DOKUZ EYLUL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

CLASSIFICATION OF SPEECH AND MUSICAL SIGNALS USING WAVELET DOMAIN FEATURES

by
Timur DÜZENLİ

July, 2010
İZMİR

CLASSIFICATION OF SPEECH AND MUSICAL SIGNALS USING WAVELET DOMAIN FEATURES

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Electrical and Electronics Engineering, Electrical and Electronics Engineering Program

by
Timur DÜZENLİ

July, 2010
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"CLASSIFICATION OF SPEECH AND MUSICAL SIGNALS USING WAVELET DOMAIN FEATURES"** completed by **TİMUR DÜZENLİ** under supervision of **ASST. PROF. DR. NALAN ÖZKURT** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Nalan ÖZKURT

Supervisor

Asst. Prof. Dr. Gülden KÖKTÜRK

(Jury Member)

Asst. Prof. Dr. Barış BOZKURT

(Jury Member)

Prof.Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

First of all, I am thankful to my supervisor, Asst. Prof. Dr. Nalan Özkurt, for her excellent guidance, support and patience to listen. I am also thankful to Dr. Hatice Doğan for her valuable comments and contributions.

I wish to extend my utmost thanks to my family for their continous support and to my friends who helped me to be patient in difficult times.

Timur DÜZENLİ

CLASSIFICATION OF SPEECH AND MUSICAL SIGNALS USING WAVELET DOMAIN FEATURES

ABSRACT

In this study, performance of wavelet transform based features for the speech / music discrimination task has been investigated. In order to extract wavelet domain features, discrete and complex wavelet transforms have been used. The performance of the proposed feature set has been compared with a feature set constructed from the most common time/frequency and cepstral domain features used in speech/music discrimination such as number of zero crossings, spectral centroid, spectral flux and Mel cepstral coefficients. In order to measure the performances of the feature sets for the speech/music discrimination, artificial neural networks have been used as classification tool. The principal component analysis has been applied to eliminate the correlated features before classification stage. Considering the number of vanishing moments and orthogonality, the best performance is obtained with Daubechies8 wavelet among the other members of the Daubechies family. According to the results the proposed feature set outperforms the traditional ones.

Keywords: speech/music discrimination, wavelet transform, Daubechies wavelet, artificial neural networks

KONUŞMA VE MÜZİK İŞARETLERİNİN DALGACIK ORTAMI ÖZNİTELİKLER KULLANARAK SINIFLANDIRILMASI

ÖZ

Bu çalışmada, müzik ve konuşma ayrımı için dalgacık dönüşümü tabanlı özniteliklerin başarımı araştırılmıştır ve zaman/frekans tabanlı öznitelikler gibi literatürde sıkça kullanılan öznitelik çıkartım yöntemleri ile karşılaştırımı yapılmıştır. Dalgacık tabanlı öznitelikleri çıkartmak için, ayrık ve karmaşık dalgacık dönüşümleri kullanılmıştır. Önerilen öznitelik setinin başarımı; sıfır geçişlerinin sayısı, izgesel merkez, izgesel akı ve mel kepstral katsayıları gibi konuşma/müzik ayrımında kullanılan en yaygın zaman/frekans ve kepstral tabanlı öznitelikler ile oluşturulmuş öznitelik seti ile karşılaştırılmıştır. Elde edilen özniteliklerin sınıflandırılmasında yapay sinir ağları kullanılmıştır. Sınıflandırma aşamasından önce birbiri ile ilişkili özniteliklerin elenmesi amacıyla temel bileşen analizi uygulanmıştır. Sönümlenen momentler ve birimdiklik değerlendirilerek, db8 dalgacığının Daubechies ailesi içindeki diğer dalgacıklardan daha yüksek başarı gösterdiği belirlenmiştir. Elde edilen sonuçlara göre, konuşma/müzik ayrımında önerilen yöntemin, önceki yöntemlere daha üstün olduğu görülmüştür.

Anahtar kelimeler: konuşma/müzik ayrımı, dalgacık dönüşümü, Daubechies dalgacığı, yapay sinir ağları

CONTENTS

Page

Page

Page

CHAPTER ONE
INTRODUCTION

Today, discrimination of speech and musical signals has been an important field due to the requirement of more efficient use of communication tools and increase in the media capabilites. The aim of a speech / music discrimination (SMD) system is to separate speech and music signals from each other by imitating the behaviour of the human ear by using efficient code and algorithms. SMD systems can be used a pre-processing stage tool for automatic speech recognition (ASR) systems, audio decoding, content based multimedia retrieval and automatic channel selection in radio broadcasts.

1.1 Speech / Music Discrimination

There have been several studies on SMD systems which use different feature extraction and classification methods. In addition, the classified material used in these studies may vary among each other.

