high for speech segments where high and low ZCR regions are interspersed. Although, ZCR for noisy regions is high but ZSTD for such regions is low due to relatively low variation in ZCR value across the frames. As can be seen in Fig. 3.2 (b), speech region has much higher ZSTD value compared to rest of the clip.

3.1.2.2 Nonsilence Ratio (NSR)

Nonsilence ratio denotes the ratio of number of nonsilent frames to the total number of frames in the clip. Silence/nonsilence decision cannot be made using volume as such an approach will result in marking the unvoiced speech regions, which are typically low-energy regions, as silent. This can be avoided if ZCR knowledge is also used to make the decision. This is based on the premise that a typical unvoiced speech region has high ZCR value. Based on the data, one can arrive at appropriate thresholds for volume and ZCR values. If either volume or ZCR value of a frame is more that the respective threshold, the frame should be marked as nonsilent. Fig. 3.2 (c) shows NSR plot for the benchmark cricket clip. Due to presence of persistent background noise, there is not much of a variation in the value of NSR. Hence, this feature is not useful in discriminating between categories like cricket and football.

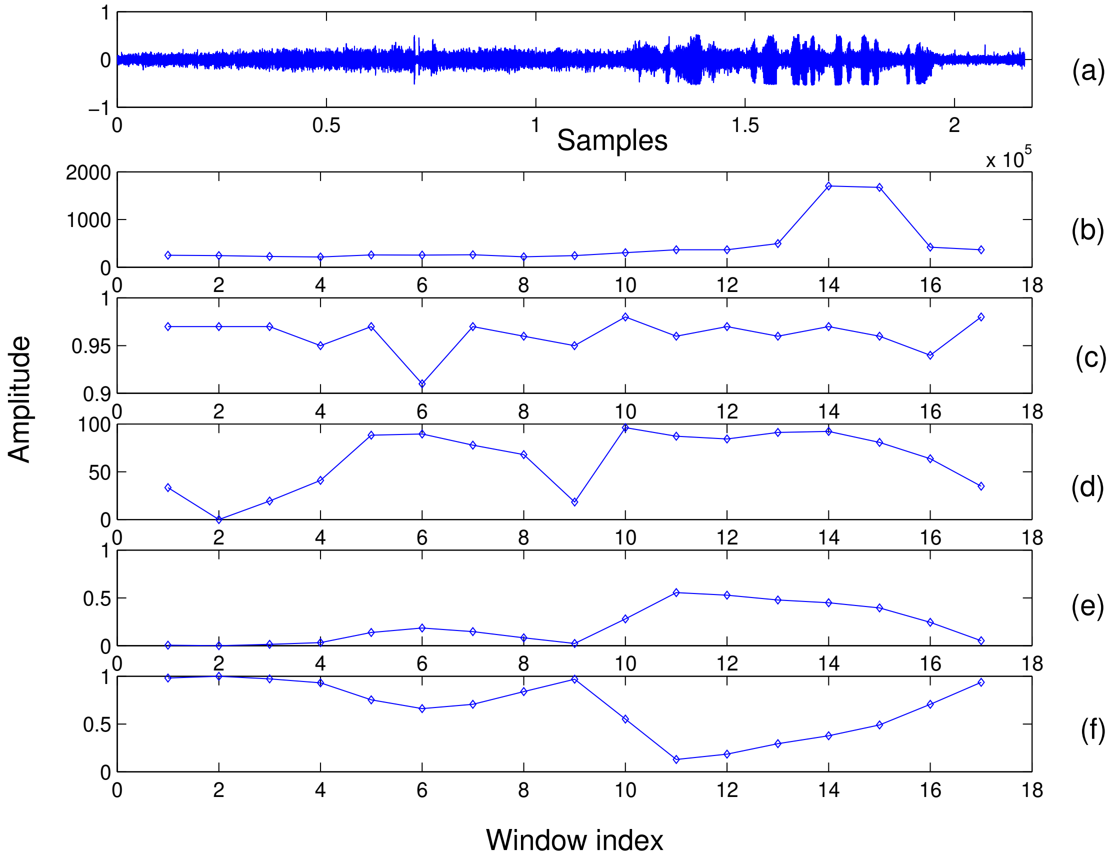


Fig. 3.2: (a) Benchmark cricket clip, (b) Zero crossing rate standard deviation, (c) Nonsilence ratio, (d) Pitch standard deviation, (e) Similar pitch ratio, (f) Nonpitch ratio.

3.1.3 Pitch-based Features

Pitch is a perceptual feature of speech. Determining the fundamental frequency (F_0) or pitch of a signal is important in many speech applications. Voiced speech involves vibration of the vocal cords and pitch refers to the fundamental frequency of such vibration or the resulting periodicity in the speech signal. It is the primary acoustic cue to intonation and stress in speech and is critical to phoneme identification in languages.

Pitch determination is not very difficult for most speech regions, but because of the nonstationary nature of speech and irregularities in vocal cord vibration, the wide range of possible F_0 values, interaction of F_0 with vocal-tract shape, and degraded speech in noisy environments, 100% accuracy has eluded the many published algorithms.

In the baseline system the Hilbert envelope [86] approach is used to get the desired pitch contour. This method is robust to degraded speech. Firstly, the residual is computed for a given audio signal based on the linear prediction model [32], where the signal is predicted based on a linear weighted sum of the past sample values. Extraction of residual using the LP analysis is explained in Appendix A.

The residual contains the nonpredictable part of the original audio signal waveform. For vowels, the spectrum of the residual is approximately flat, since the effects of the formants have been eliminated. This suggests that residual can be utilized to extract pitch information (the error signal or residual exhibits sharp pulses corresponding to the pitch periods of the vowels). These pulses (excitation peaks) are the event markers. The sequence of these events contain important perceptual information about the source of excitation. By listening to the residual of different types of audio clips, one can distinguish between speakers, music, instruments, etc.

A noisy waveform and the corresponding residual for voiced (vowel) segment of a news clip are shown in Fig. 3.3 (a), (b). It is observed that the formant regions get deemphasized in the residual. This results in better manifestation of excitation peaks in the residual.

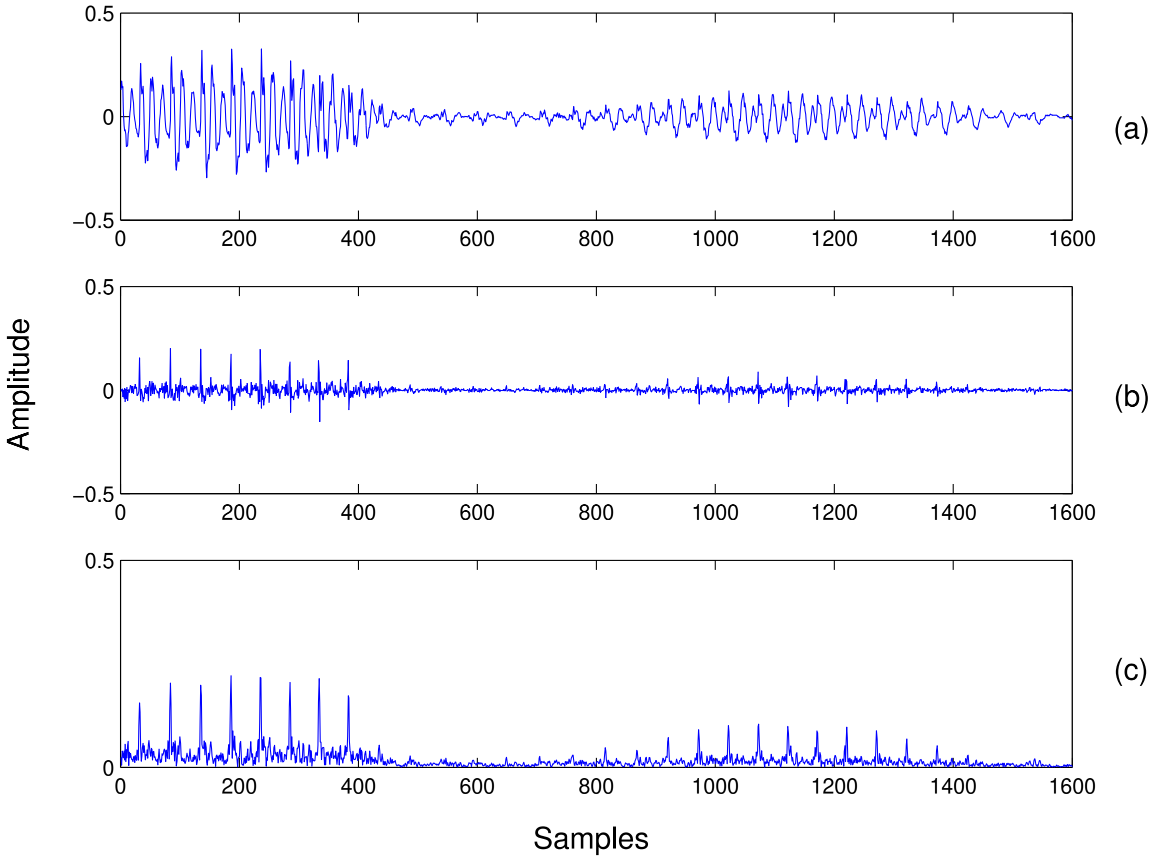


Fig. 3.3: (a) The speech waveform, (b) Residual, (c) Hilbert envelope.

This residual is used to compute the Hilbert envelope [86], where the excitation peaks are manifested in a better way. The Hilbert envelope is the magnitude of the analytical signal. The Hilbert Envelope computation removes the phase information present in the residual. This leads to better manifestation of excitation peaks, as shown in Fig. 3.3 (c), compared to the residual in Fig. 3.3 (b). The gap between excitation peaks is the pitch period. Therefore, in order to obtain the pitch contour, the excitation peaks have to be located accurately. The frame-wise autocorrelation for this is performed. As pitch is not defined for audio segments other than voiced speech, some post-processing needs to be done to set the pitch values for such regions as some sentinel. For this the use of properties (like, pitch period cannot vary in a very drastic manner) is made. Once the pitch contour is obtained, it is used to extract the desired clip level features.

3.1.3.1 Pitch Standard Deviation (PSTD)

PSTD is the standard deviation of pitch computed over all frames of a window. Formally, it can be defined as,

$$PSTD = \sqrt{\frac{\sum_{f=0}^{F-1}(P(f) - \mu_p)^2}{\sum_{f=0}^{F-1} P^2(f)}} \tag{3.8}$$

where $P(f)$ is the pitch of the f^{th} frame of the window and μ_p is the mean pitch value for the window. PSTD for speech regions is high due to interlacing of voiced and unvoiced regions (as pitch has some valid value for voiced regions while it is set to zero for unvoiced regions during post-processing). For window containing only noise etc., PSTD will generally be low as shown in Fig. 3.2 (d).

3.1.3.2 Similar Pitch Ratio (SPR)

Similar pitch ratio is the ratio of number of adjacent frames having similar pitch to the total number of frames in the window. By similar, it is meant that the two pitch values do not differ by a large extent. This feature is a measure of the density of voiced speech frames in the given window as pitch remains more or less uniform within a voiced segment. Therefore, SPR for window containing speech will be more than the ones containing some other category of audio. Fig. 3.2 (e) shows SPR for the benchmark clip. One can easily distinguish between speech and other segments as can be seen from the plot. The region containing the sound of ball hitting bat and ground also shows a slightly high SPR values due to the presence of speech-like sound.

3.1.3.3 Nonpitch Ratio (NPR)

Nonpitch ratio is the ratio of number of frames having zero pitch (i.e., the frames containing those segments of audio for which pitch is not defined) to the total number of frames in the window. Quite clearly, SPR and NPR should follow a inverse relationship i.e., the windows having high SPR should have low NPR and vice-versa. These two features may seem to be interdependent but sometimes they do carry some independent information also as is the case with a window containing speech of speakers with very different pitch. Fig. 3.2 (f) shows NPR plot for the benchmark clip. It appears as an inverted version of the SPR plot shown in Fig. 3.2 (e).

3.1.4 Spectrum-based Features

There are features which are not perceivable in the waveform directly but can be computed from its frequency counterpart. Many such features are incorporated in the feature set of the baseline system as given below,

3.1.4.1 Frequency Centroid (FC)

The centroid of a distribution is usually the weighted center of the distribution (like weighted average). On similar lines, frequency centroid is computed for each frame of the window. It is given by,

$$FC = \frac{\int_0^\infty \omega S(n)(\omega) d\omega}{\int_0^\infty S(n)(\omega) d\omega} \tag{3.9}$$

where ω denotes frequency and $S(n)(\omega)$ denotes the frequency spectrum of the given audio signal $S(n)$.

To compute the frequency centroid of a window, the weighted average of the frequency centroids of all the frames in the window is taken. Volume is taken as the weight. This is done to reduce the effect of low energy regions. Normally speech regions have lower frequency centroid compared to noisy regions as can be seen in Fig. 3.4 (b).

3.1.4.2 Bandwidth (BW)

At the frame level, the bandwidth is calculated as,

$$BW = \sqrt{\frac{\int_0^\infty (\omega - FC)^2 S(n)(\omega) d\omega}{\int_0^\infty S(n)(\omega) d\omega}} \tag{3.10}$$

where ω denotes frequency, $S(n)(\omega)$ denotes frequency spectrum and FC denotes the frequency centroid of the frame.

The weighted average of bandwidths of all frames in the window is taken to compute the window level feature. The volume is taken as the weight to compute the average. Fig. 3.4 (c) shows the bandwidth plot for the benchmark cricket clip.

