Audio Based Advertisement Segmentation Algorithm

By

Arsalaan Ahmed Shaikh

2010-NUST-MS-PHD-IT-03

Supervisor

Dr. Hammad Qureshi

NUST-SEECS

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Information Technology (MS IT)

In

School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(August 2014)

Approval

It is certified that the contents and form of the thesis entitled "**Audio Based Advertisement Segmentation Algorithm**" submitted by **Arsalaan Ahmed Shaikh** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Hammad Qureshi**

Signature: _____

Date: _____

Committee Member 1: **Dr. Khalid Latif**

Signature: _____

Date: _____

Committee Member 2: **Dr. Amir Shafi**

Signature: _____

Date: _____

Committee Member 3: **Dr. Muddassir Malik**

Signature: _____

Date: _____

Abstract

Broadcast monitoring involves evaluating whether the correct advertisement was aired and for the correct length of time and at the time previously agreed upon. Advertisement forms the bulk of the revenue of television channels hence; advertisement monitoring and auditing becomes extremely important for the industry. Lack of efficient and inexpensive monitoring technologies directly impacts growth in advertisement revenues. Hence, there is a need for accurate and automatic computer based techniques for auditing broadcast content as the advertisers only pay when the television time allocated to the advertisement is verified. Heterogeneity in aired advertisements, signal quality, sizable amount of multimedia data and demand for high accuracy makes it a challenging problem to solve.

Currently broadcast monitoring is performed manually i.e. using human effort, which is inefficient and prone to errors. Some computer based techniques have been proposed, however, these techniques lack whole industry coverage and, ability to scale to large number of broadcast channels and robustness.

In this thesis we propose a solution for broadcast monitoring. Our solution uses an audio feature extraction based advertisement segmentation algorithm. When given an input stream (broadcast transmission) the algorithm matches it against a database of advertisements and automatically detects instances of aired samples within the input stream. Several different audio features were computed such as *Spectral Entropy, Zero-Crossing Rate, Harmonic Ratio, Spectrum Basis, MFCC, GBFE, etc.* The audio features were analysed using Average Dependency & Minimum-Maximum Distance criterion. The criterion judges the ability of an audio feature to differentiate between classes of objects. The high quality features selected on the basis of

the criteria are *Gabor Filter Bank Feature (GBFE), Mel-frequency Cepstral Coefficient (MFCC) and MPEG7 Audio Flatness Mean.* Subsequently, in order to increase scalability, robustness and recognition accuracy, dimensionality of the feature vectors is reduced using sequential forward floating feature selection guided by the Average Dependency and Minimum-Maximum Distance criterion.

Experimental comparison of advertisement sequence classification was conducted to analyse the efficiency and effectiveness of the feature extraction algorithms. Real-world dataset of 216 hours of television broadcasts (belonging to three separate channels) and 28 classes of advertisements were used in the experiment. While comparing different audio features it was observed that a scalable, robust and accurate television broadcast monitoring system can be designed using Gabor Filter Bank Feature (GBFE) which yields recognition accuracy of over 99.33%. Moreover, the results achieved are comparable with available literature in terms of accuracy and are superior in terms of speed, robustness and efficiency. It is also concluded that the feature selected along with the matching technique may be used to construct a scalable and efficient system for television broadcast monitoring. However, there is room for improvement as in this work we target only audio analysis and video based matching would be required for a universal broadcast auditing system.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at National University of Sciences & Technology (NUST) School of Electrical Engineering & Computer Science (SEECS) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Arsalaan Ahmed Shaikh**

Signature: _____

Acknowledgment

First and foremost, I am immensely thankful to Almighty Allah for letting me pursue and fulfill my dreams. Nothing could have been possible without His blessings.

I would like to thank my supervisor, Dr. Hammad Qureshi, for guiding me through out my research. His sincere conversations through regular meetings and pushing me towards new heights. I really acknowledge his constant encouragements and efforts for the continuation of my research work.

I express my gratitude to my thesis Committee Members Dr. Aamir Shafi, Dr. Khalid Latif and Dr. Muddassir Malik, for their useful suggestions on my research.

I would like to acknowledge Ozone Solution Pakistan for providing all the necessary material for this thesis.

Finally, I would like to thank my parents, wife and children for their support throughout my masters, specially in the last year of my Master degree. They have always supported and encouraged me to do my best in all matters of life. I dedicate this work to them.

Arsalaan Ahmed Shaikh

Contents

List of Abbreviations

Abbreviations	Description
MFCC	Mel-frequency Cepstral Coefficient
GBFE	Gabor Filter Bank Feature
MPEG7	MPEG Content Multimedia Description Interface
AD	Average Dependency
MMD	Minimum-Maximum Distance
PAL	Phase Alternating Line
NTSC	National Television System Committee
DVB	Digital Video Broadcasting
EBMS	Electronic Broadcast Monitoring Systems
VIVA	Visual Identity Verification Auditor
DDL	Descriptor Definition Language
ZCR	Zero-Crossing Rate
FFT	Fast Fourier transform
PCC	Pearson Correlation Coefficient
SFSM	Sequential Floating Selection Method

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In the last decade or so, Pakistani broadcast industry has witnessed significant growth, particularly in its television sector. According to PEMRA (2010), the television industry starting with a handful of state owned channels has increased to 85 satellite channels along with 2000 cable channels. This increase in channels is mainly attributed to cable TV distribution which has provided access to TV programming to 75% of the urban and rural areas in the country. The primary source of revenues for over 2000 cable and satellite channels is through advertisements. As a result, there is a severe competition amongst the channel broadcasters. To maximize revenues the channel broadcasters frequently violate standard operating guidelines provided by PEMRA (2010) which leads to problems such as:

- Violation of standard maximum airtime for advertisements (17 minutes per-hour).
- Airing of advertisements on a time other than the time agreed upon.
- Airing of wrong advertisements.
- Airing of advertisements but with incorrect duration.
- Airing of copyrighted material.

Problems such as the ones mentioned above may lead to significant losses to advertisers and is the main cause for the stunted growth of the industry. Identifying and detecting such issues over a large number of channels requires an automated, scalable and efficient broadcast monitoring system.

Television broadcast monitoring entails auditing of the channel to check whether the correct segment was aired, whether it was aired at the correct time and whether it was of appropriate length or not. Advertisement air time is expensive and advertisers are reluctant to pay high rates unless an objective auditing mechanism is in place. This mechanism provides the basis on which broadcasters and advertisers resolve their claims. Television broadcasts generate large amounts of video and audio data that needs to be analyzed for effective broadcast time verification. Some solutions have been proposed such as embedding information within the broadcast signals or manually reviewing broadcasts by human operators. These systems lack whole industry coverage, proper data assessment or require implementation of broadcasting standards which is not easy to achieve.

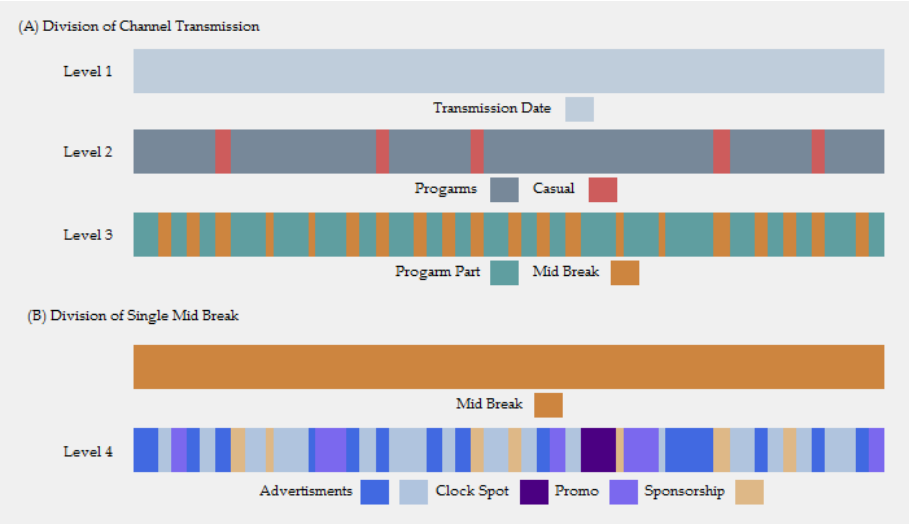

Figure 1.1: Transmission layers

In traditional broadcast monitoring, an auditing company (on request of clients) provides logs of specific advertisements that aired during trans-

mission. A transmission date is divided in to several levels as can be seen in Figure 1.1 which shows the hierarchy used by a monitoring company for reporting (Level 4 shows only some of the types of spots that are aired during mid-breaks). Program parts and advertisements are divided into further parts such as scenes, shots, seconds and frames. The logs contain the time of day the advert was aired, duration and channel of the advert. The purpose of the logs is to identify deviations from the advert broadcast plan originally agreed upon by the advertisers and broadcasters. Inconsistencies such as variations in the number of times an advertisement was shown, the timings of when it was shown (there are separate rates for different times in a day) and in the duration of the advertisements are identified. Aggressive competition amongst the broadcasters (as the growth in satellite television channels is rapidly increasing) requires the logging of all non-commercial and commercial transmission. Therefore it is a norm for a monitoring company to maintain logs of the complete day's transmission for a channel being audited. Manual auditing (i.e. using human effort) of the channels is not cost-effective as there are large number of channels and huge quantities of multimedia data. Moreover, manual auditing is error prone. To remove these errors additional quality assurance and control measures are required.

1.2 Technical Background

Channel broadcasters transmit programming based on colour encoding standards such as PAL, NTSM, DVB, etc. and these standards define broadcast properties such as frame rates, image resolution and audio modulation. For auditing, aired transmission is re-encoded and captured for the purpose of maintaining a permanent record. The transmission contains both audio and video data. Digitized audio data, as compared to video data, requires less storage and has less variation across television channels, thereby, providing better scalability and accuracy. Variations such as channels having different logos, additional information strip at bottom of the screen, difference in colour sharpness, down scaling and up scaling of advertisement and additional logos on top of the advertisements are present. A sample of these

can be seen in Figure 1.2. However, video and audio data are both equally relevant and offer a means of accurate auditing of advertisement as stipulated in Oliveira et al. (2005), Camarena-Ibarrola et al. (2009) and Gauch and Shivadas (2005).

1.3 Feature Extraction

Transmission in its raw form is very large, analysing raw data is very complex and inefficient. To reduce the size of data, a process known as feature extraction is used. Feature extraction transforms the input data into a reduced representation. The features (usually represented in the form of vectors) are then provided to different pattern recognition algorithms to differentiate one class of object from another. Using the extracted features to perform a classification task generally requires a suitable classification function. The function consigns an input object to one of a set of classes Sivagaminathan and Ramakrishnan (2007).