One of the preliminary works in this area was made by J. Saunders (Saunders, 1996). In the article, a real time system that can discriminate speech and audio signals in FM radio broadcasts has been proposed. The system has been designed to change the channel when ads begin on radio broadcast. The author notes that he could manage to reach 98% as classification performance. The distribution of zero crossing rates and an algorithm based on lop-sidedness of this distribution have been used in the feature extraction stage of the study.

In another work on decomposition of recordings, a discriminator for automatic segmentation of radiophonic musical sounds has been developed using combined supervised and unsupervised methods (Richard, Ramona, & Essid, 2007). The extracted features are grouped under four titles as temporal features (ZCR, temporal statistical moments, modulation coefficients,...), Spectral features (spectral statistical moments, spectral slope, spectral flux,...), Cepstral features (MFCC, Constant Q

transform cepstral coefficients) and Perceptual features (Relative loudness, perceptive sharpness,...). These parameters are selected using a simple feature elimination program and then support vector machines (SVM) are used for classification stage. Each time frame is labelled with one of music, speech or mixed at the end of the classification. For longer segments, a smoothing procedure is defined using unsupervised approach.

In automatic speech recognition systems (ASR), it is an essential problem to de-activate the system when there is no speech signal at the input. For these types of applications, SMD systems can be used as a pre-processing tool. A system designed for this purpose given in (Scheirer & Slaney, 1997) extracts 13 features such as 4 Hz modulation energy, Percentage of low-energy frames, spectral roll off point, spectral centroid, spectral flux, zero crossing rate, cepstrum resynthesis residual magnitude and pulse metric in the fetaure extraction stage. The authors note that they have also used variances of spectral roll off point, spectral centroid, spectral flux, zero crossing rate and cepstrum resynthesis residual magnitude to form feature vector. The performance is examined in two aspects such as frame-by-frame and long segments (2.4 sec) using different classifier schemes. It is noted in the paper that the error could be decreased to 1.4% for long segment database while the classification error for frame-by-frame segments is 5.8%. The authors also add that several radio stations have been used to collect samples. This collection contains length of 20 min. recordings and each one of these recordings contains 80 samples with length of 15 sec. for each one. At classification stage, GMM, k-NN and k-d spatial classifiers have been preferred by the authors.

A speech music discriminator system designed for radio broadcasts that has been proposed in (Pikrakis, Giannakopulos, & Theodiris, 2008) uses a multilayer procedure with three–stage structure. According to this method, the aim in the first stage is to define the speech and music segments that are seperable at first glance with high accuracy. In this stage, spectral entropy and region growing based parameters are extracted. The segments which could not be classified in the first stage are segmented with more complex methods and procedures such as Dynamic

Programming and Bayesian Networks. The last stage aims to define exact boundaries of segments. The classification is performed for different music genres and the overall performance is given as 96% in the study.

Another study given in (Matsunaga, Mizuno, Othsuki, & Hayashi, 2004) aims automatically indexing of broadcast news by suggesting a new method to define audio source intervals. The process includes two stages as determination of audio sources and post processing stage for undefined segments. The three features proposed by the authors are based on spectral cross-correlation and given as spectral stability, white noise similarity and sound spectral shape. To make comparison with previous works, two different feature sets have been used by the authors. The first feature set includes energy, pitch frequency, frequency centroid and bandwitdh. In the other set, the 3 features proposed by the authors are added to four features used in first feature set. It is claimed in the paper that the performance has increased about 6.6% after addition of 3 parameters to previous ones.

One of the application fields of speech/music discriminators is audio coding. It is important to provide low bit rate – high quality sound in applications such as wireless communications, telephone, teleconference, internet communications and digital music broadcast. However, coding of music and speech utilizes different techniques in general. An effective algorithm for music coding may not be suitable and cause problems for speech coding applications. A pre-processing stage including SMD is needed to avoid these types of problems in such applications. In a study, a SMD system which minimizes the discrimination error for coding system has been proposed using a Genetic Fuzzy System (GFS) integrated to decision stage (Exposito, Galan, Reyes, & Candias, 2007). The authors state that they have avoided many classification errors and reached 94.30% accuracy using GFS and GMM classifier. Speech samples with length of one hour in total from different accents and different genders have been collected for generating speech database. One hour for recording including different genres of music such as rock, pop etc. has been used for music database.

In another study on audio coding (Rong-Yu, 1997), average zero crossing rate has been considered at feature extraction stage for non-overlapped segments with length of 480 samples. In a similar work on multimode wideband coding of speech and musical signals (Tancerel, Ragot, Ruoppila, & Lefebvre, 2000), a SMD system has been used as pre-processing tool. In the study, the discrimination is achieved by using long term statistics in feature extraction stage and GMM for classification.

SMD systems also play an important role in multimedia applications such as content based multimedia retrieval, content compression and automatic speaker indexing.