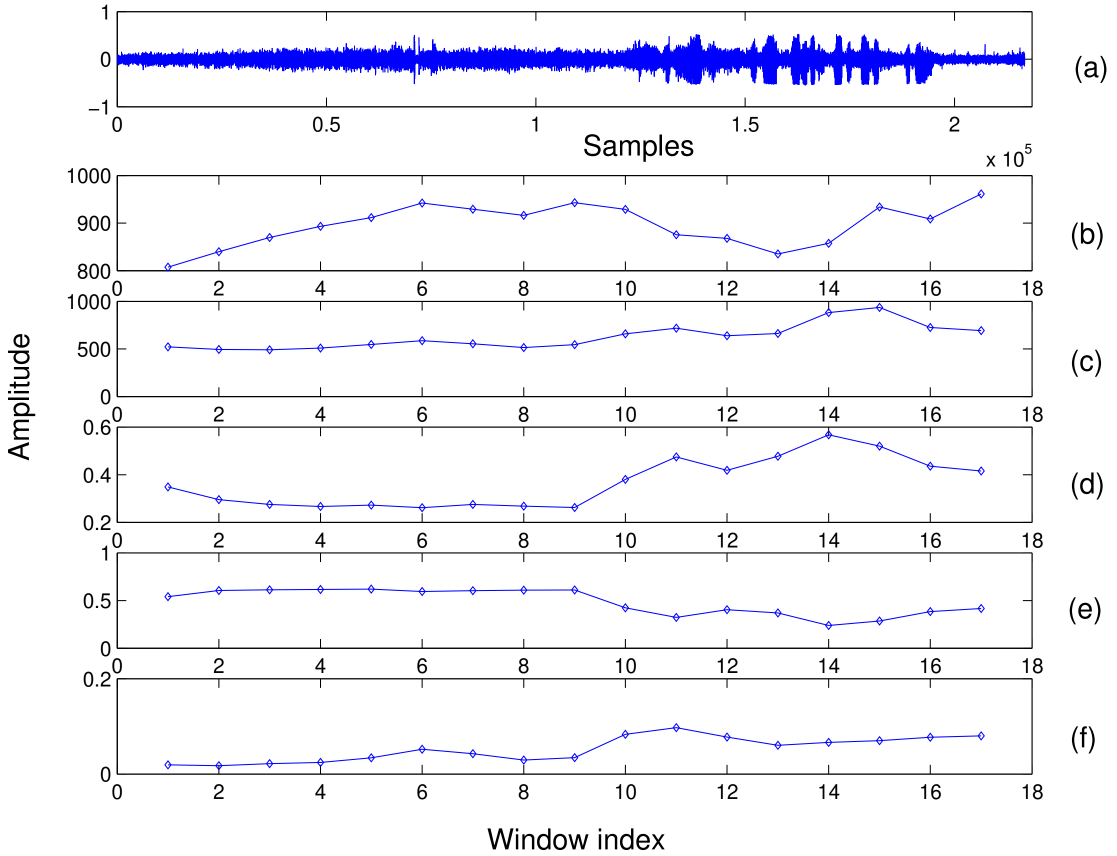


Fig. 3.4: (a) Benchmark cricket clip, (b) Frequency centroid, (c) Bandwidth, (d) Energy ratio in subband 1, (e) Energy ratio in subband 2, (f) Energy ratio in subband 3.

3.1.4.3 Energy Ratio in Subbands (ERSBs)

Energy ratios denote the ratio of energy in various frequency subbands to the total energy. The division of bands is done such that each one of the band contains the same number of critical bands which correspond to cochlear filters in human auditory system. Typically, when the sampling frequency is 22050 Hz, the frequency ranges for four subbands are $0 - 630$ Hz, $630 - 1720$ Hz, $1720 - 4400$ Hz and $4400 - 11025$ Hz. The three energy ratios are sufficient as sum of energies in four subbands will always be equal to the total energy (leaving apart the computation errors involved due to non-ideal filters).

ERSBs for each frame in the window are computed. The corresponding window level features are computed as the weighted averages of ERSBs of all the frames in the window, weighted over volume of each frame. The ERSB plots for the benchmark clip are shown in Fig. 3.4 (d) - (f).

3.2 HIDDEN MARKOV MODELS FOR AUDIO CLIP CLASSIFICATION

Once data is represented in terms of features, it can be used to build a model for carrying out the classification. In very crude terms, a rule-based classifier can simply use thresholds on few or all features in the feature set in order to decide the category of the given audio clip. This may not work as in tasks like audio clip classification where, the classes are often overlapping in the feature space. Therefore, instead of using the features directly, it is attempted to capture their temporal behavior (i.e., how the feature vectors changes in a clip for a given audio category) to make the decision. It can be noted that it is almost impossible to manually determine the audio category of a clip when the only a short duration is present or the frames of the clip are randomly shuffled. Therefore, in order to capture the temporal behavior of the feature vectors, Hidden Markov model (HMM) is used in the baseline system.

HMM belongs to the class of statistical signal models in which one tries to characterize only the statistical properties of the signal. The underlying assumption of a statistical model is that the signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined (estimated) in a precise, well-defined manner. A description of HMM is given in Appendix B.

Ergodic HMMs with 5-states are used for carrying out the audio clip classification task. Here first and last states are simply entry and exit states, respectively. So practically, the model has only three states. The amount of data required for training an HMM increases exponentially with number of states. So the choice of having only three states was made due to the limited amount of data that we had for training. For final classification, the maximum likelihood decision logic is used.

3.3 DATABASE USED FOR STUDY

The audio data used for the classification experiments was collected from TV programs broadcast in India. The TV channels used for data collection are as follows:

(a) Advertisement - Across the TV channels

(b) Cartoon - Cartoon network, STAR and Zee TV

(c) Cricket - DD sports, ESPN, STAR sports

(d) Football - DD sports, ESPN, STAR sports

(e) News - Indian news TV channels (like Sun news, DD), in closed studio environment, by Indian news readers in Hindi, English, Tamil, Telugu and Kannada languages

The two databases considered for the study are as follows:

(a) **DB1** - Data collected for each category from single TV channel, contains 200 test clips, 40 clips of each category.

(b) **DB2** - Data collected across all TV broadcast channels, contains total 1359 clips. Number of clips belonging to each category is as follows:

Advertisement - 226

Cartoon - 208

Cricket - 318

Football - 300

News - 306

DB1 has very less intra-class variability, while clips belonging to DB2 have a significant intra-class variability.

3.4 AUDIO CLIP CLASSIFICATION RESULTS USING BASELINE SYSTEM

The audio clip classification results of the system for the two databases are given in Table 3.1. Performance of the system is degraded for DB2 as compared to that of DB1. Performance of the system for database DB2 degrades drastically in the case of football category as compared to other audio categories. One reason for the same may be that football may be considered as most degraded speech, and performance of the system using spectral features is known to degrade with noisy data [87].

Table 3.1: Audio clip classification performance using system based on the spectral features.

Audio Class	Percentage of clips correctly classified	
	Database - DB1	Database - DB2
Advertisement	85.00	65.00
Cartoon	90.00	75.00
Cricket	90.00	65.00
Football	92.50	40.12
News	87.50	65.30
Average	89.00	62.08

3.5 SUMMARY

A baseline system using spectral (segmental) features is discussed in this chapter. The category of the given audio clip can also be decided using the information present at subsegmental as well as suprasegmental level of the audio signal. The importance of this excitation source information present in the audio signal is not exploited in the existing system. The extraction of audio-specific subsegmental information present in the LP residual of the signal is discussed in the following chapter.

CHAPTER 4

EXTRACTION OF AUDIO-SPECIFIC SUBSEGMENTAL INFORMATION PRESENT IN THE LP RESIDUAL USING AANN MODELS

In the previous chapter, a baseline system for audio clip classification using spectral (segmental) features is discussed. The importance of excitation information present in the audio signal is not exploited in the existing approaches for audio indexing. In this chapter, a new method for capturing audio-specific information, using source characteristics of audio is introduced.

This chapter is organized as follows: Section 4.1 introduces the audio production mechanism. Sound is produced by vibrating sources. A method to extract the source characteristics from the signal is by using the LP analysis. Section 4.2 discusses the source characteristics for audio clip classification task. Section 4.3 discusses the extraction of the LP residual from the signal, and discusses the presence of audio-specific information in the LP residual signal. Section 4.4 defines the AANN models, and discusses the ability of the AANN models to capture the audio-specific information from the residual of the audio signal. Section 4.5 discusses audio as a combination of one or more components like music and speech. Section 4.6 discusses the use of the AANN models for capturing the knowledge of audio components present in the signal, that could be used for audio clip classification. The work flow of the system is discussed in Section 4.6. Section 4.7 summarizes the chapter.

4.1 SOUND PRODUCTION MECHANISM

A signal is defined as any physical quantity that varies with time, space or any independent variable(s) [88, 89]. Audio or sound is also a signal for which the functional relation between the signal and independent variable is highly complicated. To a high degree of accuracy, a segment of audio may be represented as a sum of several sinusoids of different amplitude and frequency. A sinusoid is generated as a result of a driving force to the physical system. Hence, the production of sinusoid depends upon the driving force, system and on the fact how tightly the two elements are coupled together. From signal processing perspective the force and system can be replaced by excitation and filter representation respectively. Here, the physical system is replaced with filter and the filter coefficients represent the physical dimensions of the system.

To a large extent, the mechanism of excitation source generation is independent of the filtering process, i.e., the properties of source do not tend to be strongly influenced by the acoustic properties of the filters. Thus, it is appropriate to discuss source mechanism separately from the behavior of the system in response to the sources. To an approximation, source properties can be considered to be independent of the system. In other words, audio signal consists of two kinds of information, source information and system information. It is by the interaction of these two kinds of information, that the audio signal is generated.

Having known these facts about the audio signal, in the next section, the usefulness of the source information for the audio clip classification task is discussed.

4.2 SOURCE CHARACTERISTICS FOR AUDIO CLIP CLASSIFICATION

Most of the existing audio clip classification systems have been developed using spectral features for capturing the audio-specific information. However, as discussed in the previous section, the audio signal carries information regarding both, system as well as source. Hence, the source information can also be used to capture the audio-specific information present in the audio signal of the given category. It can be shown that the residual signal, obtained after removing the spectral information from the audio signal, has perceptually significant audio-specific information [31]. Hence, the residual of the signal could be useful in the audio clip classification task. Not much effort has gone in developing audio clip classification system using this information due to practical difficulties involved in extracting this information using existing statistical and signal processing techniques. In the present work, the LP residual signal is used to derive the source characteristics of the audio, and neural network models are used to capture audio-specific information from the residual signal. The significance of the LP residual for audio clip classification task is discussed in the next section.

4.3 SIGNIFICANCE OF THE LP RESIDUAL FOR AUDIO CLIP CLASSIFICATION

4.3.1 The Linear Prediction Residual

The LP residual is extracted from the audio signal using the linear prediction analysis [32]. In the LP analysis each sample is predicted as a linear weighted sum of the past p samples, where p represents the order for prediction.

If the present sample $s(n)$ is predicted by the past p samples, then the predicted value $\hat{s}(n)$ is given by,

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{4.1}$$

where $\hat{s}(n)$ is the predicted sample and $\{a_k\}$ are the linear prediction coefficients. The difference between the actual, and predictable sample value is termed as prediction error or residual $e(n)$, which is given by,

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{4.2}$$

The linear prediction coefficients $\{a_k\}$ are determined by minimizing the mean squared error over an analysis frame. A detailed description of the LP analysis is given in Appendix A. It has been known that the LP order (p) used for extracting the residual plays a crucial role on the performance of audio, speech and speaker recognition system [37]. The study shows that the optimal range of the LP order for the audio-specific studies is in the range 10 to 14 for an audio signal sampled at 8 kHz.

4.3.2 Presence of Audio-Specific Source Characteristics in the LP Residual

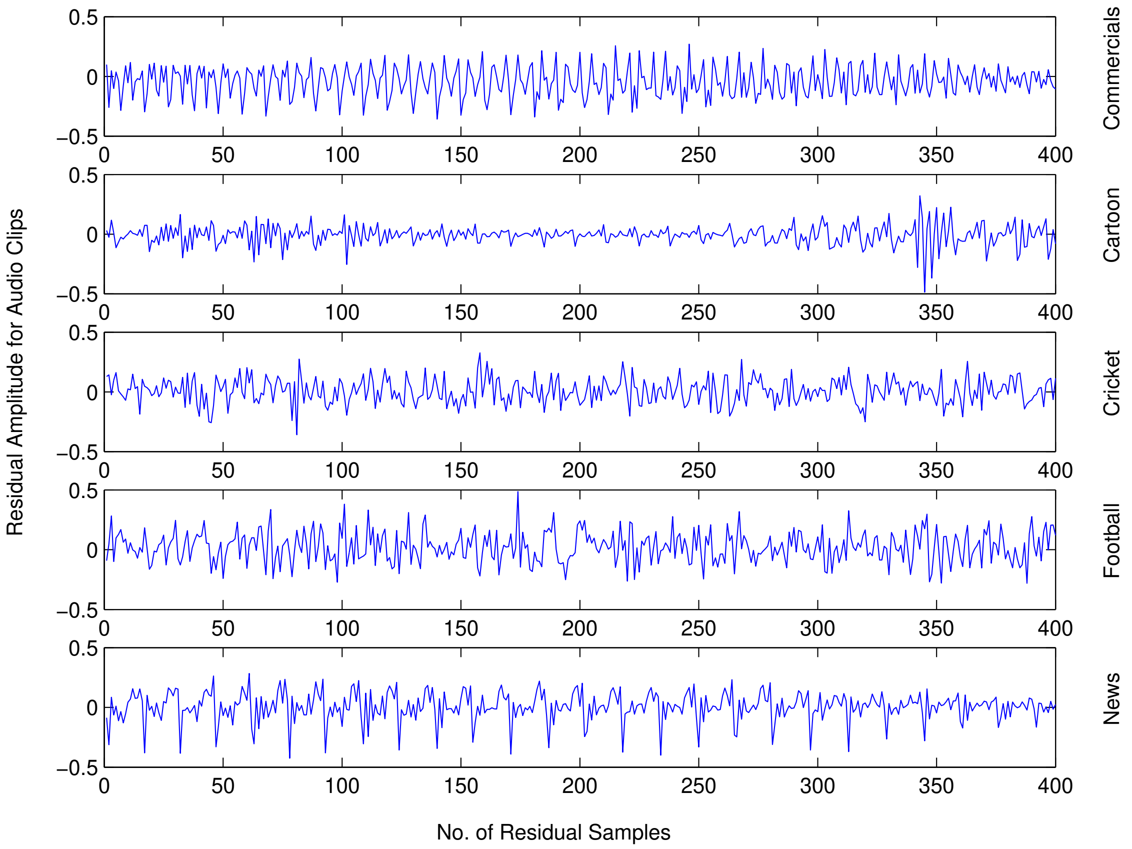


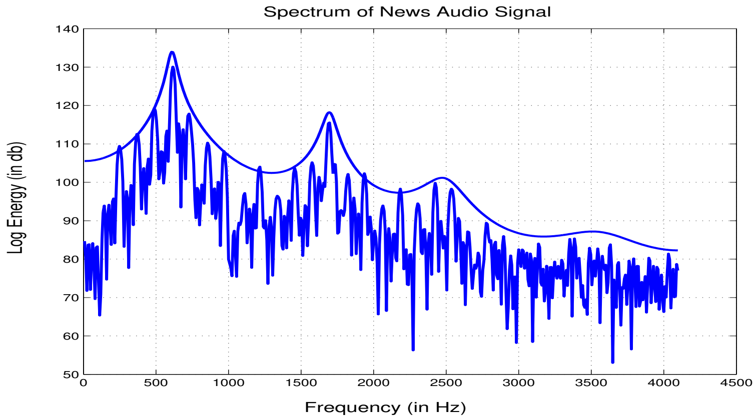
Fig. 4.1: The LP residual for segments of audio clips belonging to five different categories.