Transmission in its raw form is very large, analysing raw data is highly complex and inefficient. To reduce the size of data, a process known as feature extraction is used. Feature extraction transforms the input data into a reduced representation. The features (usually represented in the form of vectors) are then provided to different pattern recognition algorithms to differentiate one class of object from another. Using the extracted features for a classification task generally requires a suitable classification function. The function consigns an input object to one of a set of classes Sivagaminathan and Ramakrishnan (2007). Feature extraction has been used in many different audio based applications such as speech recognition Jurafsky and Martin (2000), music modelling Li et al. (2003), similar sound index Spevak and Polfreman (2001), etc.

The feature extraction process considerably reduces the size of data. However, some of the features may be irrelevant or redundant. Avoiding such features is important because they may have a negative impact on the accuracy of the algorithm. In addition, having a large feature dimension hinders the scalability of the solution by increasing the cost of acquiring the data.

Figure 1.2: Variation in identical advertisement aired on different channels.

The problem becomes increasingly important for real-time systems that need to analyse the data in the shortest amount of time; therefore defining an appropriate set of features is important. The process of adequately selecting a feature subset is a significant problem in its own right. The problem refers to the task of identifying and selecting a useful subset of features to be used to represent patterns from a larger set of often mutually redundant, possibly irrelevant, features with different associated measurement costs and/or risks Sivagaminathan and Ramakrishnan (2007).

1.4 Problem Statement

Aired television contains advertisements which are broadcast at different intervals. The advertisements are heterogeneous in terms of duration, audio and video data. The main challenge is the designing of an algorithm for segmentation of these advertisements from the transmission broadcast.

A formal statement is given as:

"Identification and selection of different audio features for the purpose of automatic detection of aired television advertisements"

1.5 Thesis Contribution

Our main contribution in this thesis is the development of an efficient, robust and scalable advertisement segmentation algorithm based on Gabor feature extraction techniques. Other significant contributions of this thesis are summarized as:

- Extraction and selection of high quality features using Gabor feature extraction techniques.
- Feature subset reduction using sequential forward floating algorithm and average distance and minimum-maximum distance criteria.
- Performance comparison of different audio features including GBFE, MFCC and MPEG7 audio descriptors.
- Performance comparison of different distance measures for classification.

1.6 Thesis Organization

This thesis focuses on the audio segmentation of advertisement, a complex sequence labeling classification problem. Rest of the thesis is organized as follows. In Chapter 2, research work already carried out relevant to television broadcast monitoring and audio feature is discussed. In Chapter 3, proposed methodology is discussed in detail along with discussion of feature extraction and feature subset selection. In Chapter 4, the experimental results obtained by adopting above mentioned methodology are discussed and analysed. Chapter 5, presents a conclusion and some directions regarding future work are also presented.

Chapter 2

Related Work

In this chapter, literature related to broadcast monitoring system is presented. However, due to the commercial nature of the system limited work is publicly available, some have been explained below. Our emphasis is to study different audio based applications, audio feature extraction techniques, feature selection algorithms and sequence labelling problems.

2.1 Broadcast Monitoring

Broadcast monitoring these days is carried out both manually and using automatic computer-based systems. Manual systems such as the one used by MediaBank (2012) employs human operators to monitor television broadcast content and creates logs in a computer-based database. As stated earlier, manual auditing is prone to error and susceptible to variations due to the training level of auditors and their skills. Since human labour and effort required is immense (due to the huge volumes of data), such systems become very expensive to operate.

2.2 Electronic Broadcast Monitoring

A lot of work has already been carried out on the development of electronic broadcast monitoring systems (EBMS). Commercial importance of the sys-

tem has led to multiple patents in the area Pitman et al. (2003),Logan A. et al. (2006) and Welsh (1989). EBMS generally fall in two categories as can be seen in Figure 2.1;

In the first category the systems are designed on pre broadcast information. Welsh (1989) proposed a video monitoring system that functions by comparing closed caption text (program content description available in NTSC signal) with stored text to detect commercial and program content. Essentially the information of what is currently on air already resides within the transmission signal. A device that decodes the signal can easily detect the aired content. The method is quite efficient and provides real time monitoring however, it requires complete change in technology in regions where the standard is not being used.

Kalker et al. (1999), presents a system for multimedia copyright infringement detection. The technology has been developed at the Philips Research Laboratories and named Visual Identity Verification Auditor (VIVA). The proposed solution is based on a steganography based technique. The technology embeds a watermark within frames of the video content i.e. programs and advertisements. The watermark is embedded in such a way that they are invisible to the viewer, robust in terms of common video processing steps (compression, digital to analogue conversion or vice versa, editing, format conversion and change of aspect ratio), low probability of false positives and false negatives and also allows real time detection.

Systems described in Welsh (1989) and Kalker et al. (1999) require the participation of broadcasters, advertisers and acceptance by the industry. It also requires both pre and post broadcast processing. Implementation of such techniques will requires lot of resources and time, especially in emergent markets such as South Asia and China where there is little or no standardization.

The second category of EBMS, analyse broadcast signal properties to automatically detect advertisement content using computer-based analysis. These systems extract audio and video properties from signals and compare them with a pre extracted feature database using pattern recognition techniques. Large amount of data is processed during feature extraction and

advertisement segmentation stages. Such systems have been described by Oliveira et al. (2005) and Camarena-Ibarrola et al. (2009).

Oliveira et al. (2005), proposes a system based on a Short-Term Fourier Transform audio feature. The transformation process results in a series of vectors (i.e. features). The system then uses a classification scheme which divides the input audio stream into blocks and classifies them into clusters based on their similarity and exploiting the temporal evolution of the signal frequency spectrum. Detection of advertisements takes place in a segmentation phase which processes the audio clusters and classifies them to an advertisement class. Experiments were conducted over a sample set of 41 hours of audio stream with 393 sample advertisements, processing the sample set took 3 hours with a low recognition rate. Furthermore, the transmission sample set used is small in terms of input audio stream which comprises of just 41 hours. However, the technique of clustering the database leads to considerable reduction in processing time of the algorithm.

In the another similar system, Camarena-Ibarrola et al. (2009). presents a system based on Multi-Band Spectral Entropy fingerprint, the broadcast signals are processed for feature extraction and are compared with a library of pre-extracted advertisements. The extraction process is applied on blocks 250 msec. The audio goes through 7 processing steps before the actual signature is attained. To validate the system, experiments were conducted on a sample set of 95 hours of audio stream with 13 different advertisements. A high recognition rate was achieved, however it was achieved at the cost of processing time as it takes 2 hours to process 24 hours of stream over the 13 adverts. In addition, the sample space contains a single transmission date for 5 different channels which may hide the problems occurred due to variation in day to day transmission signal.

Both systems contain flaws for instance they have not solved the problem of cross channel (segmentation using same source advertisement for all television channels) advertisement segmentation which limits their scalability and robustness on large number of channels. In addition, the proposed solutions have either compromised on accuracy for efficiency or vice versa.

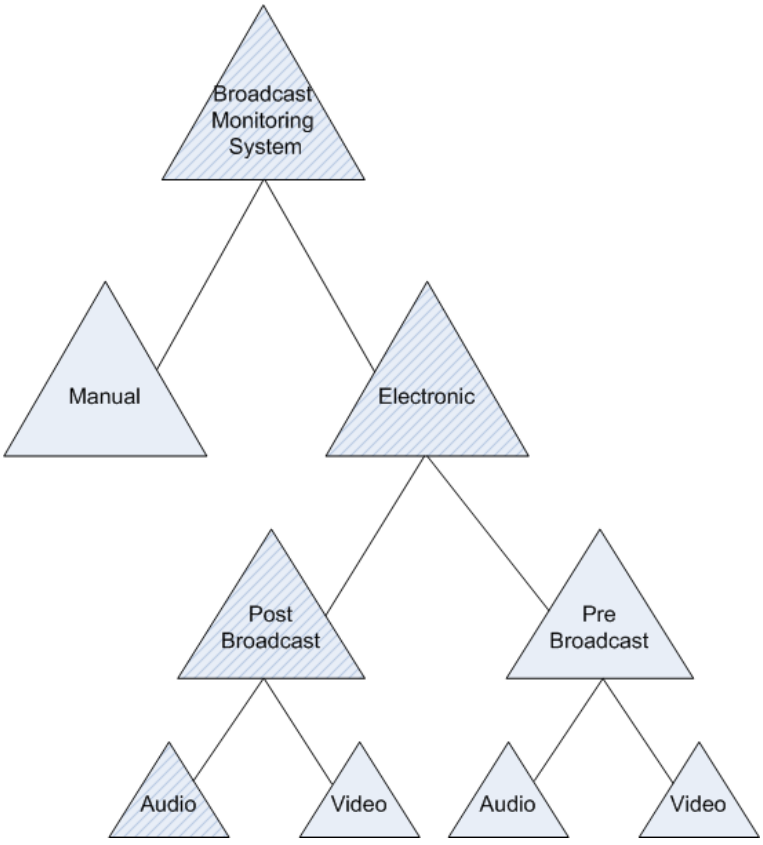

Figure 2.1: Monitoring system hierarchy

2.3 Audio Based Applications

Kim et al. (2004), presents a system whose main goal was to understand the efficiency of an audio classification and indexing system based on MPEG7 features and mel-frequency cepstral coefficients (MFCC) using Independent Component and Principal Component Analysis. The system classifies sample audio in to different classes such as Animal, People, Objects, etc. It processes an input audio through a feature extraction phase which is based on MPEG7 features, Audio Spectrum Basis and Audio Spectrum Projection. The dimensionality of the extracted audio features was reduced while trying to maintain maximum data integrity, which is one of the systems key aspects. Experiments were conducted on two separate sets of data (sounds and speech). For speech, speaker recognition and gender recognition was performed with results of up to 96% and 100% respectively. While for sound, a database was searched for similar sounds which achieved 96% recognition rate on a single level classification hierarchy.

Foote (2000), proposes an audio segmentation technique which detects significant change in the audio sequence. A self-similarity algorithm based on standard spectral analysis is used. The audio stream is chunked in very small blocks where both pre and post blocks are compared. In this way no cue (such as silence) is required for this algorithm, eliminating the flaw of having a cue present for segmentation. The method is similar to a shot boundary detection algorithm in which difference of consecutive frames in a video are detected iteratively until a frame crosses the similarity threshold. The frames containing little or no difference are bunched together into a single shot.

Many audio applications use MPEG7 based features it is important to understand the basics of the standard. Chang et al. (2001) , provide a high level view of Multimedia Content Description Interface (MPEG7). It explains all the key elements of the standard such as architectures, audio, video & multimedia standard descriptors, and description definition language along with the scope of MPEG7. The standard provides some basic high and low level descriptors, and contains the flexibility of defining new descriptors through the DDL. It does not explain how the information is extracted from

multimedia content but how the information is presented once it is extracted.

Matushima et al. (2004), explain a music searching system based on pattern recognition. The system uses MPEG7 low level descriptors such as Audio Power Type, Audio Spectrum Centroid Type, Audio Waveform Type, Audio Spectrum Envelope Type, Audio Spectrum Flatness Type and Audio Spectrum Spread Type. The low level descriptors were chosen as they are generated without human interaction. The system incorporates the following main components music genre classifier, mpeg7 descriptor generator, mpeg7 descriptor database and a search tool. Experiments were performed on 10 different music genres and a total of 160 songs. The system achieved lowest accuracy of 43% and highest of 88%. The results vary across music genres. The problem may lie in a generalized matching threshold which can be solved by using specific threshold trained for each genre.