In (El-Maleh, Klein, Petrucci, & Kabal, 2000), line spectral frequencies (LSFs) and zero crossing based parameter are used for feature extraction over length of 20 msec segments. In classification stage, in order to make comparison with previous works, the labelling has been made for length of 1 sec (50 frames) using quadratic Gaussian classifier. The feature extraction over short time segments makes study convenient for real time multimedia applications. In addition, a new feature named as Linear Prediction Zero Crossing Ratio (LP-ZCR) is proposed which is calculated using proportion of the number of zero crossings at the output of a linear prediction filter to number of zero crossings at the input. For classification, two types of classifiers are used: quadratic Gaussian classifier and nearest neighbour classifier. It is noted by the authors that speech database was created by taking samples from 5 men and 5 women speakers with 8 KHz sampling frequency and for music database, music recordings with different genres were used. 28 000 frames of speech samples (9.3 min.) and 32 000 frames of music have been used as training data.

The audio content analysis plays an important role when content-based indexing and audio retrieval are concerned. In (Lu, Zhang, & Jiang, 2002), the audio content analysis is implemented. The audio classification is done using a two-stage procedure: In the first stage, KNN Classifier and a new feature based on linear spectral pairs vector quantization (LSP-VQ) is used in order to discriminate speech and non-speech segments. In second phase of classification process, the segments

labelled as non-speech in first stage are decomposed subclasses such as music, enviromental sounds and silence. A new method is proposed using quasi-GMM and LSP correlation analysis based unsupervised speaker segmentation algorithm. The classification results are addressed in many aspects in the study.

Another study on this field is given in (Zhang & Kuo, 2001), where audio content analysis is performed for online audiovisual data segmentation and classification. The audio data taken from films and TV programs is subjected to segmentation and these segments are labelled with basic classes like as speech, music, song, environmental sounds with music in the backround, speech with music in the backround and silence. The energy function, average number of zero crossings, fundamental frequency and spectral peak tracks are calculated in feature extraction stage to make the study applicable in real time operations. The authors note that they have managed to exceed 90% as classification performance.

The system proposed in (Minami, Akutsu, Hamada, & Tonomura, 1998) can be given as an example of video indexing studies. A spectrogram based analysis that aims music detection is used for video indexing. According to authors' approach, spectrogram is taken as a gray level image and classification is made using image intensity values of this spectrogram.

The gray correlation based features are used in another publication on music speech discrimination (Gong & Xiong-wei, 2006). Unlike the previous studies, amplitude of RMS value statistics based gray correlation analysis method is used for content based indexing and retrieval of cognitive media. It is stated by the authors that this method based on geometric relation of sequences with over 90% as classification performance. In analysis section, the data is divided into segments with length of 1 sec. and gray correlation analysis is performed over these segments.

In some studies, unlike their predecessors, only one feature is preferred instead of using many features (Karneback, 2001; Wang, Gao, & Ying, 2003). It is claimed in (Karnebeck, 2001) that, the main difference between music and speech is the

bandwidth. Low frequency modulation has been used as feature in the study. Waxholm database and different types of music samples from cd recordings have been used for speech and music databases, respectively.

The other method proposed in (Wang & other., 2003) uses only a new feature based on low energy ratio and this new feature is called by the authors as modified low energy ratio. It is stated in the paper that it is possible to get higher performance results than previous works using this new parameter. Authors use news broadcasts from radio and TV channels and dialogs from movies to define speech database. For music database, instrumental songs have been used. The performance results are given as 98.4% for speech and 97% for music in the paper.

For some applications including real time operations, the efficient and faster algorithms are as important as the classification results. To meet these needs; in (Wang, Wu, Deng, & Yan, 2008), a SMD system have been proposed using hierarchical oblique decision theory to provide balance between low complexity and high accuracy. In this way, they reach to 98% accuracy with a delay of 10 msec. for each frame. 228 512 frames for music and 237 671 frames for speech have been used for extraction of parameters such as normalized spectral flux between frames, normalized spectral flux between subbands, standart deviations of energy levels, energy ratio and harmonic structure ability. Authors have suggested hierarchical oblique decision classifiers which they have trained using extracted features for classification stage. It is mentioned in the paper that this method is more flexible and simpler in terms of DSP implementation and it is possible to get more accurate results. Authors add they have achieved to get a classifcation performance of 98.3%

A system working with high speed and high accuracy proposed in (Panagiotakis & Tziritas, 2005) can manage to reach 95% accuracy with 20 msec. frame delay and it is using only two characteristics of signals such as RMS based average density of zero crossings and average frequency. In classification stage, at first a decision is given for if the present frame is silence and in the next step, the classification is made for nonsilent frames to define whether they are speech or music. Any classifier is not

used for classification. Instead, the extracted features are subjected to some tests and the final decision is given by looking at the results of these tests.