The five classes considered for the present study are advertisement/commercials, cartoon, cricket, football and news. They show variations among them. News audio (in closed studio environment) has clean speech, while speech for cartoon category differs from the news speech in terms of prosody. Music is a part of cartoon audio. Cricket and football have casual speech and other background sounds, like noise due to crowd, clapping and shouting etc. The noise is high in the case of football audio.

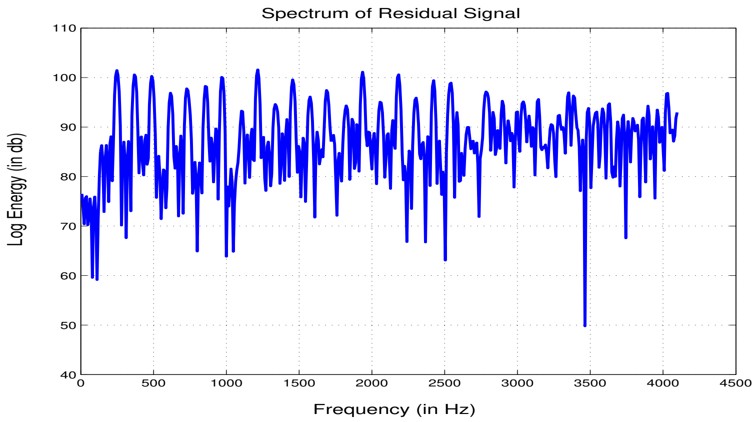
The advertisement audio is the most difficult class to study, as it has many variations within it.

The LP residual of the audio signals of the five categories considered for this study is extracted using a 12^{th} order LP analysis. The residual signals of the five different categories are shown in Fig. 4.1. It can be seen that the LP residual of different categories appear to be different. Even if the difference is subtle, the audio-specific information can be perceived while listening to the residual signal.

It is a known fact that the spectrum-based features, such as the linear prediction cepstral coefficients (LPCCs), capture the gross level information. However, the residual signal has much flatter spectral envelope (like white noise), representing the source characteristics rather than those of the system. The spectrum of the signal and the residual for a speech segment are as shown in the Fig. 4.2. Spectral representation of the residual signal might not yield complete information present in the residual. Since LP analysis extracts the second order statistical features through the autocorrelation functions, the LP residual does not contain any significant second order correlations of the audio production system. However, the excitation source characteristics are present in the LP residual. It is conjectured that the audio features may be present in the higher order relations among the samples of the residual signal. Since specific set of parameters to represent the audio information in the LP residual is not clear, and also since the extraction of such an information may involve nonlinear processing, autoassociative neural network models are proposed to capture the audio-specific information from the LP residual [41].



(a)



(b)

Fig. 4.2: (a) Spectrum of news audio signal. (b) Spectrum of residual signal.

4.4 THE AANN MODELS FOR CAPTURING AUDIO-SPECIFIC INFORMATION IN THE LP RESIDUAL

Artificial neural networks are useful for pattern recognition [41–43, 45, 90–92], because

(a) Neural network can mimic human brain.

(b) Neural network can generalize.

(c) Neural network is fault tolerant.

(d) Neural network infers solution from data without prior knowledge of regularities in the data, by extracting the regularities empirically.

(e) Neural network can be used for pattern recognition task with small sample size.

Classification and discrimination are used in a number of closely related problems [45, 91].

The AANN models are feedforward neural network (FFNN) performing an identity mapping of the input space [93], and hence the name autoassociation [42, 94]. FFNNs can be used for classification and mapping tasks [41, 95–97]. A description of FFNN is given in Appendix C. The AANN consists of an input layer, an output layer and one or more hidden layers. The number of units in the input and output layers are equal to the size of input vectors. The number of nodes in the middle hidden layer is less than the number of units in the input and output layers, and this layer is called dimension compression layer. The activation function of the units in the input and output layers is linear, whereas the activation function of the units in the hidden layers can be either linear and nonlinear.

The AANN models capture the distribution of the input spectral feature vectors in the feature space effectively [94, 98]. However, when an AANN is presented with raw data, such as samples of speech or the LP residual, the interpretation of the behavior of the AANN in terms of capturing the distribution of feature vectors is not appropriate. This is because the blocks of residual samples used as input to the network are not feature vectors. The adjacent overlapping blocks of the residual samples may have similar features, corresponding to some nonlinear (higher order) relations among the samples. The AANN trained with speech signal, captures the correlations present among the samples of the speech signal [83]. The AANN models have also been shown to capture excitation source features specific to the speaker in the LP residual [35, 38].

Speech signal contains both the second (autocorrelation) and higher order relations among the samples. If the speech signal is given directly to the network, then the second order correlations present among the samples, called system features, dominate the training of the network. If the second order correlations are removed through the LP analysis, then the network is expected to capture the implicit higher order relations in the LP residual signal. These relations may correspond to the desired audio-specific features in the excitation source.

For capturing the audio information present in the LP residual signal, a five layer AANN model as shown in Fig. 4.3 is used. The structure of the network used in this study is $40L$ $48N$ $12N$ $48N$ $40L$, where L refers to linear units, and N to nonlinear units. A hyperbolic tangent function ($tanh$) is used as the nonlinear activation function. When samples of random noise are given as input to the network, the network weights do not converge. This indicates that the network converges only when there are some relations among the samples of input signal. The structure of the network is derived using speech knowledge. The size of the input layer is kept 40, as if the length of the block exceeds a pitch period, the effect of the pitch period may also influence the training of the network. The size of the network defined is found to be suitable

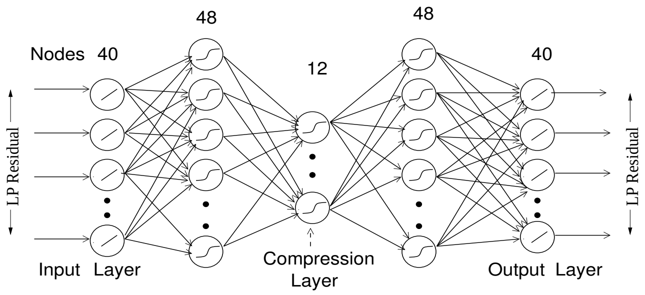


Fig. 4.3: Structure of the AANN model used for capturing audio-specific information.

for audio categories considered for study. However, it can vary for different types of audio signals, as per duration of the signal representing subsegmental information. In the next section, the audio components used to train the AANN models are defined.

4.5 KNOWLEDGE OF AUDIO COMPONENTS FOR DEFINING AUDIO CATEGORY

The audio signal for a given audio category is a combination of various audio components like speech, music and noise[1]. Alternatively, an audio category is composed of one or more audio components, which is unique to that particular audio category. This unique combination of the audio components differentiates audio of one category from that of the other. Hence, given an audio category, it is possible to name audio components that the category is composed of. Also, given an audio component, one can name the audio categories, it may belong to.

Using the knowledge of the audio components of the five categories considered for this study, each category could be described as follows:

[1]Here the term noise refers to nonspeech and nonmusic sounds, like clapping sound.

(a) *Advertisement* category is mainly composed of music and speech. There may be a few additional components like sound of a phone ringing, or sound of a car engine (depending on product the advertisement is about).

(b) *Cartoon* category is composed of music and speech. Prosody of cartoon speech is different from that of normal speech.

(c) Audio of *Cricket* category has speech and noise as components. Here noise is not white noise. The term noise is used for nonmusic and nonspeech components. Noise of cricket category has sounds like clapping, shouting and sound of ball hitting the bat. Speech component in the case of cricket audio is of casual conversation type, which includes laughing sound etc.

(d) *Football* category is also composed of speech and noise component. However, the noise component in the case of football is different from the noise in the cricket category. Here the level of noise is very high. The noise is almost continuous as opposed to a few calm regions in case of cricket category. Football noise does not have the sound of the ball hitting the bat, which is typical to the cricket category, but it has the sound of players kicking the ball.

(e) *News* category has clean speech. There is no music or noise component in news audio. Speech in the news audio has no laughing sounds, and is different from casual conversation. It has no long pauses, or filler sounds like *'ummmm'* which are present in casual conversation.

Therefore each of the five audio categories considered for this study could be defined as a unique combination of the following six audio components:

(a) Music (M)

(b) Conversational or casual speech (S1)

(c) Cartoon speech (S2)

(d) Clean speech (S3)

(e) Noise of cricket sport (N1)

(f) Noise of football sport (N2)

The following section discusses the use of component knowledge for audio clip classification.

4.6 AUDIO CLIP CLASSIFICATION STUDIES

4.6.1 Building AANN Models for Audio Components

As discussed in previous section, in an audio category, there could be one or more audio components. The knowledge of these components present in the audio is used for classification task. One AANN model is built for each of the six audio components.

For building component AANN models, data has been collected for 30 to 60 sec duration for each of the six components. The signal is sampled at 8 kHz, and is stored as 16 bit integers. Various parameters for study have been decided experimentally. The LP residual is extracted using 12^{th} order LP analysis, and the residual is normalized to unit magnitude before giving as input to the AANN models. Residual

4.6.2 AANN Models for Capturing Subsegmental Features

During classification, a test clip of 10 sec duration is used. Test data is processed in the same way as for the training data. Blocks of 40 samples of the LP residual are presented with one sample shift to the model. The output of each model is compared with its input to compute the squared error for each block. The error E_i for the i^{th} block is transformed into a confidence value C_i using C_i = exp(-λE_i), where the constant $\lambda = 1$.

In a test clip, there could be one or more audio components. The test clip is tested against all six component AANN models. It is hypothesized that the AANN models built using same components that are present in the test audio clip will give higher confidence scores as compared to other AANN models. For example, if the test clip belongs to news category, all speech component models are expected to give higher confidence scores, and music and noise models to give lower confidence score values. Again among speech models, clean speech model is expected to give highest confidence score for the test clip belonging to news category.

The average confidence scores (over a window of 10 samples, with every sample shift) of all the six component models for a segments of news test clip are shown in Fig. 4.5. As shown in the Fig. 4.5, for a test clip of news category, the values of the confidence score of the speech component AANN models are higher as compared to the values of the confidence score of noise and music component models. The clean speech component model can be seen to give the highest confidence scores.

egory of the test clip using the relation among confidence score values of the six AANN models that correspond to each component. Hence, a multilayer feedforward neural network (MLFFNN) model or multilayer perceptron (MLP) is proposed to decide the audio category using the pattern in the confidence scores. The MLP is found to be suitable for pattern classification tasks because of its ability to form complex decision surface using discriminating learning algorithms [41]. The values of the confidence scores of the six component AANN models are considered as input feature vector (component confidence vector (CCV)) for the MLP. The MLP for decision making about the category of a test audio clip using CCV is discussed in next chapter.

4.6.3 Work-flow of the System

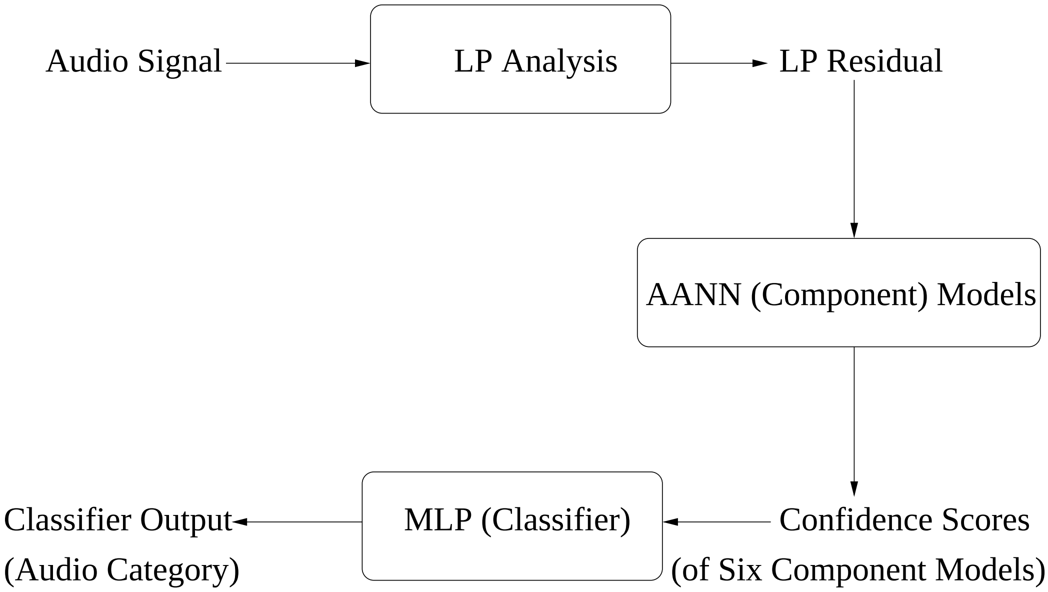


Fig. 4.7: Work-flow of the audio clip classification system using subsegmental information present in the audio signal.

Fig. 4.7 gives an overview of the work-flow of the audio clip classification system using subsegmental information in the audio signal. For a given audio test clip, the LP residual signal is extracted using LP analysis. The residual signal is given to each of the six component AANN models. Frame-wise smoothed values of the confidence scores for the six components AANN models are considered as input feature vector to the MLP classifier, which assigns a label (name of audio category) to the test audio clip. The MLP used for this task is discussed in following chapter.