Herre et al. (2001), explore different audio features to determine their performance when the audio signal is susceptible to distortion which may be the case in many real world applications. The authors argue that features such as Spectral Flatness Measure (SFM) and Spectral Crest Factors (SCF) perform better as compared to traditional features such as energy, loudness, sharpness, spectral centroid, zero crossing rates, bandwidth, pitch and MFCC. A test-bed was implemented using 1000 musical items which were subjected to distortion such as cropping, re-encoding, etc. Comparisons were made between loudness vs. SFM, SFM and SCF. Result show major differences in favour of the features proposed by the authors.

Srinivasan et al. (1999), explore different audio features that distinguish speech from music. The main goal of the authors was to classify mixed audio in segments of silence, speech and music. Average ZCR and average energy were the features chosen for this purpose. Experiments were performed on 4 audio files which had a combined duration of 1 hour, on which four ZCR variations were tested. The results achieved were mediocre at best, with each ZCR variation either having many false positives or a low recognition rate.

Lin et al. (2005), propose an improvement to an existing audio classification algorithm by using wavelet transform and a bottom up support vector machine. Sub band and Pitch properties of audio were extracted through

wavelets which were used along with Frequency cepstral coefficient (FCC). The researchers were able to reduce the dimensionality of the features while also reducing the error rate. For experimentation, Muscle Fish, a public audio library was used. The library contains 410 sounds divided into 16 classes. Results show that existing error rate of 8% was reduced to 3%. It was observed that the dimensionality of feature vector depends on the number FCC value represented by L. The minimum error rate of 3% achieved by the proposed system using L = 80 a very high dimension. While the existing system achieved error rate of 8% using L=8.

Popescu et al. (2009), explore different audio features with regards to music classification. Features such as Multi resolution wavelet analysis, spectral analysis and beat histogram were used along with MFCC, ZCR and energy. Two classifying algorithms K-NN and K-Means were used. Experimentation was performed on an audio library containing 150 audio files belonging to 5 genres. Different recognition rates were achieved on different music genres with results ranging between 62% and 100%.

Baluja and Covell (2008), propose an audio retrieval system based on wavelet audio fingerprinting. The system incorporates many different techniques such as hash tables, spectrogram images, exact and nearest neighbour matching algorithms. The system has been designed to be applicable in real time environments where quality degradation and performance of speed, memory and resource consumption are major issues. Experiments were conducted on a sample set of 1000 audio files. Signal degradation was performed on them to simulate real world distortion such as echo, equalizer, mpeg2 transcoding, GSM multirate transcoding, etc. Very high recognition rates were achieved when compared to an existing solution.

Chapter 3

Methods and Materials

The solution is primarily based on audio feature extraction; therefore it is important to understand different audio features and their extraction techniques. Below is an overview of some extraction algorithms.

3.1 Overview of Feature Extraction Algorithms

3.1.1 Mel-frequency Cepstral Coefficient (MFCC)

Mel-frequency Cepstral Coefficient (MFCC) is a feature extraction technique which emphasizes on the speech part of an audio. MFCC is a frequently used technique in applications in which speech or speaker identification is of importance such as Hasan et al. (2004) and Milner and Shao (2002). MFCC takes human perception sensitivity with respect to frequencies into considerations (Jang). Figure 3.1 shows the steps required for the extraction of the coefficient followed by a step by step explanation.

1. Audio of various broadcast is acquired in Wav format.

2. Human sound production mechanism suppresses high-frequency formants (the spectral peaks of the sound spectrum of the voice). Pre-emphasis balances this defect by applying a high-pass filter on the signal.

$$S_2(n) = S(n) - a * S(n-1) \tag{3.1}$$

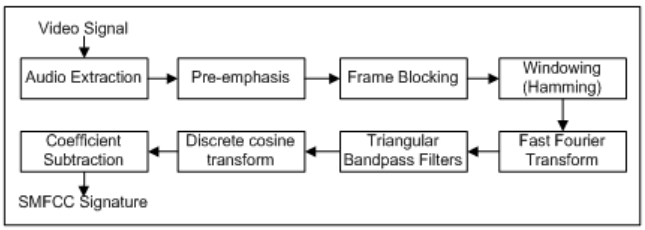

Figure 3.1: MFCC extraction process

Where $S(n)$ is the original signal, $S_2(n)$ is the transformed signal after pre-emphasis and constant a is between 0.9 to 1.0

3. The signal is segmented in to blocks of 80 milliseconds called frames with an overlapping factor of 66%. The overlapping factor ensures that a frame is created after every 27.2 milliseconds.

4. As audio signal is continuous, continuity must be maintained at the first and last point of the frame. For this the frames are multiplied by hamming window denoted by the equation 3.2.

$$w(n, a) = (1 - a) - a\cos(2pn/(N - 1)) \tag{3.2}$$

5. Following windowing, FFT is applied which converts the signal from time domain to frequency domain.

6. The FFT spectrum is equally spaced over a linear scale while voice inside the audio signal is not linear. The spectrum is converted from linear scale to a non-liner Mel scale using a set of triangular band past filters. Correspondence between linear frequency to Mel-frequency is given by the equation 3.3.

$$mel(f) = 1125 * ln(1 + f/700) \tag{3.3}$$

7. The MFC coefficients are generated by applying discrete cosine transform on the Mel spectrum generated in the previous step.

8. In the final step each frame value in a MFCC vector is subtracted by its corresponding frame value.

$$smfcc(n, m) = mfcc(n, m) - mfcc(n - 1, m) \tag{3.4}$$

Where N is the number of frames and M is the number of coefficients.

Many different versions of the MFCC extraction algorithms have been proposed by Han et al. (2006), Hossan et al. (2010) and Zhao et al. (2012), for this study the implementation proposed by Siciarz has been used.

Almost all advertisements broadcast over television contain large degree of speech; therefore MFCC extraction based technique could be applied for automatic advertisement detection.

3.1.2 MPEG Content Management Description Interface

MPEG (Moving Pictures Experts Group) content multimedia description interface commonly known as MPEG-7 is an ISO/IEC standard. The specification is developed for structuring multimedia content for fast and easy retrieval. The standard provides specification for visual, audio and multimedia descriptors; however the specification itself does not provide their implementation.

The descriptors are classified in low-level and high-level categories. Low-level descriptors describe the basic attributes of the content such as its color, size, power, energy, etc. On the other hand high-level descriptors contain a higher semantic hierarchy. The melody descriptor is one such example, various implementations are available but their performance may wary. Hence, only their results can be standardized. Moreover, high-level descriptors require human intervention such as scene definition JTC1/SC29/WG11 (2002).

An automated television broadcast monitoring system to its fullest extent cannot compromise on human intervention. Keeping these guidelines in mind only low-level descriptors have been considered. This comparative study utilizes the implementation proposed in Crysand et al. (2004). The figure 3.2

depicts the available audio descriptors. Some of the well-known descriptors are defined below.

- Audio-Fundamental-Frequency (AFF): The descriptor analyses the audio stream to project tone of the speech and harmonic pitch. This descriptor can also cross examine data from musical and melody descriptors even though it has not been specifically designed to do so. The descriptor specification includes a weight field that controls scaling of audio that shows absence of clear periodicity.

- Audio-Harmonicity (AH): Harmonicity of a sound defines how well structured it is or that they have a harmonic spectrum. Sounds such as musical instruments, human speech, etc. contain a harmonic spectrum. On the other hand, noise does not contain any harmonic spectrum. The degree of harmonicity of a sound is defined by the Audio Harmonicity descriptor, which differentiates sounds with a harmonic and non-harmonic spectrum.

- Audio-Waveform (AWF): The descriptor is generally used for illustration purposes. This allows quick display of the waveform without processing the audio signal. The waveform can also be used for fast comparison between waveforms.

- Audio-Spectrum-Basis (ASB): The descriptor has been used with great effect in automatic classification and retrieval systems. Its ability to use basis functions to reduce high-dimensional features to a compact representation plays a key role. Low-dimensional set of statistical information are essential for many real world applications that have scalability concerns.

- Audio-Spectrum-Centroid (ASC): Audio-Spectrum-Centroid Descriptor describes the center of gravity of the log-frequency power spectrum. The Spectrum-Centroid is defined as the power weighted log-frequency centroid. Spectrum centroid is an economical description of the shape of the power spectrum.

- Audio-Spectrum-Envelope (ASE): Audio-Spectrum-Envelope Desciptor describes the spectrum of the audio according to a logarithmic frequency scale. The AudioSpectrumEnvelope Descriptor describes the short-term power spectrum of the audio waveform as a time series of spectra with a logarithmic frequency axis

- Audio-Spectrum-Projection (ASP): A compact projection of spectrum features which are selected using the basis function. The dimension can be altered based on the way they are intended to be used. If time varying components are used dimension = MxNx(K+1), where M is quantity of blocks, N is the spectrum length and K is the number of functions used. On the other hand if stationary basis is used dimension becomes Nx(K+1).

- Audio-Spectrum-Flatness (ASF): Audio-Spectrum-Flatness Descriptor describes the flatness properties of the spectrum of an audio signal within a given number of frequency bands. The Audio-Spectrum-Flatness-Type describes the flatness properties of the short-term power spectrum of an audio signal.

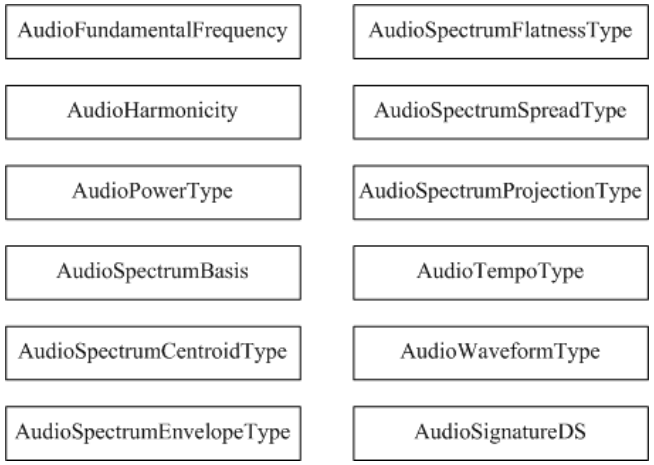

Figure 3.2: MPEF7 Audio Descriptors

3.1.3 Gabor Filter Bank Features

A relatively new audio feature has been proposed by Schädler et al. (2012) known as Gabor Filter Bank Features (GBFE). GBFE extracts spectro-temporal information from the audio signal. The features have been implemented for audio application such as automatic speech recognition systems in Schädler et al. (2012) & Schädler and Kollmeier (2012) and speaker recognition in Lei et al. (2012).