It is mentioned in (Ruiz-Reyes, Vera-Candeas, Muñoz, García-Galán, & Cañadas, 2009) that the timbral feautures used in most of previous studies are not very effective for speech/music discrimination as contrary to common thought. In this publication, different from previous studies, a robust system is proposed for speech/music discrimination using fundamental frequency estimation. For classification stage, a classical statistical pattern recognition classifier followed by a fuzzy rule based system has been used. The authors have obtained the highest success rate as 97%. However, accuracy is measured as 95% for the case where all classifiers are taken into consideration.

In other published studies on speech / music discrimination, generally the feauture extraction methods show differences and these differences are also valid for classification schemes and datasets. There are studies which make comparison between other publications in terms of feature extraction. In (Carey, Parris, & Lloyd-Thomas, 1999), it is stated that 4 types of features such as amplitudes, cepstra, pitch and zero crossings are compared in the study and cepstral and delta cepstral coefficients show higher classification performance than other parameters.

Mel frequency cepstral coefficients (MFCCs) are frequently used for feature extraction stage of speech / music discrimination applications. As an example, the first degree statistics of MFCCs are examined in (Harb & Chen, 2003) to design a SMD system. Authors of the paper have noted that they have reached 96% classification performance using only a part of 80 sec. of a dataset with length of 20 000 sec. and using neural networks as classifier. It is noted in the report that the proposed method can be applied to any radio source regardless from content of data.

When other studies that use MFCC are concerned, we encounter with speech recognition and musical genre classification applications. A study on genre classification uses features including timbral features (zero crossings, centroid, roll

off, flux, MFCC), MPEG-7 features (Audio SpectrumCentroid, Audio Spectrum Spread, Audio Spectrum Flatness, Harmonic Ratio, Modified Harmonic Ratio), Rythm features (Beat Strength, Rythmic Regularity) and other features as (RMS, Time Envelope, Low Energy Rate, Loudness, Central Moments, Predictivity Ratio) (Burred & Lerch, 2003). A feature selection algorithm which compares these features among themselves is used and a 3-component Gaussian Mixture Model is preferred as classifier by the authors. The database contains 850 files with 30 sec. length for each one and the classification results are given by comparing the direct approach with the hierarchical approach proposed by the authors.

In (Ezzaidi & Rouat, 2007), the issue is addressed from a comparison aspect between statistical theory and information theory measurements in this study on musical genre classification.

Automatic speech recognition (ASR) systems for robotics are another application field of speech / music discriminators. The study in (Choi, Song, & Kim, 2007) can be given as one of the publications for these types of applications. In this paper, a speech / music discriminator for speech recognition system of a robot has been designed as pre-processing stage by the authors. Mean of minimum cepstral distances (MMCD) are used in feature extraction stage. Speech Information Technology and Industry Promotion Center (SiTec) that contains 13 hours of recordings created by 50 different male and female speakers is used for generation of speech database. RWC Music Database Subworking group of the Real World Computing Partnership (RWCP) of Japan has provided the music database as well. The authors say that they have achieved to get a success of 99.64% and emphisize that the used dataset contains speech closely recorded speech voices and original CD tracks.

One of the popular methods used in SMD systems is Discrete Wavelet Transform (DWT) (Tzanetakis, Essl, & Cook, 2001; Didiot, Illina, Fohr, & Mella, 2010; Khan & Al-Khatib, 2006; Ntalampiras & Fakotakis, 2008). When the literature is concerned in general, it is possible to see that DWT is used commonly in many

application areas of speech and audio signal processing. The study in (Tzanetakis & other. , 2001) describes some applications of DWT to the problem of extracting information from non-speech audio. The authors make an automatic classification of various types of audio using the DWT and compare with other traditional feature extraction methods proposed in the literature. Statistics over the set of the wavelet coefficients are used in order to reduce the dimensionality of the extracted feature vectors. In this way, the mean of the absolute value of the each subband, the standart deviation of the coefficients in each subband and ratios of the mean values between adjacent subbands are used for feature extraction. A window of 65536 samples at 22050 Hz sampling rate with hop size of 512 seconds (corresponds to approximately 3 seconds) is used as input to the feature extraction process and twelve levels (subbands) of coefficients are used resulting in a feature vector with 45 dimensions. Three classification experiments are evaluated in the study as MusicSpeech, Voices and Classical.

In (Khan & other. , 2006), DWT coefficients are used in feature extraction stage of a machine learning based speech / music discriminator. The mean and variance of DWT coefficients are used as input to the classification stage. The wavelet families of Haar, Meyer and two types of Daubechies (DB2 and DB15) are investigated in the paper. It is stated by the authors that extracted features using Meyer or DB15 wavelets do not contribute much to the process of classification and the results for the Haar wavelets, however, indicate that they have performed more accurate clustering than that of DB2 wavelets. The experiments were carried out using a database of music, speech, and speech added on music data in the study where all speech and speech+music data were conversational and included examples from both genders. The audio samples were extracted from documentaries and from different movies as well. The authors evaluate the results for several classifiers such as Multilayer Perceptron (MLP) Neural Networks, Radial Basis Functions (RBF) Neural Networks ve Hidden Markov Model (HMM) classifiers.