4.7 SUMMARY

In this chapter, the significance of source characteristics for audio clip classification task has been discussed. The use of the LP residual to represent the excitation source information was demonstrated. Each audio category is defined using audio components. The AANN models are shown to capture the audio-specific excitation source information present in the residual signal. The next chapter discusses the use of the MLP for decision making task using the component confidence scores given by the AANN models. Experimental results and complementary nature of the source and system information present in the audio signal will also be discussed in the next chapter.

CHAPTER 5

AUDIO CLIP CLASSIFICATION USING SUBSEGMENTAL AND SEGMENTAL FEATURES

Previous chapter discusses the relation between an audio category and audio components such as music, speech etc. Each audio category is composed of one or more audio components in certain proportion. The audio-specific subsegmental features in the LP residual signal can be captured by the AANN models. There is a pattern in the output confidence scores of the six component AANN models, for the five audio categories. A multilayer perceptron was proposed to decide the category of a given audio clip using this pattern. Section 5.1 describes the use of the MLP to decide the category of a given audio clip. Section 5.2 discusses the results of classification using subsegmental features. In Section 5.3, the complementary nature of the information present in the subsegmental features and segmental features is discussed. Methods of combining the evidences from the two classifiers are also discussed in Section 5.3. Section 5.4 summarizes the chapter.

5.1 THE MLP FOR AUDIO CLIP CLASSIFICATION USING SUBSEGMENTAL FEATURES

5.1.1 Structure of the MLP

Feature vectors of the audio clip of a particular audio category occupy a small region in the feature space. For audio clip classification, the system has to discriminate among the five categories considered for study. The feature vectors of different audio categories are likely to occupy different regions in the feature space.

A four layer MLP classifier having the architecture shown in Fig. 5.1 is considered. The four layer MLP is suitable for pattern classification task [41]. The structure of the MLP is $6L$ $24N$ $12N$ $5N$. The first layer, which is the input layer, has six nodes. The input is the component confidence vector (CCV), which is the confidence given by the six component AANN models. As shown in Fig. 5.1, the input to the first node is the output confidence score of the music component AANN model. Inputs to the second, third and fourth nodes are confidences given by the three speech component AANN models, and input to the fifth and sixth nodes are confidence scores obtained by the two noise component AANN models. The output layer has five nodes corresponding to five audio categories (A - advertisement, C - cartoon, K - cricket, F - football, N - news). The number of nodes in the hidden layers were decided based on experimental studies.

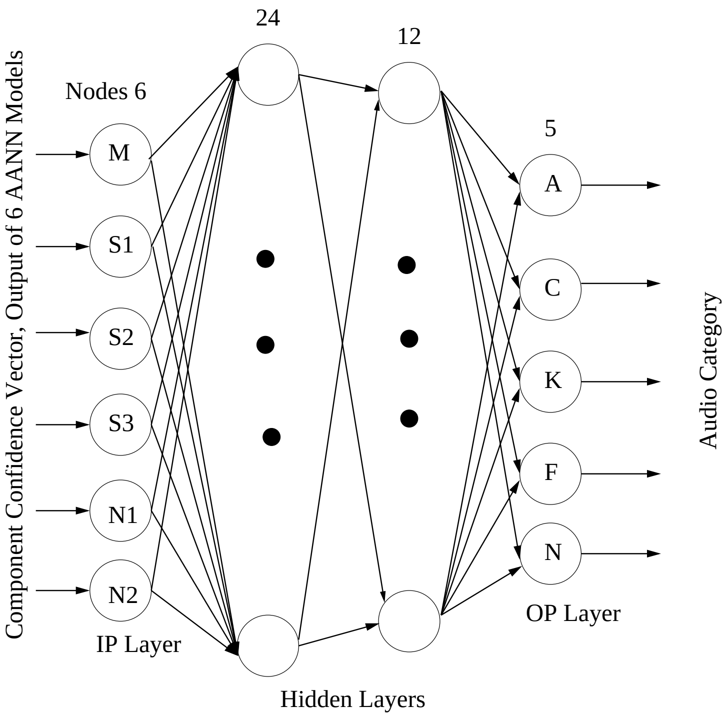


Fig. 5.1: Structure of the MLP used for audio clip classification task.

5.1.2 Training the MLP

The network is trained using backpropagation algorithm [92]. Backpropagation assumes supervised training. The objective of the training is to adjust the weights, so that application of the input feature vectors produces the desired set of outputs. A detailed description of the backpropagation training algorithm is given in Appendix D. The neural network classifier tries to capture the decision boundaries that separate the CCV of the five audio categories.

In order to train the MLP for audio clip classification, a training set has to be created consisting of input feature vectors and the corresponding output feature vectors representing the audio categories. The output vector for each audio category is defined as a five-dimensional vector, $O = \{o_1, o_2, o_3, o_4, o_5\}$, where o_i is set to 1 for CCVs belonging to the i^{th} audio class, and -1 for all other categories.

The training process is summarized as follows: The weights of the networks are initialized to small random values falling within the range $+1$ to -1 before starting the training [99]. This is done in order to prevent network from being driven to saturation [41,100]. To train the network, the training pattern consisting of input feature vector and the desired output vector are presented to the network repeatedly. The output vector of the network is compared with the desired output vector, and the error is fed backward from the output layer to the input layer. The interconnection weights are adjusted using the backpropagation algorithm to minimize the error of the output layer. The rate at which the weights are allowed to change at any given presentation is governed by the amount of error and the learning rate. Higher learning rate speeds up the convergence process, but it sometimes may result in overshooting or oscillation which may lead to nonconvergence of the network. Slower learning rate produce more reliable results at the expense of increased training time [100]. The training of the network is stopped after 2200 epochs, which was decided experimentally.

5.1.3 Testing the MLP

The test pattern is presented to the network. The classification decision is based on winner-take-all principle. In other words, during testing, the averages of the output vectors of the MLP for all the test CCVs are computed. The test audio clip is assigned to the class that corresponds to the output node that has the maximum value of the average output vector.

5.2 CLASSIFICATION RESULTS USING SUBSEGMENTAL FEATURES

Audio clip classification results for databases (DB1, DB2) are given in Table 5.1. The performance of the system for an audio category is computed as percentage of clips correctly classified, which is the ratio of the number of test audio clips of the category correctly classified to the total number of test audio clips of that category. The average performance of the system for DB1 is 86.60%. When the larger database DB2 is considered, the performance of the system comes down. Inspite of relatively lower performance for DB2, it can be seen that the audio-specific information is still captured from the LP residual signal with neural network (NN) models used for classification task.

Table 5.1: Audio clip classification performance using system based on the LP residual.

Audio Class	Percentage of clips correctly classified	
	Database - DB1	Database - DB2
Advertisement	83.00	43.50
Cartoon	88.00	45.50
Cricket	86.00	38.50
Football	90.50	72.80
News	85.50	63.30
Average	86.60	52.72

It can be noticed that degradation in performance is less in the case of news and football categories, as compared to other three audio categories. This can be due to the fact that in DB2, the variability in audio data is less in the case of news and football, as compared to the other three categories. Advertisement and cartoon clips have music, speech and other audio components. However, the type of music (depending on the musical instruments used), speech and other audio components vary from one audio clip to other clip. There is a significant intra-class variability in the audio data of the advertisement and cartoon category. Hence in the general case of advertisement and cartoon audio, the duration of the segment to be considered for capturing the subsegmental information may vary from one clip to other. Degradation in performance in the case of cricket audio is also because of the variability in the data used for testing. There is significant variation in cricket audio collected from different TV channels. Cricket audio data collected from DD sports channel has high level of noise, while noise is relatively less in the case of data collected from the ESPN channel. Also, speaking style, accent of commentator varies a lot in the case of cricket data collected across the channels. Hence, the NN models, trained using the data collected from one TV channel, may not work well for the test data collected from other TV channels. Large variability in the data makes the training of the network insufficient, due to significant overlap (of same type of patterns) among categories. Sometimes, decision making for the same is difficult for humans also without having complete domain knowledge. Hence, (1) other type of information present in the signal also needed to be considered to get a better picture, (2) as performance of the system degrades mainly because of the large variability in the test data, each of these categories can further be divided into subcategories (like cricket from DD sports and cricket from ESPN) with less intra-class variability in each subcategory. The problem can be addressed at a more finer level and a modular approach can be used for the classification task.

The subsegmental information present in the audio signal can be complementary in nature to the segmental information. The subsegmental audio-specific information is captured by using the system based on the LP residual, NN models and the segmental information is captured using the baseline system. In the following section, methods to combine the two levels of information present in the audio signal are discussed.

5.3 COMBINING EVIDENCES FROM THE TWO SYSTEMS

The complementary nature of the subsegmental and segmental information, present in the audio signal, can be explained using Fig. 5.2. Rectangle represents the set of all test audio clips. System based on spectral features and the LP residual are able to classify the audio clips as shown (using ellipse) in Fig. 5.2. Overlapping area in the ellipses shows the audio clips which are classified by using both the systems. The nonoverlapping area in these ellipses shows the audio clips which are classified by using either systems (not both). Hence, if evidence from both the systems are combined, a better performing system could be obtained. The following subsection discusses method of combining evidences from multiple classifiers.

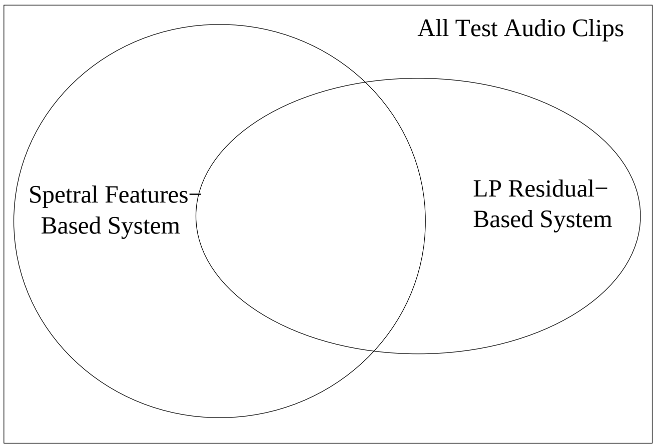


Fig. 5.2: Venn diagram showing the complementary nature of the two systems.

5.3.1 Multiple Classifiers and Combining Evidence

Traditional pattern recognition systems use a single feature descriptor and a particular classification technique. However, it is very difficult to achieve good result when there is large variability in data and noisy inputs. Recent studies [101–103] have led to a belief that by using features and classifiers of different types simultaneously, the classification performance of the system could be improved. One of the major reason for combining multiple classifiers is that for a specific pattern recognition problem, usually numerous types of features could be used to represent and recognize the pattern (in this case segmental and subsegmental). Different features are from different sources of information, and are represented in diversified forms. Therefore it is very difficult to put them together within a single classifier to make a decision. As a result there is a need for many classifiers to handle different types of features. But combining multiple classifiers is by no means the only solution for the problem involving variety of features. There are systems based on statistical and syntactical methods for this purpose [104].

Let Λ be a set of class labels $\{1, 2, \cdots, M\}$ and $c_1, c_2, \cdots, c_K$ be K classifiers. Generally the output information that various classification algorithms supply or are able to supply can be divided into three levels [101,102]:

(a) Abstract Level:

A classifier c_k gives only a label $j \in \Lambda$ as the output.

(b) Rank Level:

A classifier c_k ranks all the labels in Λ, and the label with first rank is considered as the first choice.

(c) Measurement Level:

A classifier c_k attributes each label j in Λ a measurement value to address the degree that the feature vector $\mathbf{x}$ is associated with the label j.

There are three types of problems for multiple classifier combination based upon the aforementioned three output information levels. They are summarized as follows:

(a) Type I: The combination is made based on the output information at the abstract level. Each of the K classifiers c_k assigns input $\mathbf{x}$ a label $j_k \in \Lambda$. The problem is to use these evidences to build an integrated classifier C, which gives input $\mathbf{x}$ a definite label j, where $j \in \Lambda$ [105].

(b) Type II: The combination is made based upon the output information at the rank level. For an input $\mathbf{x}$, each c_k produces a subset $L_k \subseteq \Lambda$ with all the labels in L_k ranked in a queue. The problem is to use these evidences $c_k = L_k$, $k = 1, 2 \cdots, K$ to build a classifier C such that $C(\mathbf{x}) = j$, where $j \in \Lambda$ [106].

(c) Type III: The combination is made based on the output information at the measurement level. For an input $\mathbf{x}$, each c_k produces a real vector $m_{c_k} = [m_k(1), \cdots m_k(M)]^t$ (where $m_k(i)$ denotes the degree that c_k considers that $\mathbf{x}$ has label i). The problem is to use these evidences $c_k = m_{c_k}$, $k = 1, 2, \cdots, K$ to build a classifier C with $C(\mathbf{x}) = j$, where $j \in \Lambda$.

For problem of Type I the classifiers could be different from each other in their approaches, whereas Type III requires that all the individual classifiers should be able to supply the output information at measurement level with the same scale. Type II problem requires that all classifiers should be able to supply information at rank level. The combining function basically should take advantage of the strengths of individual classifiers and avoid their weakness and improve classification accuracy.

Following subsections discuss the methods for combining the two audio clip classification systems considered in this study.

5.3.2 Abstract Level Combination of Two Systems

In present study, the information is first combined at abstract level to check the complementary nature of subsegmental and segmental information. At this level the only information considered from the two systems is the output class label. The study has only two systems, and it is difficult to go for voting or weighing the output label of one system. Hence, the output class labels are combined using *OR* logic. This is

Table 5.2: Audio clip classification performance for combined (at abstract level) system for DB2.

Audio Class	Percentage of clips correctly classified		
	Spectral features-based system	LP residual-based system	Abstract level combination
Advertisement	65.00	43.50	83.00
Cartoon	75.00	45.50	92.00
Cricket	65.00	38.50	87.50
Football	40.12	72.80	86.00
News	65.30	63.30	86.30
Average	62.08	52.72	86.96

similar to considering the whole area under two ellipses in Fig. 5.2. Hence, if either of the systems is able to classify the clip, it is considered as correct classification, and misclassification is considered only when both systems fail. Since the objective of this study is to get the evidence from other system when one system fails, the *OR* logic is considered. Abstract level classification results are given in Table 5.2.