GBFE uses a set of 41 filters (Gabor Filters) which are based on human physiology Schädler et al. (2012). During the feature extraction process the log Mel-spectrum of the audio signal is calculated, which is filtered using the Gabor Filters resulting in a 311-dimensional feature vector.

Similar to MFCC, GBFE is motivated from aural perception of human speech. As speech is an integral part of aired advertisements, GBFE can be an applicable feature for automatic television broadcast monitoring. One concern is the large dimension of the features which may limit the scalability of the monitoring system.

3.2 Methods

In a manual monitoring system, human operators review recorded transmission and manually mark occurrence of an advertisement. The solution proposed is an automatic system which is based upon feature extraction using audio properties and then a pattern recognition algorithm is used to match the advertisement. If a match (based upon a predefined criterion) is found, the system creates a log of it in the database. The proposed solution is explained in two sections, an audio feature assessment and selection segment in which different audio features were studied and evaluated, followed by an advertisement segmentation algorithm that uses a sliding window method for fast processing. Both of these sections are explained below in detail.

3.2.1 Audio Feature Assessment

The solution is designed specifically to be implemented in regions which are lagging behind in the advancement of broadcast monitoring technology. The quality of transmission varies from channel to channel and from time to time. Table 3.1 shows difference of audio attributes of an identical advertisement that aired on two separate channels. Similar differences can also be found within the transmission of same channels at different instances of time. Figure 3.3 illustrates the waveform of 10 samples each of 2 advertisements that aired at separate instances in time. It can be seen that the waveforms are

Audio Attribute	**Channel-A**	**Channel-B**
DC offset	-0.000021	-0.000002
Min level	-0.248169	-0.359497
Max level	0.224548	0.355316
Pk lev dB	-12.11	-8.89
RMS lev dB	-28.5	-24.7
RMS Pk dB	-22.61	-18.06
RMS Tr dB	-61.5	-58.01
Bit-depth	14/16	15/16

Table 3.1: Audio difference in identical advertisement

symmetric for most of the part, however deviations can been seen which are more elaborate at the end of the waveform. These deviations are more prominent for adverts with low durations which are caused due to signal distortion and broadcasters airing short duration ads. Due to such uncertainty in transmission quality it is very important to select a robust feature.

Compared to classification problems such as similar sound indexing, segmentation of advertisement for television broadcast monitoring is more complex. The sequential nature of television broadcasts with the added constraint of un-supervised classification, high recognition rates, robustness and scalability makes this sequential classification problem a challenging task.

A television broadcast stream generally contains program content along with promotional and commercial advertisements. An audio feature must adequately distinguish one advertisement from other broadcast content. The broadcast is also subject to different levels of signal degradation, therefore the audio features (to a certain extent) must be flexible in negating the disruption.

This section assesses audio features based on the average dependency and minimum-maximum distance criteria (explained below). These benchmarks establish the capability of a feature to amply differentiate between different aired content. Furthermore, feature subset selection (based on Average Dependency criteria) is also discussed.

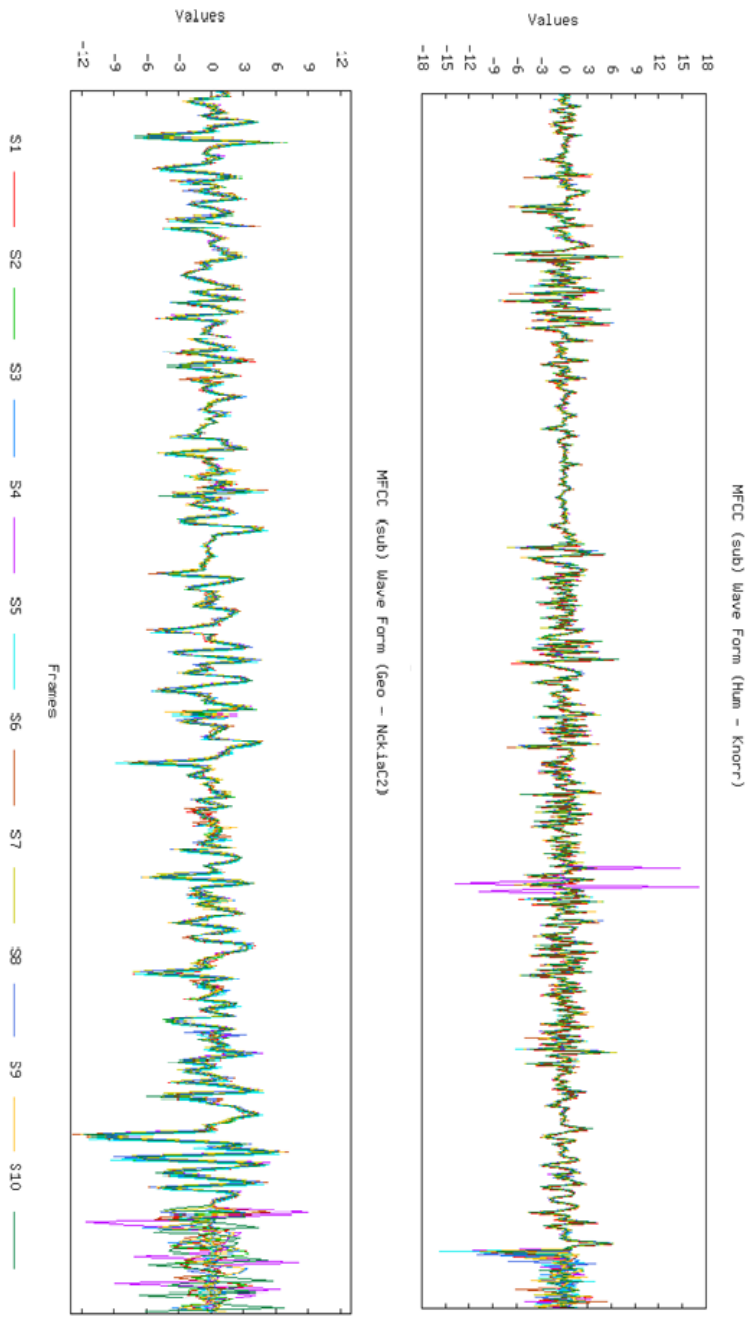

Figure 3.3: Advertisement Waveforms

3.2.1.1 Average Dependency Criteria

Consider two sets A & N and a dependency function D. Set A consists of m number of advertisements belonging to the same class (same advertisements aired at different instances) while set N consists of $m-1$ elements belonging to several different classes (program or advertisement content other than that of set A). The dependency function D determines the correlation, similarity or distance between the members of set A such as $D(A_i, A_j)$ where $i \neq j$. The function also calculates the dependency between the members of sets A and N such as $D(A_i, N_j)$.

Using the dependency function D, a true dependency matrix T and false dependency matrix F are calculated. Matrix T is formed by calculating dependency between all elements of set A while matrix F is formed by calculating the dependency between all elements of sets A and N. The matrices are shown in Figure 3.4. Once the matrices are formed the average $avgT$ and $avgF$ of all elements of T and F are calculated respectively. Finally by subtracting $avgT$ and $avgF$ the Average Dependency (AD) is calculated.

$T =$

$D(A1,A2)$	$D(A1,A3)$	$D(A1,A4)$	$D(A1,Am)$
$D(A2,A1)$	$D(A2,A3)$	$D(A2,A4)$	$D(A2,Am)$
$D(A3,A1)$	$D(A3,A2)$	$D(A3,A4)$	$D(A3,Am)$
$D(A4,A1)$	$D(A4,A2)$	$D(A4,A3)$	$D(A4,Am)$
$D(Am,A1)$	$D(Am,A2)$	$D(Am,A3)$	$D(Am,Am\text{-}1)$

$F =$

$D(A1,N1)$	$D(A1,N2)$	$D(A1,N3)$	$D(A1,Nm\text{-}1)$
$D(A2,N1)$	$D(A2,N2)$	$D(A2,N3)$	$D(A2,Nm\text{-}1)$
$D(A3,N1)$	$D(A3,N2)$	$D(A3,N3)$	$D(A3,Nm\text{-}1)$
$D(A4,N1)$	$D(A4,N2)$	$D(A4,N3)$	$D(A4,Nm\text{-}1)$
$D(Am,N1)$	$D(Am,N2)$	$D(Am,N3)$	$D(Am,Nm\text{-}1)$

$$avgT = \frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{m-1} Tij\right)}{m * m - 1}$$

$$avgF = \frac{\left(\sum_{i=1}^{m}\sum_{j=1}^{m-1} Fij\right)}{m * m - 1}$$

Figure 3.4: True and false dependency matrix

To evaluate a feature, AD is calculated and then averaged across k number of advertisement classes. A feature is judged based on its ability to attain a large AD across different classes of advertisements.

3.2.1.2 Minimum-Maximum Distance Criteria

The transmission broadcast has variation in signal quality, therefore there lies a possibility that some elements belonging to T may have lower values than that of elements in F (even though overall AD is high). To restrict this possibility the Minimum-Maximum Distance (MMD) criterion is used. MMD is calculated by subtracting the minimum element of T with the maximum element of F. A large positive MMD value attained by an audio feature establishes its degree of robustness.

3.2.1.3 Distance & Similarity Functions

The homogeneous similarity or distance between subjects is an essential cog in many scientific fields such as machine intelligence, decision support systems, forecasting and more so to pattern recognition. Classification algorithms such as Nearest Neighbor, K-nearest Neighbor, etc. are based on distance functions. In the case of advertisement segmentation an approach is required to specify that advertisement $'A1'$ is close to $'A2'$, however faraway from $'B1'$. Identifying an appropriate (low processing cost and high classification accuracy) distance function may enable the design of a robust and scalable advertisement segmentation algorithm. The functions defined below are commonly used in classifiers and support vector machines.

- Euclidean Distance: also known as Euclidean metric is one of the most commonly used distance functions it calculates the basic distance between two points.

- Manhattan Distance: Distance calculated over two points along the right angle axis is known as Manhattan Distance.

- Minkowski Distance: is the generalization of Euclidean and Manhattan distance. The distance function is used when a variable has wide range of values and when small number of boundary values impacts the overall result.

- Pearson Correlation Coefficient: Allows the identification of positive and negative associations between linear objects. It is represent by a value ranging between -1 and +1. A value boarding to -1 represents a negative relationship while value of +1 indicates a perfect positive relationship. 0 represents no association.

- RBFKernel Distance: A well-known kernel function (based on Euclidean distance) which is primarily used in support vector machine classification. Output values range between 0 (completely similar) and 1 (meaning no similarity).

- Spearman Footrule Distance: It is similar to Manhattan distance in terms of calculation, however Manhattan distance uses the variables quantitative value while Footrule distance is calculated on the rank of the variables quantitative value.

- Chebychev Distance: Consider two vectors A and B with dimension x. Then Chebychev distance can be represented as $Max\{|A1 - B1|, |A1 - B1|...|Ax - Bx|\}$. Essentially the maximum distance between individual properties becomes distance between the two objects.

- Cosine Distance: Calculates the angular distance between two vectors where the values are bounded between 0 and 1.