In (Didiot & other. , 2010), a wavelet based parameterization for a SMD sytem has been proposed. The authors state that DWT parameters must be preferred rather

than Fourier Transform based features for applications which use non-stationary signals like music and speech sounds. The results are evaluated for three wavelet family and numerous vanishing moments. Static, dynamic and long term parameters are investigated in the classification stage of the system.

It has been presented an effective approach which addresses the issue of speech/music discrimination using DWT in (Ntalampiras & Fakotakis, 2008). Multiresolution analysis is applied to the input signal by the authors while the most significant statistical features are calculated over a predefined texture size. For implementation, speech/music discrimination is based on six statistical measurements including mean, variance, minimum value, maximum value, standard deviation and the median taken from the low frequency information of the signal. Both male and female speech is obtained from the TIMIT database and an EBU music collection is used for music database. The classification results are obtained for 4 wavelet families given as Haar (Daubechies 1), Daubechies 4, Symlets 2 and Biorthogonal 3.7. The authors note that Haar must be used in the task of speech / music discrimination. They also add that it has demonstrated very good performance achieving 91.8% recognition rate despite the fact that the system is based solely on wavelet signal processing.

1.2 Aim of Thesis

In the literature, many successful methods including time domain, frequency domain and time/frequency domain have been proposed to be used at feature extraction stages of speech / music discrimination systems. Since it provides compact representation of signals in both time and frequency domains, discrete wavelet trnasform (DWT) stands out among other methods.

The first aim of this study is to further examine the capabilities of DWT for SMD by considering the feature extraction strategies, the properties of different wavelets and the length of the analysis window.

It is known that DWT suffers from the lack of shift invariance and oscillatory behavior. As complex wavelet transform (*C*WT) proposes an acceptable solution to these problems, it also provides compact representation for nonstationary signals. The second aim of this thesis is to observe if *C*WT is a convenient method for SMD systems by proposing a new *C*WT based parameterization system at feature extraction stage. The dual tree method which constructs approximately analytical wavelets will be used for the implementation of the *C*WT in the thesis. In order to make comparison, performance results of *C*WT and DWT based classification over other two methods such as time/fequency based features and DWT based energy features will be examined.

1.3 Outline of thesis

The thesis is organised in to 5 chapters as follows:

Chapter 2 is a detailed review of features used in the thesis. In this chapter, four different feature extraction methods are described and the advantages of proposed method is stated at the end of this section. In Chapter 3, a brief information about artificial neural networks (ANN) is given since it has been used as classification tool in the thesis. It is also mentioned about the principal component analysis (PCA) that

used for pre-classification stage. Chapter 4 is the most important section of thesis since it contains results of the experiments performed in this study. At the beginning of chapter, a detailed information on the material used in the thesis is presented and the results are examined. In the last chapter of the thesis, a comparative discussion is made about expected and encountered results. The benefits and advantages of thesis is discussed as well in this chapter.

CHAPTER TWO
FEATURES FOR SPEECH / MUSIC DISCRIMINATION

In this chapter, the related theoretical back ground of the features used in the thesis will be given.

2.1 Time/Frequency Domain Features and Mel Cepstral Coefficients

The time domain features such as number of zero crossings and frequency domain features such as low energy ratio, spectral centroid, spectral roll-off and spectral flux are commonly used for music/speech discrimination. Also, Mel frequency cepstrum coefficients are shown to be successful in music/speech classification and recognition applications. For comparison, a feature vector constructed from these features has been used for classification as the first method of this thesis.

2.1.1 Number of Zero Crossings

It is a time-domain feature which represents the number of zero crossing in a frame. It is a useful feature in music and speech discrimination since it is a measure of the dominant frequency in the signal (Saad, El-Adawy, Abu-El-Wafa, & Wahba, 2002; Scherier & other, 1997). The number of zero crossings are calculated as

$$Z_t = \frac{1}{2}\sum_{n=2}^{N}\left[\text{sgn}(x(n)) - \text{sgn}(x(n-1))\right] \qquad (2.1)$$

where $x(n)$ is the n^{th} component of the frame of length N.

2.1.2 Low Energy Ratio

This feature gives the number of the frames of where the effective or root mean square (RMS) energy is less than the average energy. The RMS energy for each frame is determined as

$$X_{RMS} = \sqrt{\frac{1}{K}\sum_{k=1}^{K} X_k^2} \qquad (2.2)$$

where X_k is the magnitude of k^{th} frequency component in the frame. Since the energy distribution is more left-skewed than for music, this measure will be higher for speech (Scherier & other, 1997).