5.3.3 Rank and Measurement Level Combination of the Two Systems

Abstract level combination shows the complementary nature of subsegmental and segmental information for audio clip classification task. However, the abstract level combination method uses least information given by the two systems, it uses only class label given by the system. It does not use any information about the confidence with which system gives the class label. Hence, a method using rank and measurement level combination is proposed for the task. The method is summarized as follows:

(a) Normalize output confidence values of the two systems in the range of 0 to 1.

(b) Add class-wise confidence values of the two systems.

(c) Select the maximum of the class-wise sum of confidence values.

(d) The category corresponding to this maximum confidence value (1^{st} rank) is assigned to the test audio clip.

Table 5.3: Audio clip classification performance for combined (at rank and measurement level) system for DB2.

Audio Class	Percentage of clips correctly classified		
	Spectral features-based system	LP residual-based system	Rank and measurement level combination
Advertisement	65.00	43.50	92.47
Cartoon	75.00	45.50	98.55
Cricket	65.00	38.50	88.67
Football	40.12	72.80	90.10
News	65.30	63.30	95.10
Average	62.08	52.72	92.97

Classification results using rank and measurement level combination of the two systems are given by Table 5.3. This method outperforms simple abstract level combination (as given in Table 5.2). Reason for that can be explained with the help of following example. Let the class-wise normalized output confidence values for the two systems and its sum be given in Table 5.4. It could be seen that as per the first system the category of the clip is 'Advertisement', while as per second system it is 'Cricket'. This is an example when both systems are misclassifying the test clip. However, on adding the confidence values, the winning class is 'Cartoon', which is the actual category of test clip. Hence hidden performance, when both systems are failing, is captured by this method.

Analysis of confidence values shows that if both models are giving highest confidence for the same category name, it is selected, and misclassification is also accepted in this case. However, when both models are not agreeing, second-best confidence values of the systems play a role in deciding the system classifying the clip correctly. Also, it helps in bringing this hidden performance out as explained above using Table 5.4. Best performance is obtained by giving equal weight to both systems.

Table 5.4: An example of class-wise normalized output confidence values of the two systems and its sum.

Category name ⇒	Advertisement	Cartoon	Cricket	Football	News
Confidence values of first system ⇒	1	0.8	0.5	0.4	0
Confidence values of second system ⇒	0.2	0.9	1	0.3	0
Sum of Confidence values of two systems ⇒	1.2	1.7	1.5	0.7	0

5.4 SUMMARY

The MLP have been shown to be suitable for audio clip classification task using audio-specific subsegmental information captured by the AANN models from the LP residual signal. It was shown that the information captured by the system based on the

LP residual and the system based on spectral features are complementary in nature. Methods to combine the two levels of information present in the audio signal have been proposed. The combined system outperforms individual systems. Apart from subsegmental and segmental information, suprasegmental information is also present in the audio signal, which could also be useful for the task. The following chapter discusses the presence of audio-specific suprasegmental information in the LP residual signal.

CHAPTER 6

SUPRASEGMENTAL FEATURES FOR AUDIO CLIP CLASSIFICATION

The presence of audio-specific complementary information in subsegmental and segmental features for audio clip classification was discussed in the previous chapters. The other level of information contained in the audio signal is the suprasegmental information, that is the variation of the signal features over long duration. This behavioral characteristics of audio is perceived in the LP residual signal also. This chapter discusses the presence of suprasegmental information in the LP residual signal for audio clip classification task. Section 6.1 discusses the presence of suprasegmental information in the LP residual signal. The excitation peaks that can be noticed in the LP residual signal are seen clearly in the Hilbert envelope [86] of the LP residual. Section 6.2 discusses the Hilbert envelope, and discusses the presence of audio-specific suprasegmental information in the Hilbert envelope. Methods to extract suprasegmental information from the Hilbert envelope are discussed in Section 6.3. This section also discusses the utility of the suprasegmental information for the audio clip classification task. Section 6.4 summarizes the chapter.

6.1 SUPRASEGMENTAL FEATURES IN THE LP RESIDUAL SIGNAL

Behavioral characteristics of one audio category differ from that of the other. The behavioral characteristics can also be perceived while listening to the LP residual signals of various audio categories. In Fig 6.1, the LP residual signals for five audio categories are shown. In Fig 6.1, one may notice some differences in the patterns in the residual signals of five audio categories. Sometimes this difference may not be noticed in the waveform, but could be perceived while listening to the residual signal. Hence, it is worthwhile to explore these features for audio clip classification task.

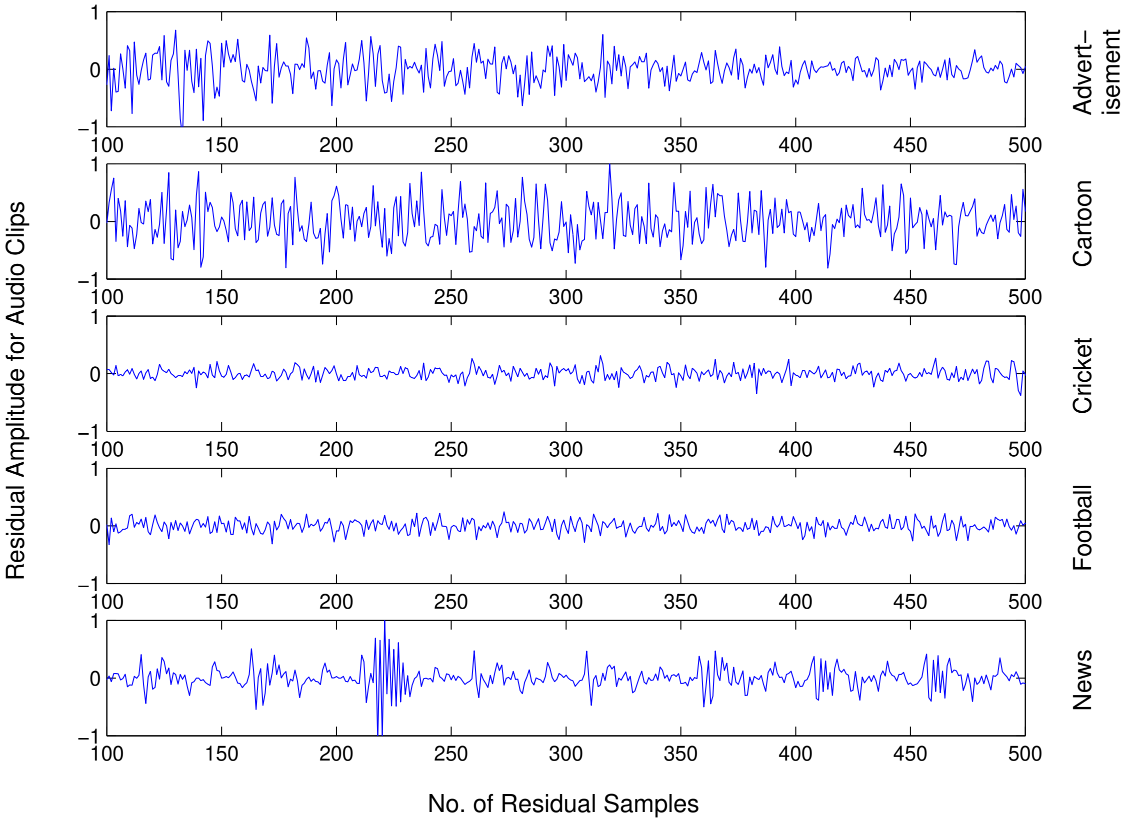


Fig. 6.1: The LP residual for the segments of audio clips belonging to five audio categories.

Patterns in the LP residual signal are in form of sequence of excitation peaks. These excitation peaks can be considered as event markers. The sequence of these events contain important perceptual information about the source of excitation and behavioral characteristics of audio. By listening to the residuals of different types of audio clips, one can distinguish between speakers, music, instruments, etc. Excitation peaks can further be enhanced by taking the Hilbert envelope of residual signal. The Hilbert envelope computation removes the phase information present in the residual, thereby leading to better identification of the excitation peaks. Following section discusses a method to calculate the Hilbert envelope of the LP residual signal. The patterns in the Hilbert envelopes of the five audio categories can also be seen.

6.2 THE HILBERT ENVELOPE OF THE LP RESIDUAL SIGNAL

6.2.1 Computation of the Hilbert Envelope from Residual Signal

The residual signal is used to compute the Hilbert envelope, where the excitation peaks show up prominently. The Hilbert envelope is defined as,

$$h_e(n) = \sqrt{e^2(n) + h^2(n)} \tag{6.1}$$

where $h_e(n)$ is the Hilbert envelope, $e(n)$ is the LP residual and $h(n)$ is the Hilbert transform of the residual. The Hilbert transform of a signal is the 90^0 phase shifted version of the original signal. Therefore, the Hilbert envelope represents the magnitude of the analytic signal,

$$x(n) = e(n) + ih(n) \tag{6.2}$$

where $x(n)$ is the analytic signal, $e(n)$ is the residual and $h(n)$ is the Hilbert Transform of the residual.

The Hilbert envelope computation removes the phase information present in the residual. This leads to emphasis of the excitation peaks. The excitation peaks are further emphasized by using the neighborhood information of each sample in the Hilbert envelope. The modified Hilbert envelope is computed as,

$$h_{em}(n) = \frac{h_e^2(n)}{\sum_{k=n-l}^{k=n+l} \frac{h_e(k)}{(2l+1)}} \tag{6.3}$$

where $h_{em}(n)$ is the modified Hilbert envelope, $h_e(n)$ is the Hilbert envelope and l is the number of samples on either side of the neighborhood of current sample n.

6.2.2 Presence of Suprasegmental Features in the Hilbert Envelope

Fig. 6.2 shows the residual, the Hilbert envelope and the corresponding modified Hilbert envelope for a noisy audio (speech) segment. It can be noticed that the excitation peaks are clearly visible in the modified Hilbert envelope[1]. The gap between excitation peaks is the pitch period in case of speech signal, which varies for different audio signals. The categories considered for the study are combination of various audio components. Hence the patterns in the excitation peaks in the Hilbert envelope for different audio categories are also a combination of events of the audio components.

As shown in Fig. 6.3, the pattern over a longer duration of segment, in excitation peaks is different for different audio categories, hence it could be utilized for audio

[1] For this study modified Hilbert envelope is considered. So in following text whenever Hilbert envelope is mentioned, it actually refers to the modified Hilbert envelope.

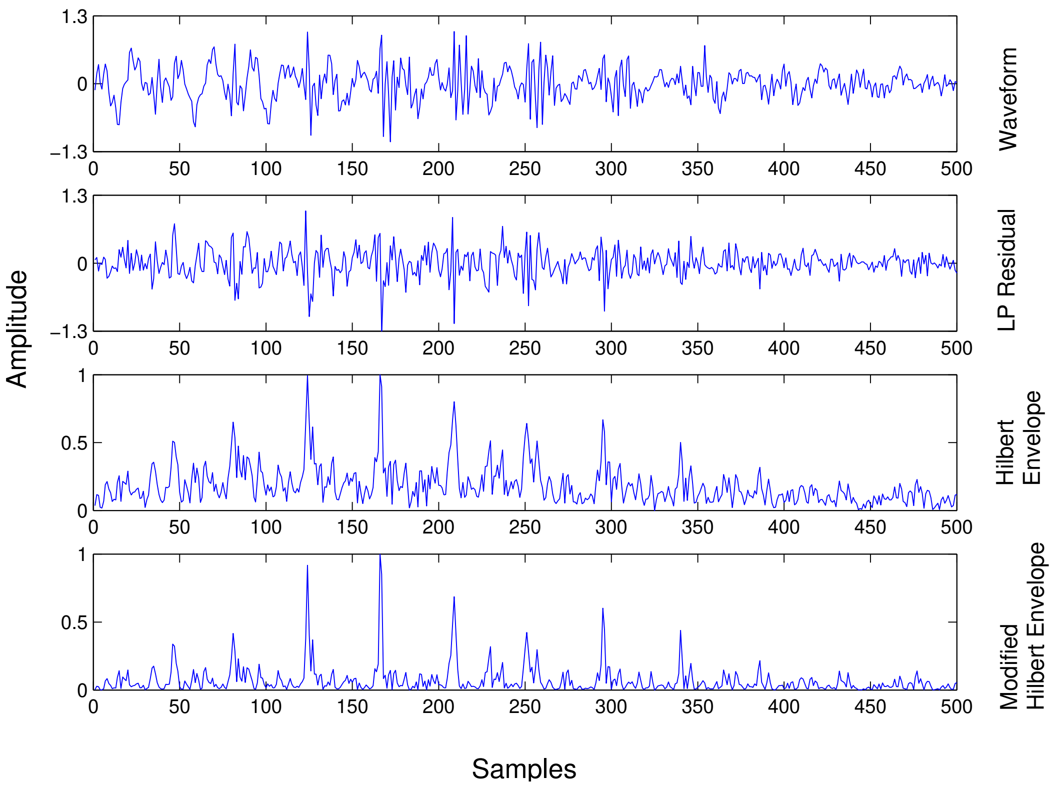


Fig. 6.2: The waveform, LP residual, Hilbert envelope, modified Hilbert envelope of the residual signal for a noisy audio segment.

clip classification task. The following section emphasizes the presence of audio-specific features in this pattern in the Hilbert envelope of audio signal.