Function	**AD**	**MMD**
EuclideanDistance	0.137443	-0.015738345
MaxProductSimilarity	0.001534	-0.000845642
MinkowskiDistance	0.137443	-0.015738345
RBFKernelDistance	0.000737	-8.76186E-05
SpearmanFootruleDistance	5.39E-05	-5.49021E-06
PearsonCorrelationCoefficient	0.528812	0.297704368
ChebychevDistance	0.025409	0.002311407
CosineSimilarity	0.075795	0.03943881

Table 3.2: Ad and MMD values for dependency functions

3.2.1.4 Assessment

To assess the audio features 15 sets of advertisements were prepared. The cardinality of each set of advertisement was 15 (making a total of 225 sam-

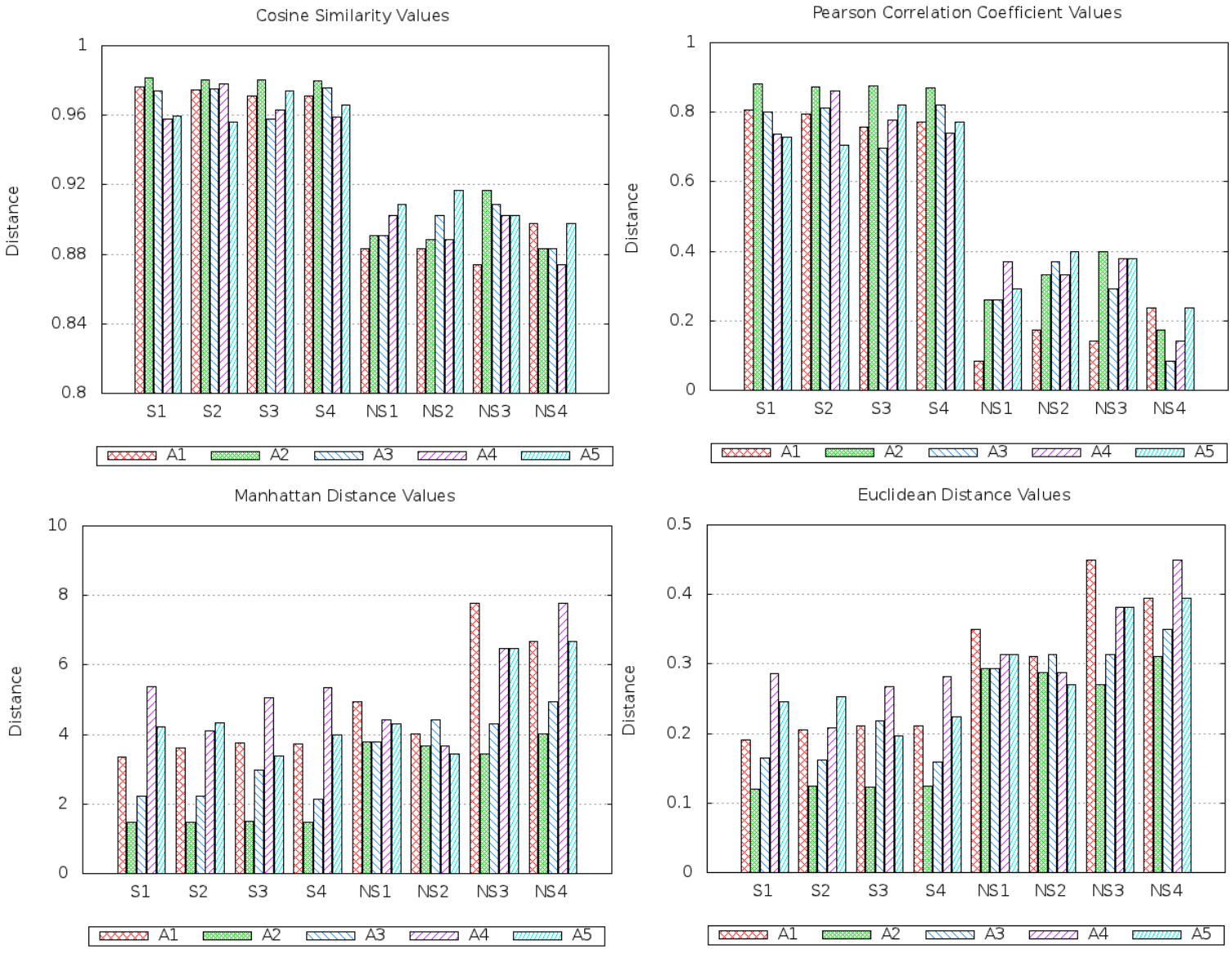

Figure 3.5: Dependency matrix values for different functions

ples). The samples were cropped from 216 hours of recorded transmission belonging to 3 different television channels. The transmission contained quality variation which is also reflected in the cropped advertisement samples. Selected samples were heterogeneous in nature encompassing sequences of speech, music, speech-music overlay and signal variation. The set was also prepared from the same transmission; however the samples were cropped from sequences other than that of the advertisement sets.

For the dependency function Pearson Correlation Coefficient (PCC) was used. Affectivity of PCC in data analysis and detecting linear relationships has already been established by Mueen et al. (2010). PCC measures the strength of association or co-dependency of two objects. Its result lies between -1 (strong negative association) and 1 (strong positive association). To establish PCC as an appropriate dependency function a comparative analysis was conducted against many well-known distance functions using the prepared sample set. Uniformity of PCC against Manhattan, Euclidean and Cosine distance functions is clearly visible in Figure 3.5. Moreover, PCC's superior AD and MMD values of 0.52 and 0.29 (view Table 3.2) respectively validate the appropriateness of PCC.

Table 3.3 shows the dimensions and the AD & MMD values calculated on the audio features. It can be observed that all features have yielded positive AD values, however majority of them are close to 0. Contiguity towards 0 shows the feature's inability to distinguish between true and false samples. Moreover, features with strong AD values are shown to have a weak MMD.

Figures 3.6 and 3.7 show AD values (only 5 advertisement classes for clear illustration) of Audio Flatness Mean and Zero-crossing Rate. Although both features have similar overall AD values, Zero-Crossing Rate has a low MMD. Considering uniformity in AD values, Audio Flatness Mean is a more suitable feature.

An automatic broadcast monitoring system based on features that have low MMD or AD value will yield large quantity of false positive and false negative results.

It can be stated that none of the features (using current implementations and dimension) can be used for a highly accurate, robust and scalable broad-

Audio Feature	**Dim**	**AD**	**MMD**
Audio Flatness Variance	16	0.5557	0.3103
Audio Flatness Mean	16	0.5555	0.3717
Audio Fundamental Frequency	1	0.1521	-0.0170
Harmonic Ratio	1	0.3904	0.0346
Upper Limit Harmonicity	1	0.0424	-0.0275
Audio Power Type	1	0.3461	-0.0263
Audio Spectrum Basis	8	0.5069	-0.1376
Audio Spectrum Projection	9	0.0068	-0.0194
Audio Spectrum Centroid	1	0.3784	0.0247
Audio Spectrum Envelope	34	0.2957	0.0871
Audio Spectrum Spread	1	0.3344	-0.0382
Spectrum Basis	8	0.5069	-0.1376
Audio Wave Form	2	0.0998	-0.0067
Spectral Entropy	1	0.4055	-0.1166
Zero-crossing Rate	1	0.4464	-0.1011
GBFE	311	0.0859	0.0192
SMFCC	10	0.0774	-0.5840

Table 3.3: Ad and MMD values for each audio feature

cast monitoring system. Table 3.3 indicates that Audio Flatness Variance and Audio Flatness Mean are the only relatively strong candidates, however much larger AD and MMD value will be required for a concrete automatic broadcast monitoring system.

3.2.1.5 Feature Subset Selection

The impact of large feature dimensions on accuracy and processing time has been stated in Pourhabibi et al. (2011); Sivagaminathan and Ramakrishnan (2007); Mitra et al. (2002); Robnik-ikonja and Kononenko (2003). Features such as Audio Flatness Mean, GBFE, Audio Spectrum Envelope, SMFCC , etc. have large dimensions (as seen in Table 3.3), which may have impacted their AD and MMD values.

Different feature subset selection algorithms can be found in literature such as SWR Colak and Isik (2003); Yang and Honavar (1998) and Branch & Bound algorithm Pudil et al. (1994). However, According to Jain and

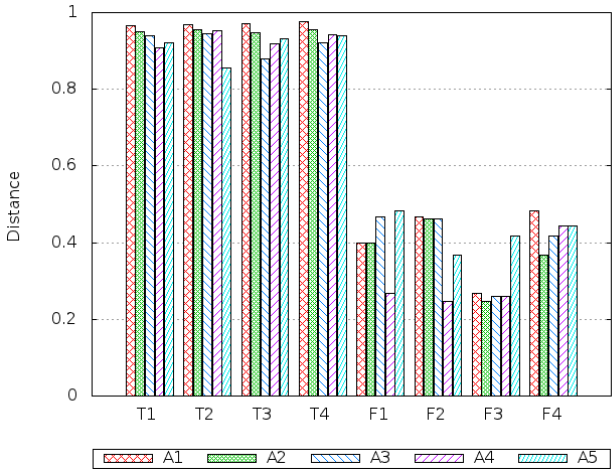

Figure 3.6: Dependency matrix values for Audio Flatness Mean

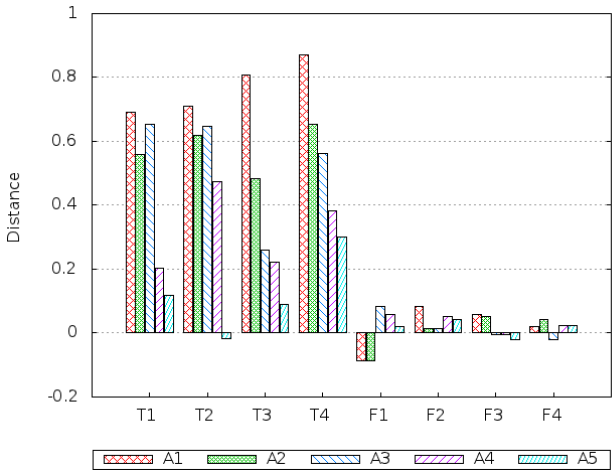

Figure 3.7: Dependency matrix values for Zero-crossing Rate

Zongker (1997) sequential floating search method (SFSM) is one of the most effective algorithms. In short SFSM starts with a null feature set and, for each step, the algorithm adds a dimension to the current set if it satisfies some criterion function. The process continues until each of the feature dimensions are accepted or rejected by the criterion function.