2.1.3 Spectral Centroid

This is the measure of the center of mass of the frequency spectrum and calculated as

$$SC = \frac{\sum_{k=1}^{K} f_k X_k}{\sum_{k=1}^{K} X_k} \qquad (2.3)$$

where X_k is the magnitude of the component in the frequency band f_k (Saad & other., 2002; Scherier & other., 1997).

2.1.4 Spectral Roll-off

This feature is important in determining the shape of the frequency spectrum. The spectral roll-off point R_k is the frequency where the 95% of the spectral power lies below as summarized in

$$\sum_{k=1}^{R_k} X_k^2 = 0.95 \sum_{k=1}^{K} X_k^2 \qquad (2.4)$$

where X_t^k is the magnitude of the component of the k^{th} frequency. Since the most of the energy is in the lower frequencies for speech signals, R_k has lower values for speech (Saad & other., 2002; Scherier & other., 1997).

2.1.5 Spectral Flux:

It represents the spectral changes between adjacent frames and calculated as

$$SF_t = \sum_{k=1}^{K} \left(X_k^t - X_k^{t-1} \right)^2 \qquad (2.5)$$

where X_t^k is the k^{th} frequency component of the t^{th} frame. Then the average of the all frames are calculated. The music has a higher rate of changes than speech, thus this value is higher for music (Saad & other., 2002; Scherier & other., 1997).

2.1.6 Mel Frequency Cepstrum Coefficients (MFCC)

The Mel frequency spectrum is the linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency (Zheng, Zhang, & Song, 2001). The Mel scale is inspired from the human auditory system in which the frequency bands are not linearly spaced. Thus the sound is represented better. The calculation of the MFCC includes the following steps:

1. The discrete Fourier transform (DFT) transforms the windowed speech segment into the frequency domain and the short-term power spectrum $P(f)$ is obtained.

2. The spectrum $P(f)$ is warped along its frequency axis f (in hertz) into the mel-frequency axis as $P(M)$ where M is the mel-frequency,

$$M(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right) \qquad (2.6)$$

3. The resulted warped power spectrum is then convolved with the triangular band-pass filter $P(M)$ into $\theta(M)$. The convolution with the relatively broad critical-band masking curves $\theta(M)$ significantly reduces the spectral resolution of $\theta(M)$ in comparison with the original $P(f)$, which allows for the down sampling of $\theta(M)$.

$$\theta(M_k) = \sum_M P(M - M_k)\psi(M), \text{ k=1,…,K} \qquad (2.7)$$

Then K outputs $X(k) = \ln(\theta(M_k))$; $k = (1...K)$ are obtained. In the implementation, $\theta(M_k)$ is the average instead of the sum.

4. The MFCC are computed as

$$MFCC(d) = \sum_{k=1}^{K} X_k \cos\left[d\left((k-0.5)\frac{\pi}{K}\right)\right], \text{ k=1,…,D.} \qquad (2.8)$$

2.2 Wavelet Transform

Although it is not the most effective way of representing a signal, sometimes it is important to provide representation of a signal in terms of its spectrum or Fourier Transform. It is well known that speech and music signals contain a combination of several frequencies and they show different characteristics for different time locations. However, Fourier Transform does not show changes in the structure of frequency domain, that is, it shows only global frequency content independently from time information. In this way, if a stationary signal is in question, then Fourier Transform can be useful. For non-stationary signals, the transform must be performed locally using analysis windows (Heil & Walnut, 1989). In Figure 2.1, the representation schemes for different transformations are given. As it can be seen in (a), Fourier Transform does not perform any windowing for transformation of signal. On the other hand, in (b) and (c), STFT and Wavelet transforms use windows to analyze the signal and this property makes them an appropriate tool for processing of non-stationary signals.

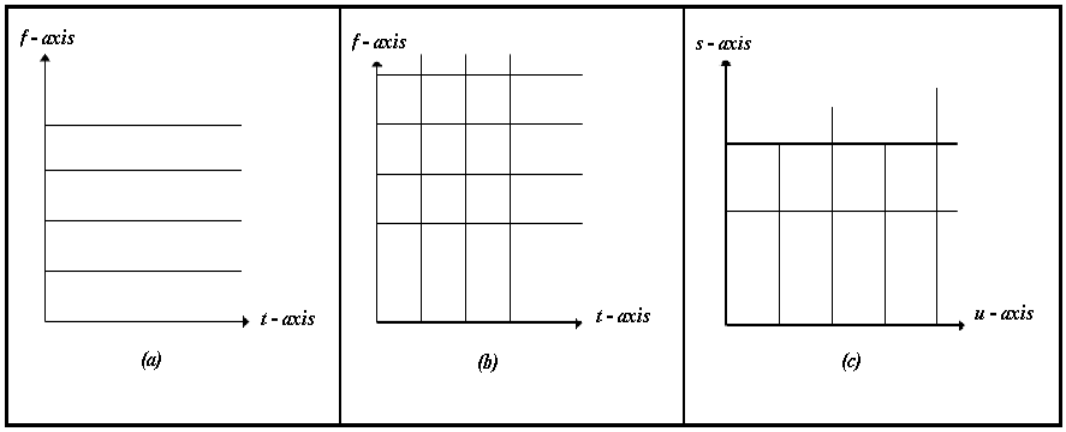


Figure 2.1 Different time-frequency representations for the three transforms: (a) Fourier Transform, (b) STFT and (c) wavelet transform (Chun-Lin, 2010)