6.3 SUPRASEGMENTAL FEATURES IN THE HILBERT ENVELOPE

As discussed in previous section, there is audio-specific information at the suprasegmental level in the LP residual signal, which can be perceived by listening to the signal, and it can also be observed in the Hilbert envelope of the LP residual signal, as shown in Fig. 6.3. One method to capture this pattern in the Hilbert envelope is by taking

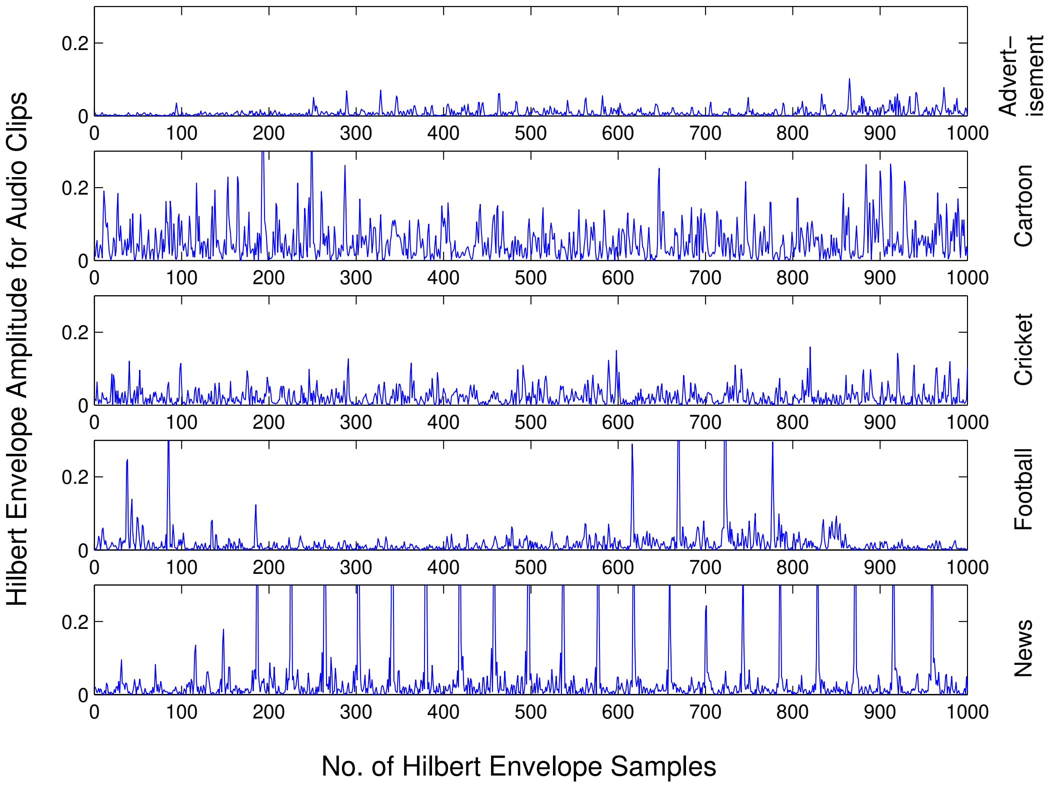


Fig. 6.3: The Hilbert envelope of the residual signal for the segments of audio clips belonging to five audio categories.

the autocorrelation of the Hilbert envelope. For a segment of 100 ms of the Hilbert envelope the autocorrelation sequence is calculated. The reason for choosing 100 ms window size for calculation of the autocorrelation is that the long term characteristics of the audio signal are of interest. The window is further shifted by 50 ms, and calculation of the autocorrelation is repeated till whole length of the audio clip is considered. These autocorrelation sequences (starting from 3^{rd} sample from the center peak to 400^{th} sample, normalized with respect to the central peak) are plotted in Fig. 6.4 for a clip for each of the five audio categories considered.

Mean, variance and skewness of autocorrelation sequences along frame sequence axis for each of 3^{rd} to 203^{th} sample are calculated. The variance of these statistics for a clip is plotted as a point in 3-dimension. These statistics of autocorrelation sequences of the Hilbert envelope is plotted for all 1359 test clips (belonging to DB2) in Fig. 6.5. It can be seen that clips belonging to different categories occupy different but overlapping regions in this plot. The overlapping area in the plot implies that the component knowledge in audio is observed to play a major role in this study also. The (parts of) audio clips having similar audio components have been observed to have the same kind of pattern in autocorrelation sequence, and occupying the overlapping area in 3-dimensional plot. However, inspite of the overlapping region, the clips belonging to different audio categories can be observed to occupy different regions in 3-dimensional plot, as in Fig. 6.5. It can be seen in Fig. 6.5 that the clips belonging to news category are lying in the right hand side region of the plot (shown using 'x'), and clips belonging to football category are lying in the central region of the plot (shown using 'o'). Clips of cartoon category are lying in the left hand side region of the plot (shown using '$\diamond$'), and clips of cricket category are lying in central most region of the plot (shown using '*'). A significant overlap in the plot can be noticed in case of cricket and football clips, inspite of a nonoverlapping area occupied by the clips of the two categories. Advertisement clips are lying in central-bottom most area in the plot (shown using '.'). This emphasizes the presence of audio-specific suprasegmental information in the LP residual signal. However, the use of the suprasegmental information as an additional evidence for audio clip classification task is still to be explored.

CHAPTER 7

SUMMARY AND CONCLUSIONS

In this thesis, a global view of an audio indexing system is introduced. Classification of audio into various categories is an important step for a practical audio indexing system to handle the large volume of nontext data. Further, there are three levels of information in the audio signal, namely, subsegmental, segmental and suprasegmental. The objective of this study was: (1) To extract perceptually significant audio-specific features from the residual signal for the audio clip classification task, (2) to use the information present at various levels of audio signal to classify it into one of the following predefined category. The classes considered for study were: Advertisement, cartoon, cricket, football and news. Five major studies conducted are as follows:

In first study, a baseline system using spectral (segmental) features and HMM models was built for the audio clip classification task.

The linear prediction analysis is used to the extract the residual of the signal. The second study was to extract the subsegmental features from the LP residual signal. Component knowledge was used to represent an audio category. The AANN models have been shown to be suitable to find the presence of a component in the given test audio signal, and hence, to capture audio-specific subsegmental features.

In third study, a MLP has been used for decision making task using the audio-specific subsegmental features captured by the AANN models from the LP residual. A four layer MLP was used for the study. A system using these NN models was presented. Performance of the system was evaluated using two databases collected from TV broadcast.

Information present at various levels of audio signal may be complementary in nature. Hence, in the fourth study, methods to combine the evidences from the system based on the LP residual and NN models, and the baseline system based on spectral features, were proposed. It was shown that combining the two systems leads to a better performance than individual systems.

The fifth study was conducted to show the presence of audio-specific suprasegmental features also in the LP residual signal. The pattern in the excitation peaks in the LP residual for different audio categories was enhanced by taking the Hilbert envelope of the residual signal. The statistics of the peak distribution in the autocorrelation sequences of the Hilbert envelopes were noticed to be different for the five audio categories. This study shows the presence of suprasegmental features in the LP residual signal, but how to use these feature for classifying audio is yet to be explored.

7.1 MAJOR CONTRIBUTIONS OF THE PRESENT WORK

(a) Extraction of audio-specific subsegmental information, from the LP residual of the audio signal using AANN and MLP models, by using the audio component knowledge has been presented for the audio clip classification.

(b) The complementary nature of subsegmental and segmental information present in the audio signal is studied for the audio clip classification task.

(c) The presence of audio-specific suprasegmental information in the LP residual signal has also been shown in the present work.

7.2 SCOPE FOR FUTURE WORK

(a) Various parameters, such as structure of the AANN models, have been derived based on speech knowledge. These parameters are to be studied carefully for the kind of audio and the application in hand.

(b) The present work has discussed the presence of audio-specific suprasegmental information in the LP residual signal. Methods to study the complementary nature of this information to the information captured by existing systems are to be explored. Methods to combine the three levels of information in the audio signal are to be studied to get a better performing system.

(c) The proposed framework can be extended for audio segmentation, classification of other audio categories and other audio indexing applications.

(d) The presented audio clip classification system can be extended to a multimedia indexing system.

APPENDIX A

LINEAR PREDICTION ANALYSIS

In LP analysis, an all-pole model is assumed for the system producing the signal $s(n)$. A p^{th} order all-pole model assumes that sample value at time n can be approximated by linear combination of past p samples. i.e.,

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{A.1}$$

If $\hat{s}(n)$ denotes the prediction made by the all-pole model then, the prediction error is given by,

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{A.2}$$

This error is nothing but the LP residual of the given signal.

For a frame of size m samples, the mean square of prediction error over the whole frame is given by,

$$E = \sum_{m} e^2(m) = \sum_{m} [s(m) - \sum_{k=1}^{p} a_k s(m-k)]^2 \tag{A.3}$$

Optimal predictor coefficients will minimize this mean square error. At minimum value of E,

$$\frac{\partial E}{\partial \mathbf{a_k}} = 0 \ , \qquad k = 1, 2, ..., p. \tag{A.4}$$

Differentiating Eqn. A.3 and equating to zero we get,

$$\mathbf{R}\ \mathbf{a} = \mathbf{r} \tag{A.5}$$

where, $\mathbf{a} = [a_1\ a_2 \cdots a_p]^T$, $\mathbf{r} = [r(1)\ r(2) \cdots r(p)]^T$, and $\mathbf{R}$ is a Toeplitz symmetric autocorrelation matrix given by,

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & r(p-1) \\ r(1) & r(0) & \cdots & r(p-2) \\ \vdots & & \ddots & \vdots \\ r(p-1) & & \cdots & r(0) \end{bmatrix} \tag{A.6}$$

Eqn. A.5 can be solved for prediction coefficients using Durbin's algorithm as follows:

$$E^{(0)} = r[0] \tag{A.7}$$

$$k_i = \frac{r[i] - \sum_{j=1}^{L-1} \alpha_j^{(i-1)} \cdot r[|i-j|]}{E^{(i-1)}} \qquad 1 \leq i \leq p \tag{A.8}$$

$$\alpha_i^i = k_i \tag{A.9}$$

$$\alpha_j^i = \alpha_j^{(i-1)} - k_i \cdot \alpha_{i-j}^{(i-1)} \tag{A.10}$$

$$E^{(i)} = (1 - k_i^2) \cdot E^{(i-1)} \tag{A.11}$$

The above set of equations are solved recursively for $i = 1, 2, ..., p$. The final solution is given by

$$a_m = \alpha_m^{(p)} \qquad 1 \le m \le p \tag{A.12}$$

where, a_m's are linear predictive coefficients (LPCs).

Cepstral coefficients can be extracted from the predictor coefficients using recursive algorithm as follows.

$$\begin{aligned} c_0 &= \ln \sigma^2 && \text{(A.13)} \\ c_m &= a_m + \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad 1 \le m \le p && \text{(A.14)} \\ &= \sum_{k=1}^{m-1} \frac{k}{m} \cdot c_k \cdot a_{m-k} \qquad\qquad m > p && \text{(A.15)} \end{aligned}$$

APPENDIX B

HIDDEN MARKOV MODEL

In some sense, HMM is an extension of the concept of Markov models to include the case where the observation is a probabilistic function of the state i.e., the resulting model (i.e., HMM) is a double embedded stochastic process with an underlying stochastic process that is not observable (it is hidden), but can only be observed through another set of stochastic processes that produce the sequence of observations.

Elements of an HMM

An HMM is characterized by the following :

(a) N, the number of states in the model. Although the states are hidden, for many applications there is often some physical significance attached to the states or to the set of states of the model. Generally the states are interconnected in such a way that any state can be reached from any other state (an ergodic model); however, other possible interconnections of states are often of interest. The individual states are denoted as $S = \{S_1, S_2, ..., S_N\}$, and the state at time t as q_t.

(b) M, the number of distinct observation symbols per state i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the system being modeled. The individual symbols are denoted as $V = \{V_1, V_2, ..., V_M\}$.

(c) The state transition probability distribution $A = \{a_{ij}\}$ where,

$$a_{ij} = P[q_{t+1} = S_j | q_t = S_i], \quad 1 \leq i, j \leq N$$

For the special case where any state can reach any other state in a single step, we have $a_{ij} \geq 0$ for all i, j. For other types of HMMs, we would have $a_{ij} = 0$ for one or more (i, j) pairs.

(d) The observation symbol probability distribution in state j, $B = \{b_j(k)\}$, where,

$$b_j(k) = P[V_k \quad at \quad t | q_t = S_j], \quad 1 \leq j \leq N, \quad 1 \leq k \leq M$$

(e) The initial state probability distribution $\pi = \{\pi_i\}$ where,

$$\pi_i = P[q_1 = S_i], \quad 1 \leq i \leq N$$

Given appropriate values of N, M, A, B and π, the HMM can be used as a generator to give an observation sequence,

$$O = O_1 O_2 O_3 ... O_T$$

(where each observation symbol O_t is one of the symbols form V, and T is the number of observations in the sequence) as follows:

(a) Choose an initial state $q_1 = S_i$ according to the initial state distribution π.

(b) Set $t = 1$.

(c) Choose $O_t = V_k$ according to the symbol probability distribution on state S_i i.e., $b_i(k)$.

(d) Transit to a new state $q_{t+1} = S_j$ according to the state transition probability distribution for state S_i i.e., a_{ij}.

(e) Set $t = t + 1$, return to step 3 if $t \leq T$; otherwise terminate the procedure.

The above procedure can also be used as a model for how a given observation sequence was generated by an appropriate HMM. It can be seen from the above discussion that a complete specification of an HMM requires specification of two model parameters (N and M), specification of observation symbols, and the specification of three probability measure A, B and π. For convenience, the compact notation is used,

$$\lambda = (A, B, \pi)$$

to indicate the complete parameter set of the model.

To do the classification, the following problem is confronted in the context of HMMs. Given the observation sequence $O = O_1 O_2 ... O_T$, and a model $\lambda = (A, B, \pi)$, how to efficiently compute $P(O|\lambda)$, the probability of the observation sequence, given the model? The problem can also be viewed as one of scoring how well a given model matches a given observation sequence. For example, if the case is considered in which it is tried to choose among several competing models, the solution to the above problem allows the choice of the model which best matches the observations.

APPENDIX C

FEEDFORWARD NEURAL NETWORK

A feedforward neural network, as described in Fig. C.1, consists of:

(a) Neurons also called as artificial neurons or units or nodes.

(b) Input neurons which receive the data from the outside world and propagate them to the other units in the neural network.

(c) Output neurons which forward the data to the outside world.

(d) Hidden neurons which have no connection to the exterior of the neural network.

(e) Links that connects the neurons in a neural network. The neurons are also regarded as the processing units and the links as information transfer lines between them.

(f) Layers are defined for connections and neurons. The input neurons form the input layer and the output neurons form the output layer. The neurons in between input and output layer are grouped into hidden layers.

Notation and Terminology

(a) L is the number of layers of neurons in a neural network. The input layer has the index 0, layer L is the output layer of the network, and intermediate or hidden layers l are numbered consecutively $(0 < l < L)$, if they exist.