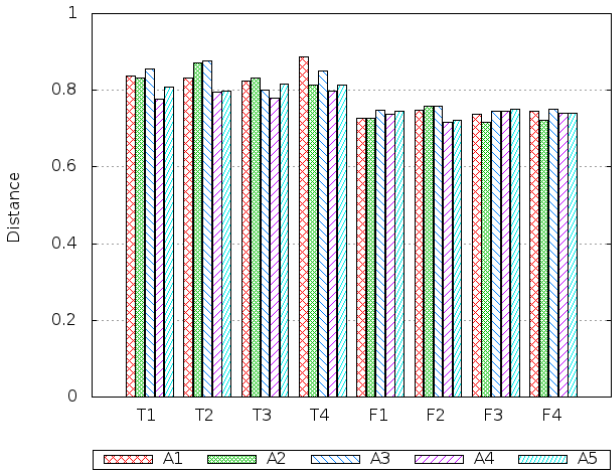

Figure 3.8: Dependency matrix values for GBFE (Before SFSM)

The criterion function used is a minimum Average Dependency (AD) value. On each step of SFSM, AD is calculated on the next dimension of the audio feature along with the dimensions that have already satisfied the criterion in previous iterations. If the resulting AD is greater than the minimum Average Dependency (AD) value the dimension is included into the selected feature dimension set and minimum Average Dependency (AD) is set to the calculated AD. Initially minimum Average Dependency (AD) of the audio feature is set to its overall AD shown in Table 3.3. The effect of SFSM can be seen in Figures 3.8 and 3.9 which show that AD values of GBFE calculated before and after the feature subset selection respectively.

Table 3.4 shows the reduced dimension size and AD & MMD values calculated on the optimum dimension subset of each audio feature. It can be observed that applying feature dimension reduction process has significantly increased the AD of all audio features. The increase in AD has also reflected

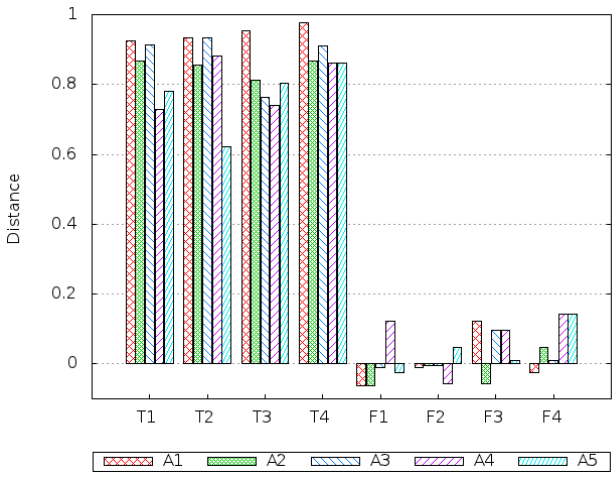

Figure 3.9: Dependency matrix values for GBFE (After SFSM)

in substantial rise in MMD values for some of the features. Moreover, the process has reduced the dimension size to such an extent that processing time will be significantly less than if the original dimensions were used.

The feature assessment based on Average Dependency and Minimum-Maximum Distance criteria has yielded three potentially strong feature candidates Audio Flatness Mean, GBFE and SMFCC. The advertisement sequence classification experiments that were performed on the three audio features are discussed in the next section.

Audio Feature	**Dim**	**AD**	**MMD**
Audio Flatness Variance	1	0.7533	-0.0039
Audio Flatness Mean	**3**	**0.7536**	**0.4102**
Audio Spectrum Basis	2	0.8185	0.0153
Audio Spectrum Projection	1	0.5276	0.0467
Audio Spectrum Envelope	1	0.3679	-0.1155
Spectrum Basis	2	0.8184	0.0142
Audio Wave Form	1	0.4101	-0.0121
GBFE	**8**	**0.8236**	**0.4777**
SMFCC	**1**	**0.7403**	**0.4766**

Table 3.4: Ad and MMD values for each audio feature (After SFSM)

3.2.1.6 Feature Vector Subtraction

Even with the selection of an optimum set of feature dimensions, they may be insufficient in their original form because of the quality difference between channels. One of problems that occur due to transmission quality is highlighted in Figure 3.10 which shows a difference in amplitude in identical advertisements across channels. To overcome this problem coefficient value of each audio feature vector such as $feat(n, m)$ is subtracted by the coefficient value of the preceding feature vector $feat(n-1, m)$. Where n is the vector number and m is the dimension number. The effects of subtraction can be seen in Figure 3.11 which shows the waveform of the same advertisements depicted in Figure 3.10 after subtraction. The advertisements are now close to identical.

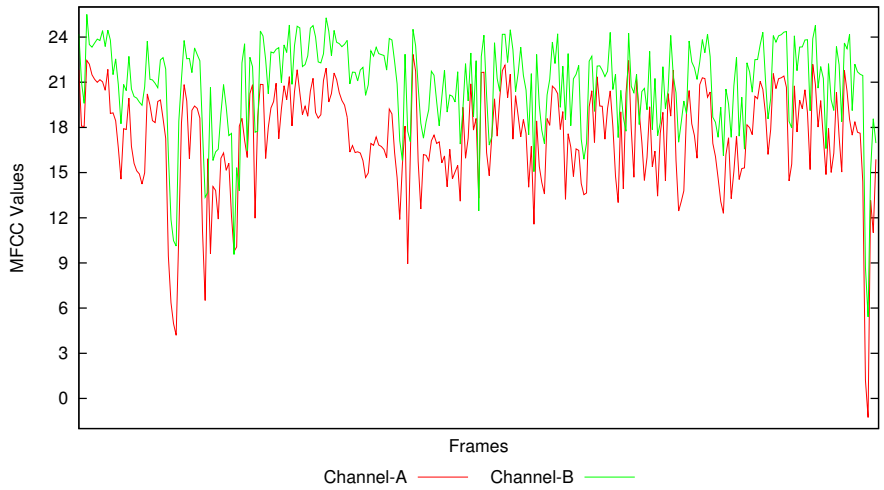

Figure 3.10: MFCC representation of identical advertisements

3.2.2 Advertisement Sequence Classification

In this section two phase algorithm for advertisement sequence classification is proposed, the algorithm pseudocode is provided in Figure 3.12. In the first phase, a sliding window technique is used. The window of limited length (comprising of starting frames of the source advertisement) slides across the

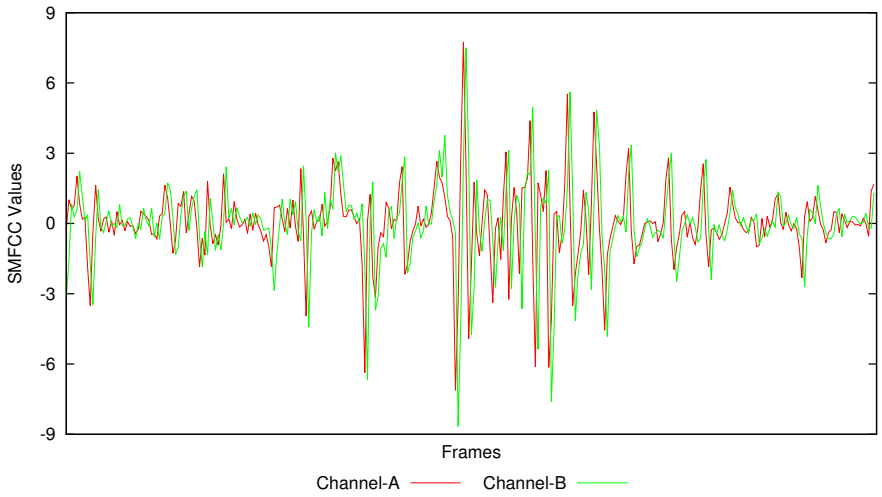

Figure 3.11: Representation of identical advertisements after subtraction

transmission frames sequentially with a sliding step of 1. At each step starting frames of source advertisement are matched (using dependency function) with the transmission frames starting from the current iteration. If the result is true, phase two is initiated. In phase two all frames of the source advert are matched with the transmission frames starting from the current iteration.

- *'AD'* a sequence of vectors $[V_1, V_2, ...V_n]$ containing features of the advert that must be detected. Each vector represents a single frame which varies in size depending on the feature.
- *'TRANS'* a sequence of vectors $[V_1, V_2, ...V_m]$ containing features of a transmission, in which advert will be detected. Each vector represents a single frame which varies in size depending on the feature.
- *'windowsize'* Cardinality of the array of start frames used in phase 1.
- *'RESULT'* an array containing start position of matched advertisements.

Figure 3.13 graphically explains the sequence classification process. In this example source advertisement contains five frames $A[0..4]$ and the trans-

```
SLIDING-WINDOW (AD, TRANS, windowsize)

1   ADStart ← copy[AD(1 to windowsize)]
2   for j ← 1 to length[TRANS] – windowsize
3       TransW ← copy[TRANS(j to j+windowsize)]
4       res ← Calculate-Frame-PCC(ADStart, TransW)
5           if res = TRUE
6               TransC ← copy[TRANS(j to j+Length[AD])]
7               matched ← Calculate-Frame-PCC(AD, TransC)
8               if matched = TRUE
9                   add j in RESULT

Calculate-Frame-PCC (AD, TRANS)

1   pcc ← Pearson-Correlation-Coefficient(AD, TRANS)
2       if pcc >= th
3           return TRUE
4       else
5           return FALSE
```

Figure 3.12: Pseudocode of sliding window algorithm

mission contains $T[0..N-1]$ frames . An occurrence of the source advertisement starts at frame 4 in the transmission. At $i = 0$, using $windowsize = 2$, the starting two frames of the source advertisement are matched with the starting two frames of the transmission. As the occurrence starts at frame 4, $Calculate - Frame - PCC()$ returns $false$, the window then slides on to the next frame in the transmission and the process starts again. Similarly at $i = 1$ and $i = 2$, $Calculate - Frame - PCC()$ returns $false$. But at $i = 3$, the function returns $true$ as occurrence of the source advertisement has started. Phase two is initiated and all five frames of source advertisement are matched with five transmission frames, $Calculate - Frame - PCC()$ returns $true$ and the timings are logged to a database.

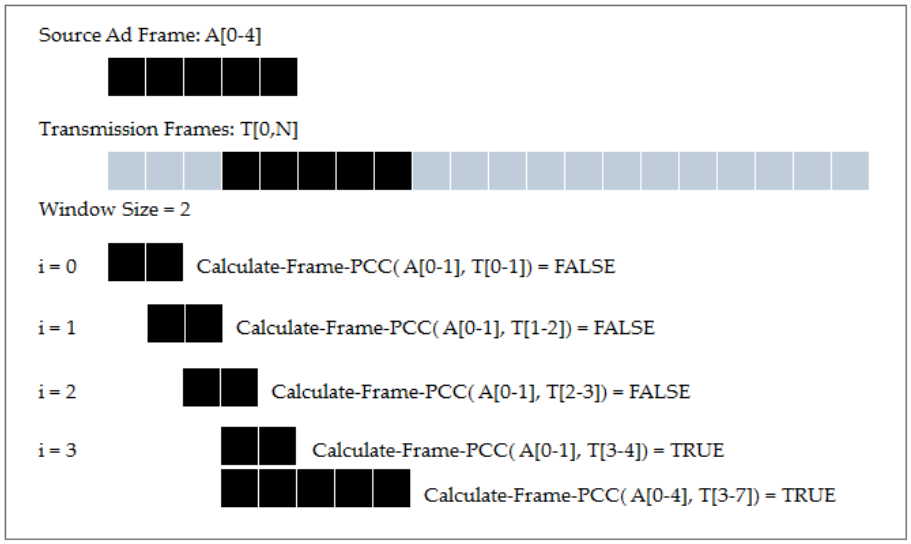

Figure 3.13: sliding window algorithm explained

Chapter 4

Results And Discussions

In this chapter we summarize the experiments performed in order to investigate the usefulness of the proposed selected features (Audio Flatness Mean, MFCC and GBFE) and the sliding window advertisement segmentation algorithm. Initially, we explain the experimental setup of real-world data, followed by description of performance attributes that are used to evaluate the feature extraction techniques. Subsequently, the results attained from the experiments are stated followed by a discussion evaluating the systems performance.