Short time Fourier transform (STFT) and wavelet transform (WT) can be given as examples for methods that use windows to analyse the signals locally. STFT use costant length windows for analysis and this sometimes causes problems in terms of representation. WT uses windows which can scale their sizes adaptively to provide good resolution in time and frequency domain. Both STFT and WT use the correlation between the signal and analysis function (Chun-Lin, 2010). As it is shown in Figure 2.2, continuous wavelet transform is performed using translated and scaled versions of a mother wavelet. The transformation is represented for two different scaling values such as s=5 and s=20.

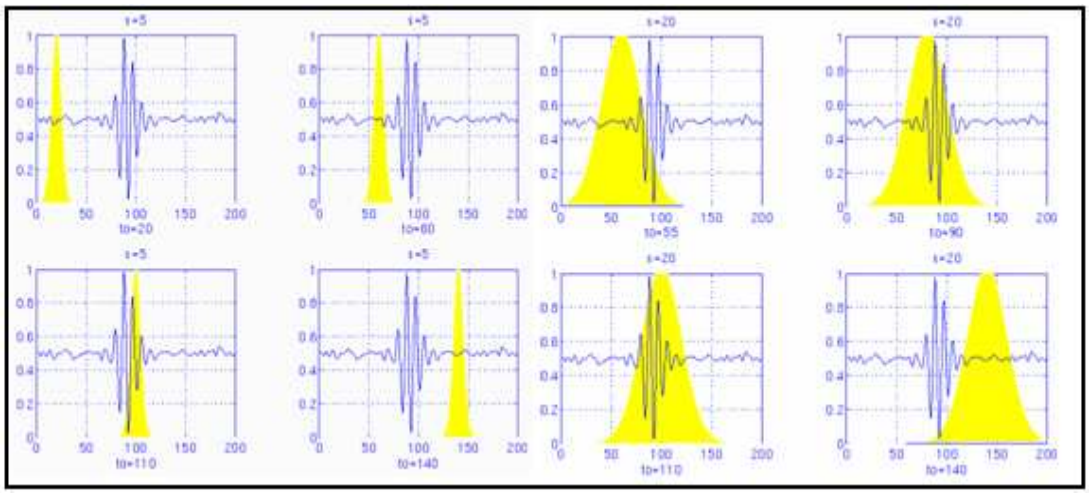


Figure 2.2 Continous wavelet transform for a non-stationary signal for different scaling parameters. (Sumbera, 2001)

To perform Contionuos Wavelet Transform, the convolution between the signal and analysis function is calculated as analogous to Fourier transform. The only difference between two methods is that wavelets are used instead of sinusoids in wavelet transform. Wavelets are functions which oscillate locally and they are limited in time domain. Wavelet functions contain parameters which allow to shifting and scaling of windows and in this way, they provide a better resolution both in time and frequency domain than STFT (Merry, 2005).

Another implementation for wavelet transforms is performed with filter banks and is named as Discrete Wavelet Transform (DWT). DWT subjects a signal to some filtering process using filter banks and decompose it to coefficients called as detail and approximation. These coefficients provides a good representation of signals with giving frequency information and time location of that frequency component.

2.2.1 The Continous Wavelet Transform

A mother wavelet function limited in time domain $\psi(t) \in L^2(R)$ is defined where limited in time domain refers to taking values in a limited region over time axis. These wavelets are normalized and also have zero mean property (Chun-Lin, 2010).

Mathematically, these properties are given as

$$\int_{-\infty}^{\infty} \psi(t)dt = 0$$
$$\|\psi(t)\|^2 = \int_{-\infty}^{\infty} \psi(t)\psi^*(t)dt = 1 \qquad (2.9)$$

The mother wavelet has the capability of forming the basis set denoted as

$$\left\{ \psi_{s,u}(t) = \frac{1}{\sqrt{s}} \psi\left(\frac{t-u}{s}\right) \right\}, u \in R, s \in R^+ \qquad (2.10)$$

where u and s are translating and scaling parameters, respectively. The translating parameter in the equation shows the region that is being analyzed. $\{\psi_{u,s}(t)\}$ is obtained orthonormally which is ensured by multiresolution property.