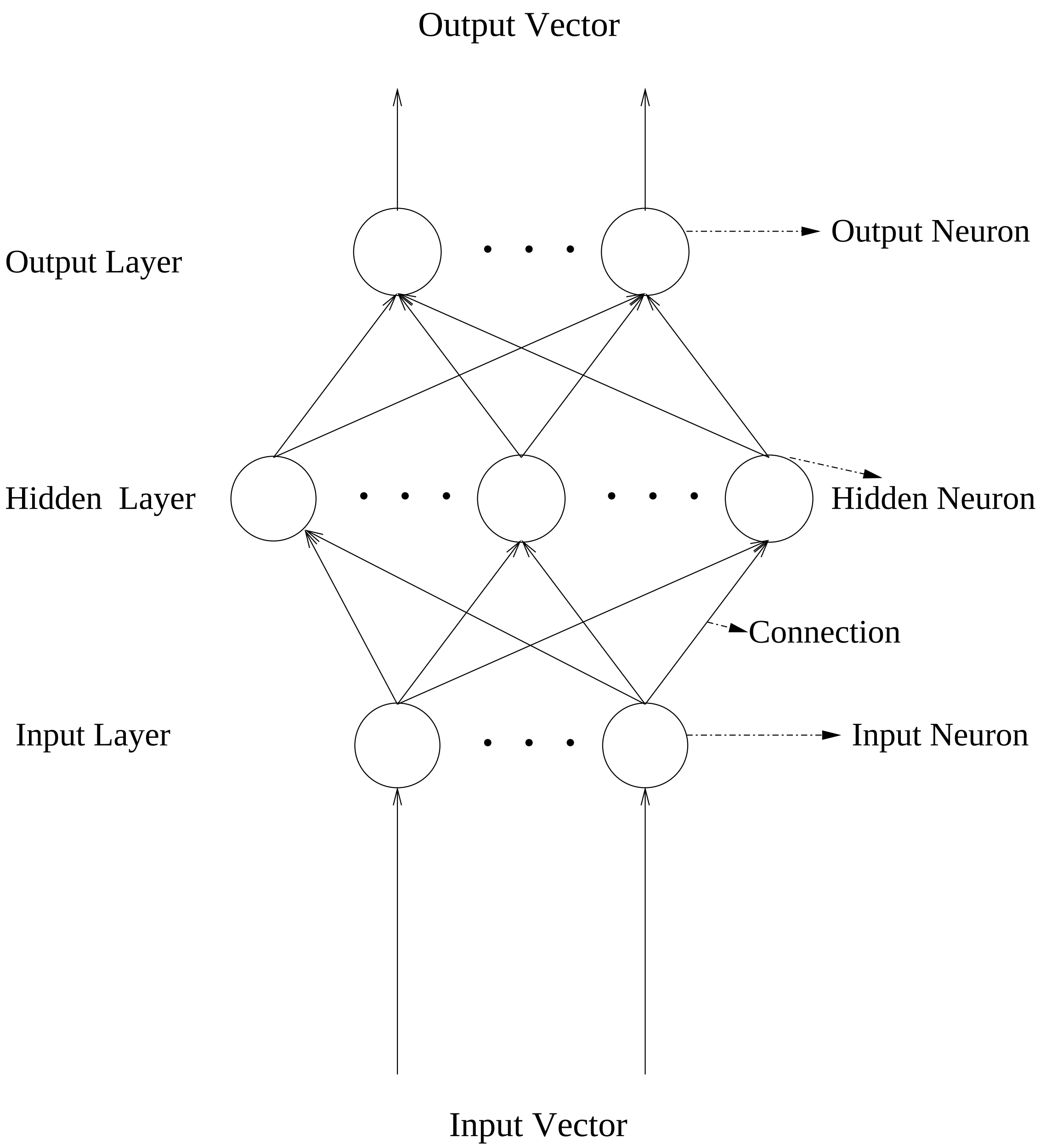


Fig. C.1: Basic feedforward neural network.

(b) N_l is the number of neurons in layer l, with its neurons numbered 1 upto N_l.

(c) C_l is the number of connections in layer l of a feedforward neural network. It equals the number of weights W_l in the same layer.

(d) $c^p_{l,i}$ is the output of the i^{th} connection in a layer l, when pattern p is presented.

(e) $w_{l,i,j}$ is the weight multiplied with the output of connection i of layer l and neuron j in layer l. Neuron j in layer l sums over all these products and applies a activation function $\phi(.)$.

(f) $\phi(.)$ is the activation function. Mainly four different types of activation functions are used.

  - The identity function (linear) identity(.), i.e., $f(x) = x$.
  - The logistic function,

$$f_\beta(x) = 1/(1 + exp(-\beta \cdot x)) \tag{B.1}$$

  - The hyperbolic tangent tanh(.) function,

$$tanh(x) = 2 \cdot f_2(x) - 1 \tag{B.2}$$

  - The hard limiting threshold or heavieside function,

$$threshold_t(x) = \begin{cases} 0 & : \quad x \leq t \\ 1 & : \quad x > t \end{cases} \tag{B.3}$$

  - The Gaussian activation function which is often called radial basis function.

The first derivative of hard limiting threshold function does not exist, which makes it unsuitable for the backpropagation algorithm. So when hard limiting threshold function is used either the perceptron learning algorithm is applied or the error used to update is calculated from another sigmoid function, and the threshold is only used during the recall phase [92].

(g) β is the gain parameter (also called as steepness).

(h) $a^p_{l,j}$ is the activation value of neuron j in layer l ($j > 0$) after the pattern p was propagated through the network. For $l = L$, this variable is also called the network output y^p_j.

(i) d^p_j is the target value of pattern p for output neuron j. The value of p ranges between 1 to the number of patterns P.

(j) x^p_i is the i^{th} element of input pattern p.

(k) η_l is the learning rate used for the weights in layer l. If the network uses a unique global learning rate, then the index l is omitted.

(l) $E(w)$ is the error function, objective function or cost function for a set of target vectors $\mathbf{d}^p$ and the output vectors $\mathbf{y}^p$ of the network with a certain instantiation of weights. The sum of the squared error is most often used,

$$E(w) = \sum_{p,j} (d^p_j - y^p_j)^2/2 \tag{B.4}$$

(m) $\Delta w_{l,i,j}$ is the amount by which a weight is changed and is calculated from the first derivative of $E(w)$ with respect to weight $w_{l,i,j}$ multiplied by learning rate.

APPENDIX D

BACKPROPAGATION ALGORITHM

The feedforward neural network is trained using error–backpropagation algorithm [92]. The backpropagation algorithm consists of the following steps:

(a) Initialization: The weights and the biases are initialized with random values [92].

(b) Pattern presentation: The neurons or nodes of the input layer are linear, i.e., the activation function is identity(.). The input vector $\mathbf{x}^p$ of a pattern p is used to initialize the activation values of the neuron in the input layer and its corresponding target vector $\mathbf{d}^p$ is presented,

$$a_{0,i} = x_i^p, \qquad i = 1, 2, \cdots, N_0 \tag{C.1}$$

(c) Forward propagation: During this phase, the activation values of the neurons are propagated layer wise through the network starting at input layer. The activation value $a_{l,j}$ of the neuron j in layer l is,

$$a_{l,j} = \phi(\beta \cdot h_{l,j}) \tag{C.2}$$

$$h_{l,j} = \sum_{i=0}^{N_l - 1} w_{l,i,j} \cdot c_{l,i} \tag{C.3}$$

where $\phi(.)$ is differentiable. The output of the connection is given below,

$$c_{l,i} = a_{l-1,i} \tag{C.4}$$

(d) Backward propagation of error: In this phase, the weight changes $\Delta w_{l,i,j}$ are calculated from the derivative of the error function to the weights,

$$\Delta w_{l,i,j} = \eta_l \cdot \partial E / \partial w_{l,i,j} \tag{C.5}$$

For each neuron an error signal $\delta_{l,j}$ is calculated, starting at the output layer, and then propagated back through the network towards the input layer.

The error signal for each neuron in the output layer is,

$$\delta_{L,j} = \beta \cdot \phi^{'}(\beta \cdot h_{l,j}) \cdot (d_j - y_j) \tag{C.6}$$

The error signal for each neuron other than the neurons in the output layer is,

$$\delta_{l,j} = \beta \cdot \phi^{'}(\beta) \cdot \sum_{k=0}^{k=N_{l+1}} \delta_{l+1,k} w_{l+1,j,k}, \qquad 1 \leq l \leq L-1 \tag{C.7}$$

(e) Weight update: After the calculation of all error signals, the weights and the biases are updated on-line as given in equation (C.8),

$$w_{l,i,j} = w_{l,i,j} + \eta_l \cdot \delta_{l,j} \cdot c_{l-1,i} \tag{C.8}$$

Usually the weight update is done on-line, that is after each pattern presentation. The counterpart of this is the off-line backpropagation, where all Δw are summed over the whole pattern set and afterwards applied to the network [107].

(f) Convergence test: Once the network is trained with all the patterns, and weights are updated, if the mean square error for the training data and the network output is too big then go back to pattern presentation in step 2. The training is stopped either after the error becomes less then a tolerable error threshold or after certain number of iterations [41, 92].

BIBLIOGRAPHY

[1] http://www.sims.berkeley.edu/research/projects/how-much-info/summary.html.

[2] http://www.bbcworldwide.com/.

[3] G. Salton, M. J. McGill, *An Introduction to Modern Information Retrieval,Patterns in Pattern Recognition: 1968-1974.* McGraw-Hill, 1983.

[4] N. Leavitt, "Let's Hear It for Audio Mining," *IEEE Computer*, pp. 23–25, Oct. 2002.

[5] P. J. Moreno, J. M. V. Thong, B. Logan and G. J. F. Jones, "From Multimedia Retrieval to Knowledge Management," *IEEE Computer*, pp. 58–66, Apr. 2002.

[6] C. Saraceno and R. Leonardi, "Audio as a Support to Scene Change Detection and Characterization of Video Sequences," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 4, (Munich, Germany), pp. 2597–2600, Apr. 1997.

[7] Report available at http://www.musclefish.com.

[8] E. Wold, T. Blum, D. Keslar and J. Wheaton, "Content-based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, pp. 27–36, Jul. 1996.

[9] Report available at http://www.mp3.com.

[10] B. Feiten and T. Ungvary, "Organizing Sounds with Neural Nets," in *Proc. Int. Computer Music Conf.*, (San Francisco, CA, USA), 1991.

[11] B. Feiten and S. Gunzel, "Automatic Indexing of a Sound Database using Self-organizing Neural Nets.," *Computer Music Journal*, vol. 18(3), pp. 53–65, Fall 1994.

[12] J. T. Foote, "Content-based Retrieval of Music and Audio," in *Multimedia, Storage and Archiving Systems II, Proc. of SPIE*, vol. 3229, (Dallas, TX), pp. 138–147, Nov. 1997.

[13] A. Ghias, J. Logan and D. Chamberlin, "Query by Humming-musical Information Retrieval in an Audio Database," in *Proc. ACM Multimedia Conf.*, (San Fancisco, CA), pp. 231–235, Nov. 1995.

[14] G. R. Doddington, "Speaker Recognition - Identifying People by their Voices," *Proc. IEEE*, vol. 73, pp. 1651–1664, Nov. 1985.

[15] T. Lambrou, P. Kudumakis, R. Speller, M. Sandler and A. Linney, "Classification of Audio Signals using Statistical Features on Time and Wavelet Transfer Domain," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 6, (Seattle, WA,), pp. 3621–3624, May 1998.

[16] B. Matityaho and M. Furst, "Neural Network Based Model for Classification of Music Type," in *Proc. 18th Conv. Electrical and Electronic Engineers*, (Israel), pp. 4.3.4/1–5, Mar. 1995.

[17] J. Saunders, "Real-time Discrimination of Broadcast Speech/Music," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 2, (Atlanta, GA), pp. 993–996, May 1996.

[18] Z. Liu, J. Huang and Y. Wang, "Classification of TV Programs based on Audio Information using Hidden Markov Model," in *Proc. IEEE 2nd Workshop Multimedia Signal Processing*, (Redondo Beach, CA), pp. 27–32, Dec. 1998.

[19] A. K. Jain, R. P. W. Duin and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, Jan. 2000.

[20] U. Grenander, *General Pattern Theory.* Oxford Univ. Press, 1993.

[21] L. N. Kanal, "Patterns in Pattern Recognition: 1968-1974," *IEEE Trans. Information Theory*, vol. 20, pp. 697–722, Nov. 1974.

[22] G. Nagy, "State of the Art in Pattern Recognition," *Proc. IEEE*, vol. 56, pp. 836–862, May 1968.

[23] P. E. Ross, "Flash of Genius," *Forbes*, pp. 98–104, Nov. 1998.

[24] S. Watanabe, *Pattern Recognition: Human and Mechanical.* New York: Wiley, 1985.

[25] S. Watanabe, *Frontiers of Pattern Recognition.* New York: Academic Press, 1972.

[26] G. T. Toussaint, "The Use of Context in Pattern Recognition," *Pattern Recognition*, vol. 10, no. 3, pp. 189–204, 1978.

[27] T. M. Cover, "The Best Two Independent Measurements are not the Two Best," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 4, pp. 116–117, Jan. 1974.

[28] J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, Mar. 1998.

[29] K. Woods, W. P. Kegelmeyer Jr. and K. Bowyer, "Combination of Multiple Classifiers Using Local Accuracy Estimates," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 405–410, Apr. 1997.

[30] R. M. Haralick, "Decision Making in Context," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 5, pp. 417–418, Mar. 1983.

[31] G. Aggarwal, A. Bajpai, A. N. Khan and B. Yegnanarayana, "Exploring Features for Audio Indexing," in *Inter-Research Institute Student Seminar*, (IISc Bangalore, India), Mar. 2002.

[32] J. Makhoul, "Linear Prediction: A Tutorial Review," *in Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[33] B. Yegnanarayana, S. R. M. Prasanna and K. S. Rao, "Speech Enhancement using Excitation Source Information," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Orlando, FL, USA), May 2002.

[34] K. S. Reddy, *Source and System Features for Speaker Recognition.* MS dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Sep. 2001.

[35] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, "Autoassociative Neural Network Models for Speaker Verification using Source Features," in *Proc. Int. Conf. Cognitive and Neural Systems*, (Boston, USA), May 2002.