4.1 Experimental Setup

Experiments were conducted on 216 hours of captured transmission from 3 different television channels. Recordings were provided by a local monitoring company in a lossy format wma@128kbps distributed over 1296 files of 10 minutes each. A total of 28 advertisements were cropped from the transmissions files which served as the source for advertisement detection. Each advertisement was selected to represent diverse characteristics of audio as shown in Table 4.1. Manual logs of aired advertisements were also provided by the local monitoring company which acted as the ground truth. MFCC, Audio Flatness Mean and GBFE features of each of the 1296 files were extracted to a database along with the features of cropped source ad-

vertisements.

SNo.	Brand	Dur (sec)	Speech	Music	Speech/Music
1	Ariel	20	Y	N	N
2	Blueband	20	N	N	Y
3	Brite	48	N	Y	Y
4	Comfort	40	Y	N	Y
5	Dove	30	N	N	Y
6	Mobilink	20	Y	Y	Y
7	Knorr	30	Y	Y	Y
8	Kurkure	20	N	N	Y
9	Max Bar	30	Y	N	Y
10	Nokia	15	Y	N	N
11	NokiaX	30	Y	Y	N
12	Panteen	20	Y	N	Y
13	Pepsi	30	Y	Y	Y
14	Sheild	45	Y	N	Y
15	Zong	30	N	Y	Y
16	Colgate	15	Y	Y	Y
17	Telenor	45	Y	N	Y
18	Palmolive	25	N	N	Y
19	Bata	20	N	N	Y
20	Fair & Lovely	10	Y	N	Y
21	Q Mobile	35	N	Y	Y
22	Mortein	10	Y	N	N
23	Carecream	35	Y	N	Y
24	Malaysia Palm Oil	30	N	N	Y
25	Orient Dispenser	25	N	Y	Y
26	Orient Microwave	25	N	Y	Y
27	Orient Refrigerator	30	N	Y	Y
28	Government	30	N	N	Y

Table 4.1: Advertisement Samples

4.2 Performance Evaluation Attributes

Robustness, scalability and high recognition rates are some of the characteristics an automatic broadcast monitoring system should contain. In existing

literature, current systems do not perform adequately with regards to all the performance attributes. To assess the performance of our solution, the proposed feature extraction techniques are compared with respect to the following attributes.

- **High Recognition Rate:** maintaining high recognition rate for an automatic broadcast monitoring system is very important. Variation of 5% to 10% may result in loss of revenue in the range of hundreds of thousands for channel operators. However, an error margin of +/- 2% is accepted by the industry. It was observed in literature that high levels of accuracy negatively impact other performance attributes.

- **Minimum False Positives:** Similar to low accuracy, high false positives are not acceptable for monitoring systems. Any amount of false positive detections may nullify the credibility of the generated reports. Furthermore, the advertiser may have to bear losses in the amounts paid to the broadcaster.

- **Short Processing Time:** The processing time directly impacts the scalability of the systems. When we consider Pakistans broadcast industry containing 2000 cable and 80 satellite channels, having a short processing time becomes extremely essential to analyse the huge amount of data.

- **Impact of Signal Distortion & Noise:** Signal distortion is the term often used to describe undesirable change in a signal and refers to changes in a signal due to the nonideal characteristics of the transmission channel and missing samples. While noise can be defined as unwanted signal that interferes with the communication of another signal. In cases where there is high level of distortion and noise, positive detections may be impossible. However, the system should be immune to limited distortion and should correctly segment the advertisement.

- **Impact of Signal Degradation:** The transmission broadcast is captured in multimedia codecs which are optimized for storage space to

reduce the amount of resources. The original signal may be transcode, re-encoded and cropped before it is archived. Using codecs such as mp3 and wma which are lossy in nature may cause loss in information that may be important for advertisement detection and segmentation.

4.3 Results

Optimization achieved through feature selection and reduction based on Average Dependency & and Minimum-Maximum Distance criteria and feature vector subtraction is highlighted in the attained results. The recognition accuracy and false positive rates attained before dimension reduction and feature subtraction can be seen in Table 4.2. While results attained after optimization of feature vectors is highlighted in Table 4.3 and Figure 4.1. Recognition rates attained on individual advertisement samples can be seen in Table 4.4.

	CH1 Acc (%)		**CH2 Acc (%)**		**CH3 Acc (%)**	
Feature	**True +**	**False +**	**True +**	**False +**	**True +**	**False +**
SMFCC	65.74	20.4	72.92	25.2	74.3	16.64
Mean	83.21	15.19	86	13.54	92	10.81
GBFE	100	100+	100	100+	100	100+

Table 4.2: Recognition rates attained before feature optimization

	CH1 Acc (%)		**CH2 Acc (%)**		**CH3 Acc (%)**	
Feature	**True +**	**False +**	**True +**	**False +**	**True +**	**False +**
SMFCC	98.43	0.004	98.54	0.004	100	0
Mean	95.81	0.015	96.11	0.021	100	0.052
GBFE	99.19	0.002	98.8	0	100	0

Table 4.3: Recognition rates attained after feature optimization

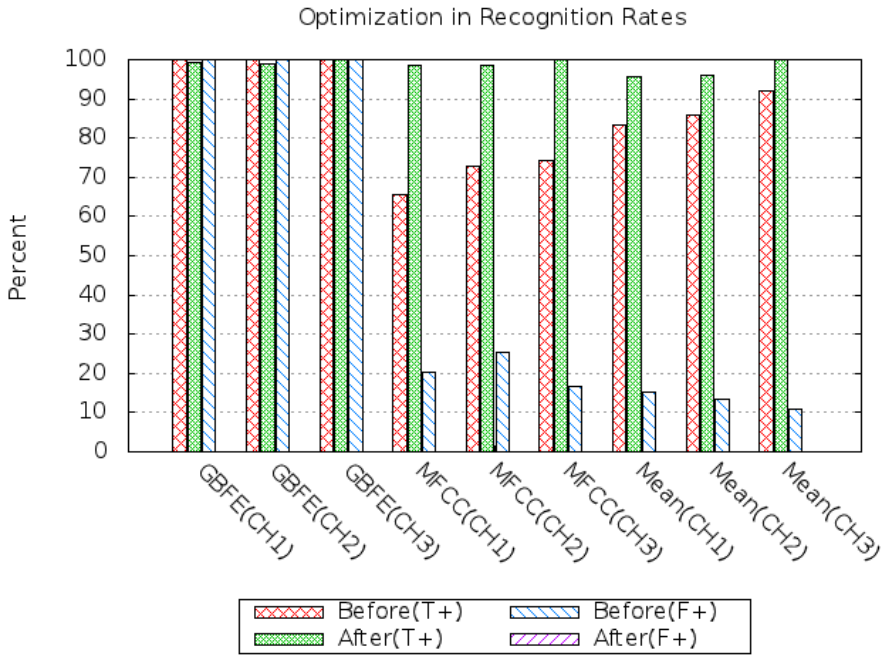

Figure 4.1: Recognition rates before/after comparison

4.4 Discussion

In this section we discuss the results stated in the previous section with respect to evaluation attributes explained earlier.

- **High Recognition Rate:** MFCC, Audio Flatness Mean and GBFE attain a very high accuracy as can been seen in Table 4.3. Overall the audio features achieve a recognition rate of 98.99, 97.31 and 99.33 percent respectively. These recognition rates are comparable with available literature. Figure 4.1 shows that techniques used for feature dimension reduction and vector subtraction considerably increase the performance of MFCC and Audio Flatness Mean feature. However, GBFE extraction technique maintains its superiority over other audio features even with a slight decrease in its recognition rate. It also meets the minimum criteria acceptable to the industry.

- **Minimum False Positives:** Feature dimension not only impacts the processing time but may also have negative impact on accuracy of the solution. This fact is highlighted in Figure 4.1 which shows that false positives rates are negligible once the dimensions are reduced and high quality features are selected. The major improvement was shown by GBFE extraction technique which is comparable with available literature.

- **Short Processing Time:** The processing time depends on many different factors such as feature vector dimension, frame blocking size, overlapping ratio, size of advertisement, etc. Processing time has been a major concern for automatic broadcast monitoring which has limited the scalability of existing solutions. By reducing the feature dimension and using a two phase segmentation algorithm we were able to increase efficiency tenfold. Audio Flatness Mean, MFCC and GBFE required a processing time of 0.23, 0.18 and 2.5 seconds per transmission hour respectively as seen in Table 4.5.

- **Impact of Noise & Distortion:** Using real-world data as sample for the experiment has provided data inherent flaws including distortion

and noise. These flaws and their impact on the signal wave form can be witnessed in Figure 4.2. The figure shows the wave forms of 10 identical advertisements aired at different instances of time. It can be observed that the wave forms are approximately identical, however samples 4 and 5 spike the otherwise symmetrical wave form at different intervals. No change in the results was observed when the distorted samples were used for advertisement segmentation.

- **Impact of Signal Degradation:** To measure the effects of signal degradation the sample advertisements were subjected to simulated signal degradation. The samples were cropped from their original source and transcoded from WMA to MP3 at a lower audio bit rate using the lossy lame encoder. No change in results was observed when the degraded samples were used for advertisement segmentation.