[36] C. S. Gupta, S. R. M. Prasanna and B. Yegnanarayana, "Autoassociative Neural Network Models for Online Speaker Verification using Source Features from Vowels," in *Proc. Int. Joint Conf. Neural Networks*, (Honolulu, Hawaii, USA), May 2002.

[37] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, "Source Information from Linear Prediction Residual for Speaker Recognition," *communicated to J. Acoust. Soc. Amer.*, 2002.

[38] B. Yegnanarayana, K. S. Reddy and S. P. Kishore, "Source and System Features for Speaker Recognition using AANN Models," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Saltlake City, Utah, USA), May 2001.

[39] M. S. Ikbal, H. Misra and B. Yegnanarayana, "Analysis of Autoassociative Mapping Neural Networks," in *Int. Joint Conf. on Neural Networks*, (Washington, USA), Jul. 1999.

[40] J. M. Zachariah, B. Yegnanarayana, S. R. M. Prasanna and C. S. Gupta, "Combining Evidence from Source, Suprasegmental and Spectral Features for a Text-dependent Speaker Verification System," *communicated to IEEE Trans. Speech Audio Processing*, 2002.

[41] B. Yegnanarayana, *Artificial Neural Networks.* Prentice-Hall of India, New Delhi, India, 1999.

[42] C. M. Bishop, *Neural Networks for Pattern Recognition.* New York: Oxford University Press Inc., 1995.

[43] B. Cheng and D. M. Titterington, "Neural Networks: A Review from Statistical Perspective," *Statistical Science*, vol. 9, no. 1, pp. 2–54, 1994.

[44] S. Geman, E. Bienenstock and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, vol. 4, no. 1, pp. 1–58, 1992.

[45] B. Ripley, *Pattern Recognition and Neural Networks.* Cambridge: Mass.: Cambridge Univ. Press, 1996.

[46] B. Ripley, *Statistical Aspects of Neural Networks*, pp. 40–111. Networks on Chaos: Statistical and Probabilistic Aspects, eds. U. Bornndorff-Nielsen, J.Jensen and W. Kendal, Chapman and Hall, 1993.

[47] T. Kohonen, *Self-Organizing Maps*, vol. 30. Springer Series in Information Sciences, 1995.

[48] A. K. Jain, R. C. Dubes and C.-C. Chen, "Bootstrap Techniques for Error Estimation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 9, pp. 628–633, May 1987.

[49] R. Setiono and H. Liu, "Neural-Network Feature Selector," *IEEE Trans. Neural Networks*, vol. 8, pp. 654–662, May 1997.

[50] B. Yegnanarayana and P. S. Murthy, "Enhancement of Reverberant Speech using LP Residual," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 267–281, May 2000.

[51] J. Makhoul, F. Kubala, R. Leek, D. Lui, L. Nguqen, R. Schwartz and A. Srivastava, "Speech and Language Technologies for Audio Indexing and Retrieval," *Proc. IEEE*, vol. 88, pp. 1338–1353, Aug. 2000.

[52] T. Zhang and C.-C. J. Kuo, "Heuristic Approach for Generic Audio Data Segmentation and Annotation," in *Proc. ACM Int. Multimedia Conf.*, pp. 67–76, Nov. 1999.

[53] T. Zhang and C.-C. J. Kuo, "Hierarchical Classification of Audio Data for Archiving and Retrieving," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. 6, (Phoenix, AZ), pp. 3001–3004, Mar. 1999.

[54] S. Pfeiffer, S. Fisher and W. E. Elsberg, "Automatic Audio Content analysis," *ACM Multimedia*, Apr. 1996.

[55] B. Matityaho and M. Furst, "Neural Network-based Model for Classification of Music Type," in *Proc. 18th Conv. Electrical and Electronic Engineers*, (Tel Aviv, Israel), pp. 4.3.4/1–5, Mar. 1995.

[56] L. Wyse and S. Smoliar, "Toward Content-based Audio Indexing and Retrieval and New Speaker Discrimination Technique," *Inst. Syst. Sci.*, Dec. 1995.

[57] D. Kimber and L. Wilcox, "Acoustic Segmentation for Audio Browsers," in *Proc. Interface Conf.*, (Sydney, Australia), Jul. 1996.

[58] Z. Liu, J. Huang, Y. Wang and T. Chen, "Audio Feature Extraction and Analysis for Scene Classification," in *IEEE Signal Processing Soc. Workshop on Multimedia Signal Processing*, (Princeton, NJ), pp. 1–6, Jun. 1997.

[59] N. V. Patel and I. K. Sethi, "Audio Characterization for Video Indexing," in *Storage and Retrieval for Image and Video Databases (SPIE)*, (San Jose, USA), pp. 373–384, Feb. 1996.

[60] J. Boreczky and L. Wilcox, "A Hidden Markov Model Framework for Video Segmentation Using Audio an Image Features," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Seattle, WA), pp. 3741–3744, May 1998.

[61] M. Naphade, T. Kristjansson, B. Frey, et al., "Probabilistic Multimedia Objects (MULTIJECTS): A Novel Approach to Video Indexing and Retrieval in Multimedia Systems," in *Proc. IEEE Conf. on Image Processing*, vol. 3, (Chicago, IL), pp. 536–540, Oct. 1998.

[62] J. J. Wolf, "Efficient Acoustic Parameters for Speaker Recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044–2055, Mar. 1972.

[63] R. C. Lummis, "Speaker Verification by Computer using Speech Intensity for Temporal Registration," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. AU-21, no. 2, pp. 80–89, 1973.

[64] B. S. Atal, "Automatic Speaker Recognition based on Pitch Contours," *J. Acoust. Soc. Amer.*, vol. 52, pp. 1687–1697, Dec. 1972.

[65] M. J. Carey, E. S. Parris, H. Lloyd-Thomas and S. Bennett, "Robust Prosodic Features for Speaker Identification," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 3, (Philadelphia, PA), pp. 1800–1803, Mar. 1996.

[66] A. E. Thyme-Gobbel and S. E. Hutchins, "On Using Prosodic Cues in Automatic Language Identification," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 3, (Philadelphia, PA), pp. 1768–1771, Mar. 1996.

[67] E. Scheirer and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminatory," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Munich, Germany), pp. 1331–1334, Apr. 1997.

[68] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[69] G. Smith, H. Murase and K. Kashino, "Quick Audio Retrieval using Active Search," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Seattle, WA), pp. 3777–3780, May 1998.

[70] M. P. Cook, *Modeling Auditory Processing and Organization.* Cambridge, UK: Cambridge Univ. Press, 1993.

[71] S. R. Subramanya, A. Youssef, B. Narahari and R. Simha, "Use of Transforms for Indexing in Audio Databases," in *Proc. Third Int. Conf. on Computational Intelligence and Multimedia Applications*, (New Delhi, India), pp. 278–284, Sep. 1999.

[72] L. Lu and H.-J. Zhang, "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. Speech, Audio Processing*, vol. 10, pp. 504–514, Oct. 2002.

[73] S. Chen and P. Gopalakrishnan, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, (Landsdowne, VA), Feb. 1998.

[74] P. C.Woodland, T. Hain, S. Johnson, T. Niesler, A. Tuerk and S. Young, "Experiments in Broadcast News Transcription," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Seattle, WA), pp. 909–912, May 1998.

[75] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke and R. Gadde, "The Development of SRIs 1997 Broadcast News Transcription System," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, (Landsdowne, VA), Feb. 1998.

[76] G. Guo and S. Z. Li, "Content-based Audio Classification and Retrieval by Support Vector Machines," *IEEE Trans. on Neural Networks*, vol. 14, pp. 209–215, Jan. 2003.

[77] M. Siegler, U. Jain, B. Raj and R. Stern, "Automatic Segmentation, Classification and Clustering of Broadcast News Audio," in *Proc. DARPA Speech Recognition Workshop*, (The Westfields Conference Center, Chantilly VA), Feb. 1997.

[78] H. Gish, M. H. Siu and R. Rohlicek, "Segregation of Speakers for Speech Recognition and Speaker Identification," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, (Toronto, Canada), pp. 873–876, May 1991.

[79] J. Rissanen, "Universal Coding, Information, Prediction, and Estimation," *IEEE transactions on information theory*, vol. IT-30, Jul. 1984.

[80] S. Z. Li, "Content-based Audio Classification and Retrieval Using the Nearest Feature Line Method," *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 619–625, Sep. 2000.

[81] H. Wactlar, A. Hauptmann and M. Witbrock, "Informedia: News-on-demand Experiments in Speech Recognition," in *Proc. ARPA Speech Recognition Workshop*, (Arden House, Harriman, NY), Feb. 1996.

[82] S. Wegmann, F. Scattone, I. Carp, L. Gillick, R. Roth and J. Yamron, "Dragon Systems 1997 Broadcast News Transcription System," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, (Landsdowne, VA), Feb. 1998.

[83] A. V. N. S. Anjani, *Analysis of Autoassociative Neural Network Models for Processing of Degraded Speech.* MS dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Jun. 2000.

[84] C. S. Gupta, *Significance of Source Features for Speaker Recognition.* MS dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Apr. 2003.

[85] Y. Wang, Z. Liu and J.-C. Huang, "Multimedia Content Analysis using both Audio and Visual Clues," *IEEE Signal Processing Magazine*, vol. 17, pp. 12–36, Nov. 2000.

[86] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.

[87] NIST, "Speaker Recognition Workshop Notebook," *Proc. NIST 2000 Speaker Recognition Workshop, University of Maryland, USA*, Jun 2000.

[88] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms and Applications.* New Delhi: Prentice-Hall of India, 1997.

[89] A. V. Oppenheim and A. S. Willsky, *Signals and Systems.* New Delhi: Prentice-Hall of India, 1998.

[90] B. Yegnanarayana, S. P. Wagh and S. Rajendran, "A Speaker Verification System using Prosodic Features," in *Proc. Int. Conf. Spoken Language Processing*, (Yokohama, Japan), pp. 1867–1870, Sep. 1994.

[91] B. D. Ripley, "Neural Networks and Related Methods for Classification," *J. Roy. Statist. Soc. Ser. B*, vol. 56, pp. 409–456, Nov. 1994.

[92] G. Thimm, *Optimization of High Order Perceptrons.* PhD dissertation, EPFL, Communication Systems Division, EPFL, Lausanne, 1997.

[93] M. A. Kramer, "Nonlinear Principal Component Analysis using Autoassociative Neural Networks," *AIChE Journal*, vol. 37, pp. 233–243, Feb. 1991.

[94] K. I. Diamantaras and S. Y. Kung, *Principal Component Neural networks: Theory and Applications.* New York: John Wiley and Sons Inc., 1996.

[95] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New Jersey: Prentice-Hall Inc., 1999.

[96] B. Yegnanarayana, *Artificial Neural Networks for Pattern Recognition*, vol. 19. Sadhana, Apr. 1994.

[97] J. M. Zurada, *Introduction to Artificial Neural Network Systems.* Singapore: Information access and distribution, 1992.

[98] B. Yegnanarayana and S. P. Kishore, "AANN-An Alternate to GMM for Pattern Recognition," *Neural Networks*, vol. 15, pp. 459–469, Apr. 2002.

[99] S. E. Fahlman, "An Empirical Study of Learning Speed in Back-propagation Networks," Technical Report CMU-CS-88-162, School of Compute Science, Carnegie Mellon University, Pittsburgh, PA, USA, Sep. 1988.

[100] S. Haykin, *Neural Networks: A Comprehensive Foundation.* New York: Macmillan College Publishing Company, 1994.

[101] K. Chen, L. Wang and H. Chi, "Methods of Combining Multiple Classifiers with Different Features and their Applications to Text-independent Speaker Identification," *Int. J. Pattern Recognition and Artificial Intelligence*, vol. 11, no. 3, pp. 1–18, 1997.

[102] L. Xu, A. Krzyzak, C. Y. Suen, "Method of Combining Multiple Classifiers and Their Application to Handwritten Numeral Recognition," *IEEE Trans. on Systems, Man and Cybernetics*, vol. 22, pp. 418–435, May 1992.

[103] D. Genoud, G. Gravier, F. Bimbot and G. Chollet, "Combining Methods to Improve Speaker Verification Decision," Technical Report IDIAP-RR-96-02, IDIAP, Martigny, Switzerland, 1996.

[104] B. Duerr, W. Haettich, H. Tropf and G. Winkler, “A Combination of Statistical and Syntactical Pattern Recognition Applied to Classification of Unconstrained Handwritten Numerals,” *Pattern Recognition*, vol. 12, no. 3, pp. 189–199, 1980.

[105] D. Mazurov, I. Krivonogov and S. Kazantsev, “Solving Optimization and Identification Problems by the Committee Methods,” *Pattern Recognition*, vol. 20, no. 4, pp. 371–378, 1987.

[106] T. K. Ho, J. J. Hull and S. N. Srihari, “Decision Combination in Multiple Classifier Systems,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp. 66–75, Jan. 1994.

[107] D. E. Rumelhart, G. E. Hinton and R. T. Williams, “Learning Internal Representations by Error Propagation,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (D. E. Rumelhart and J. L. McClelland and the PDP research group, ed.), vol. 1, pp. 318–362, Cambridge: MA: MIT press. Reprinted in Anderson and Rosenfield, 1988.

LIST OF PUBLICATIONS

PRESENTATIONS IN CONFERENCES

- Anvita Bajpai and B.Yegnanarayana, "Audio Clip Classification using LP Residual and Neural Networks Models," *accepted for The 12th European Signal Processing Conference (EUSIPCO - 2004)*, (Vienna-Austria), 6-10 Sep. 2004.
- Anvita Bajpai and B. Yegnanarayana, "Exploring Features for Audio Indexing using LP Residual and AANN Models," *accepted for The 17th International FLAIRS Conference (FLAIRS - 2004)*, (Miami Beach, Florida), 17-19 May 2004.
- Anvita Bajpai and B. Yegnanarayana, "Exploring Features for Audio Clip Classification using LP residual and AANN Models," *International Conference on Intelligent Sensing and Information Processing (ICISIP - 2004)*, (Chennai, India), 4-7 Jan. 2004.
- Gaurav Aggarwal, Anvita Bajpai, A. Nayeemulla Khan and B. Yegnanarayana, "Exploring Features for Audio Indexing," in *Inter-Research Institute Student Seminar (IRISS)*, (Indian Institute of Science, Bangalore, India), Mar. 2002.