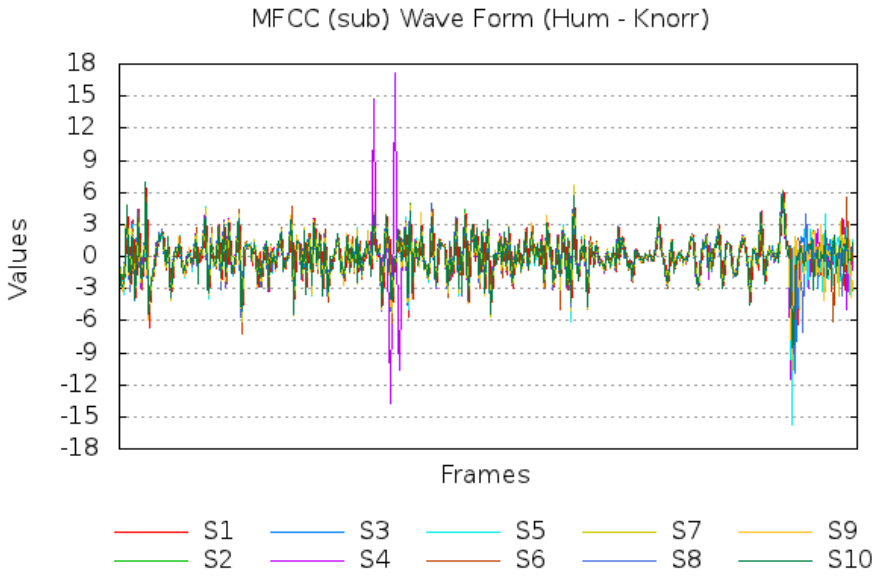

Figure 4.2: Distortion impact on wave form

	CH1 Acc (%)			CH2 Acc (%)			CH3 Acc (%)		
Ad	**MFCC**	**Mean**	**GBFE**	**MFCC**	**Mean**	**GBFE**	**MFCC**	**Mean**	**GBFE**
A1	96	96	100	100	100	100	100	100	100
A2	100	100	100	100	100	83	100	100	100
A3	97	97	100	100	100	100	100	100	100
A4	100	100	100	100	85	92	100	100	100
A5	100	100	100	100	100	100	100	100	100
A6	100	100	100	100	100	100	100	100	100
A7	100	100	100	100	88	100	100	100	100
A8	100	100	100	100	100	100	100	100	100
A9	100	100	100	100	100	100	100	100	100
A10	98	98	100	100	100	100	100	100	100
A11	89	85	100	96	100	96	100	100	100
A12	100	100	100	100	100	100	100	100	100
A13	100	100	100	100	100	100	100	100	100
A14	100	52	100	100	100	100	100	100	100
A15	100	100	100	90	100	93	100	100	100
A16	100	66	100	92	76	100	100	100	100
A27	100	100	100	100	100	100	100	100	100
A18	100	100	100	85	100	100	100	100	100
A19	100	100	100	100	100	100	100	100	100
A20	77	100	77	94	94	100	100	100	100
A21	100	100	100	100	100	100	100	100	100
A22	97	86	100	100	75	100	100	100	100
A23	100	100	100	100	100	100	100	100	100
A24	100	100	100	100	100	100	100	100	100
A25	100	100	100	100	70	100	100	100	100
A26	100	100	100	100	100	100	100	100	100
A27	100	100	100	100	100	100	100	100	100
A28	100	100	100	100	100	100	100	100	100

Table 4.4: Advertisement wise recognition rates

Audio Feature	**Before**	**After**
GBFE	98.7	2.5
Audio Flatness Mean	2.88	0.23
MFCC	2.14	0.18

Table 4.5: Processing Time Difference

Chapter 5

Conclusion And Future Work

This chapter presents a conclusion of the study with an overall summary of the research findings and possible future work.

5.1 Conclusion

Segmentation of advertisements within aired transmission for the purpose television broadcast monitoring lies in the sequence labeling category of pattern recognition. The problem area is highly challenging due to the share volume of data and the need for a scalable, robust and highly accurate system. This thesis investigates different audio feature based techniques along with distance measure classification as a potential solution to the problem. In this thesis, we have introduced a sequence labeling algorithm for advertisement segmentation i.e. using GBFE, MFCC and MPEG7 based feature extraction techniques along with Pearson Correlation Coefficient as a distance classifier. Our proposed algorithm is tested on a real-world sample set of 216 hours of aired transmission belonging to different local television channels and 28 unique advertisements. A set of high quality features were extracted and their dimension were reduced using sequential forward floating method based on an Average Dependency and Minimum-Maximum distance criterion. An overall accuracy of 98.99, 97.31 and 99.33 was achieved employing Subtracted Mel-frequency cepestral coefficients, MPEG7 audio flatness

mean and Gabor features respectively. Compared to other similar systems reported in the literature review, our system is superior, as it provides high recognition rate while maintaining low processing time by using high quality features and simple distance classification. Moreover, the proposed system has a negligible false positive rate. In addition to achieving our primary goal of advertisement segmentation, this thesis also provides performance comparison of several different audio features and distance measures. The comparisons were based on the Average Dependency and Minimum-Maximum Distance criteria employed on 225 advertisement samples. It was found that audio features GBFE, SMFCC and MPEG7 audio flatness mean performed well as compared to ZCR and Spectral Entropy, while distance measures Pearson Correlation Coefficient and Cosine Similarity were superior to Euclidean Distance and Minkowski Distance.

5.2 Future Work

Broadcast monitoring has been an active research field for many years, however due to the commercial nature of the problem limited literature is publically available. In this research we have achieved the following goals.

- A comparable recognition rate.
- A greater degree of reliability by achieving a low false positive rate.
- Higher scalability by maintaining a low processing time.

Although we have achieved very high recognition rates, there is room for improvement specifically for advertisements with durations less than 7 seconds. This can be achieved by using different feature extraction strategies and combining them together. Furthermore, improvements can be made to increase the accuracy in ascertaining the exact duration of each segmented advertisements. In addition the processing time can also be reduced further particularly for Gabor features. This research can be used as ground work for a copyright infringement detection system specifically for detection of illegal broadcast of copyrighted songs.

Bibliography

Baluja, S. and Covell, M. (2008). Waveprint: Efficient wavelet-based audio fingerprinting. *Pattern recognition*, 41(11):3467–3480.

Camarena-Ibarrola, A., Chvez, E., and Tellez, E. (2009). Robust radio broadcast monitoring using a multi-band spectral entropy signature. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*.

Chang, S.-F., Sikora, T., and Purl, A. (2001). Overview of the mpeg-7 standard. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6).

Colak, S. and Isik, C. (2003). Feature subset selection for blood pressure classification using orthogonal forward selection. In *Bioengineering Conference, 2003 IEEE 29th Annual, Proceedings of*, pages 122–123.

Crysand, H., Tummarello, G., and Piazza, F. (2004). Mpeg-7 encoding and processing: Mpeg7audioenc+ mpeg7audiodb. In *3rd Musicnetwork Open Workshop, Munich*.

Foote, J. (2000). Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 1.

Gauch, J. and Shivadas, A. (2005). Identification of new commercials using repeated video sequence detection. In *Image Processing,IEEE International Conference on*.

Han, W., Chan, C.-F., Choy, C.-S., and Pun, K.-P. (2006). An efficient mfcc extraction method in speech recognition. In *Circuits and Systems, 2006. ISCAS 2006. Proceedings. 2006 IEEE International Symposium on*, pages 4 pp.–.

Hasan, M., Jamil, M., and Rahman, M. (2004). Speaker identification using mel frequency cepstral coefficients. *variations*.

Herre, A., Allamanche, E., and Hellmuth, O. (2001). Robust matching of audio signals using spectral flatness features. In *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 127–130.

Hossan, M., Memon, S., and Gregory, M. (2010). A novel approach for mfcc feature extraction. In *Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on*, pages 1–5.

Jain, A. and Zongker, D. (1997). Feature selection: evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(2):153–158.

Jang, R. Audio signal processing and recognition.

JTC1/SC29/WG11, I. O. F. S. I. (2002). Coding of moving pictures and audio, mpeg-7 overview.

Jurafsky, D. and Martin, J. (2000). An introduction to natural language processing, computational linguistics, and speech recognition.

Kalker, T., Depovere, G., Haitsma, J., and Maes, M. J. (1999). Video watermarking system for broadcast monitoring.

Kim, H.-G., Moreau, N., and Sikora, T. (2004). Audio classification based on mpeg-7 spectral basis representations. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(5):716 – 725.

Lei, H., Meyer, B., and Mirghafori, N. (2012). Spectro-temporal gabor features for speaker recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4241–4244.

Li, T., Ogihara, M., and Li, Q. (2003). A comparative study on content-based music genre classification. New York. ACM.

Lin, C.-C., Chen, S.-H., Truong, T.-K., and Chang, Y. (2005). Audio classification and categorization based on wavelets and support vector machine. *Speech and Audio Processing, IEEE Transactions on*, 13(5).

Logan A., J., Goldhor, R., Goessling, D., et al. (2006). Apparatus and methods for broadcast monitoring. US Patent 7,055,166.

Matushima, R., Hiramatsu, D., Silveira, R., Ruggiero, W., da Costa, C., Monteiro, M., and Hatori, C. (2004). Integrating mpeg-7 descriptors and pattern recognition: an environment for multimedia indexing and searching. In *WebMedia and LA-Web, 2004. Proceedings*, pages 125–132.

MediaBank, P. (2012). Media monitoring system.

Milner, B. and Shao, X. (2002). Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model. In *International Conference on Spoken Language Processing (ICSLP)*. Citeseer.

Mitra, P., Murthy, C., and Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3):301–312.

Mueen, A., Nath, S., and Liu, J. (2010). Fast approximate correlation for massive time-series data. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 171–182, New York, NY, USA. ACM.

Oliveira, B., Crivellaro, A., and César, Jr, R. M. (2005). Audio-based radio and tv broadcast monitoring. In *Proceedings of the 11th Brazilian Symposium on Multimedia and the web*, New York. ACM.

PEMRA (2010). Pakistan electronic media regulatory authority, annual report.

Pitman, M., Fitch, B., Abrams, S., Germain, R., et al. (2003). System for monitoring broadcast audio content. US Patent 6,574,594.

Popescu, A., Gavat, I., and Datcu, M. (2009). Wavelet analysis for audio signals with music classification applications. In *Speech Technology and Human-Computer Dialogue, 2009. SpeD '09. Proceedings of the 5-th Conference on*, pages 1–6.

Pourhabibi, T., Imani, M. B., and Haratizadeh, S. (2011). Feature selection on persian fonts: A comparative analysis on gaa, gesa and ga. *Procedia Computer Science*, 3:1249–1255.

Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119 – 1125.

Robnik-ikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69.

Schädler, M. R. and Kollmeier, B. (2012). Normalization of spectro-temporal gabor filter bank features for improved robust automatic speech recognition systems. In *INTERSPEECH*.

Schädler, M. R., Meyer, B. T., and Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition. *The Journal of the Acoustical Society of America*, 131:4134.

Siciarz, Z. Aquila open source and cross-platform dsp (digital signal processing) library.

Sivagaminathan, R. K. and Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Syst. Appl.*, 33(1):49–60.

Spevak, C. and Polfreman, R. (2001). Sound spotting–a frame-based approach. In *International Symposium on Music Information Retrieval, Bloomington, IN, USA*.

Srinivasan, S., Petkovic, D., and Ponceleon, D. (1999). Towards robust features for classifying audio in the cuevideo system. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, pages 393–400, New York, NY, USA. ACM.

Welsh, R. (1989). Video monitoring system. US Patent 4,857,999.

Yang, J. and Honavar, V. (1998). Feature subset selection using a genetic algorithm. In Liu, H. and Motoda, H., editors, *Feature Extraction, Construction and Selection*, volume 453 of *The Springer International Series in Engineering and Computer Science*, pages 117–136. Springer US.

Zhao, H., Zhao, K., Liu, H., and Yu, F. (2012). Improved mfcc feature extraction combining symmetric ica algorithm for robust speech recognition. *Journal of multimedia*, 7(1):74–81.