It is possible to map a one dimensional signal $f(t)$ to the two dimensional coefficients $Wf(s,u)$ that contain time and frequency information using this transform. These two parameters are used to locate a certain frequency (scaling parameter s) at a particular time instant (translating parameter u).

Continous wavelet transform is given as

$$Wf(s,u) = < f(t), \psi_{s,u} > \qquad (2.11)$$

$$= \int_{-\infty}^{\infty} f(t)\psi^{*}_{s,u}(t)dt$$

$$= \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{s}}\psi^{*}(\frac{t-u}{s})dt$$

The inverse continous wavelet transform is given as

$$f(t) = \frac{1}{C_{\psi}}\int_{0}^{\infty}\int_{-\infty}^{\infty} Wf(s,u)\frac{1}{\sqrt{s}}\psi\left(\frac{t-u}{s}\right)du\frac{ds}{s^2} \qquad (2.12)$$

where C_{ψ} is defined as

$$C_{\psi} = \int_{0}^{\infty}\frac{|\psi(w)|^2}{w}dw < \infty \qquad (2.13)$$

This equation is also called the admissibility condition where $\psi(w)$ is the Fourier transform of the mother wavelet $\psi(t)$ (Chun-Lin, 2010).

Continuous wavelet transform is calculated by taking discrete samples for the scaling parameter s and translation parameter u and the resulting wavelet coefficients are called wavelet series (Merry, 2005).

Wavelet series can be calculated as

$$Xwt_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt \quad \text{with } \psi_{m,n} = s_0^{-m/2}\psi(s_0^{-m}t - nu_0) \qquad (2.14)$$

where integers m and n control the wavelet dilatation and translation.

2.2.2 The Discrete Wavelet Transform

The continous wavelet transform uses functions that contain parameters such as translating and scaling to make multiresolution analysis. However, DWT performs this analysis by using multiresolution filter banks and specific wavelet filters (Merry, 2005).

2.2.2.1 Filter Banks

Filter banks refer to collection of filters which decompose the signals into different frequency bands. The discrete signals are applied to analysis filter bank and decomposed to their frequency components filtering by $L(z)$ and $H(z)$, low-pass and high-pass filters, respectively. The outputs of the filters represent the same frequency content with input by coming together, but the amount of samples are doubled. So, the outputs of filters in analysis filter bank are subjected to downsampling by a factor 2.

The signals are upsampled by a factor 2 as contrary to analysis filter bank and passed through the synthesis filters $L_0(z)$ and $H_0(z)$ in reconstruction process. Summing of outputs of these synthesis filters yields the reconstructed signal $y[k]$ as given in Figure 2.3.

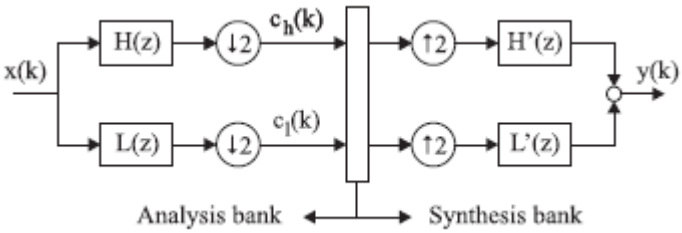


Figure 2.3 Two channel filter bank (Merry, 2005)

2.2.2.2 Perfect reconstruction

The filter banks should be biorthogonal to satisfy perfect reconstruction property (Merry, 2005). To ensure satisfying of this property, aliasing and distortion must be prevented by some design criteria (Strang & Nguyen, 1997). In the two channel filter bank given in Figure 2.3, the signal is decomposed into two frequency bands using low-pass $L(z)$ and high-pass $H(z)$ filters. There will not be loss of information if the filters have sharp-edge structure, however, it is not possible to implement these types of filters in practice since always a transition band exists. This case causes amplitude and phase distortion in each of the channels (Schneiders, 2001). For a two channel filter bank, aliasing can be avoided by designing the filters of the synthesis filter bank as (Strang & other., 1997)

$$\begin{aligned} L'(z) &= H(-z) \\ H'(z) &= -L(-z) \end{aligned} \qquad (2.15)$$

A product filter $P_0(z) = L'(z)L(z)$ is defined to prevent distortion. This distortion can be tackled if (Schneiders, 2001)

$$P_0(z) - P_0(-z) = 2z^{-N} \qquad (2.16)$$

In the equation, N is given as the overall delay in filter banks.

The perfect reconstruction filter bank can be designed in two steps:

1. A low-pass filter P_0 satisfying the equation given above is designed.

2. $P_0(z)$ is factored into $L'(z)L(z)$ and $H'(z)$ and $H(z)$ are calculated using equations given above.

2.2.2.3 Multiresolution Filter Banks

In previous section, a two channel decomposition has been presented which uses low-pass and high-pass filters that give approximation and detail coefficients at their outputs, respectively. A three-level filter bank is shown in Figure 2.4.

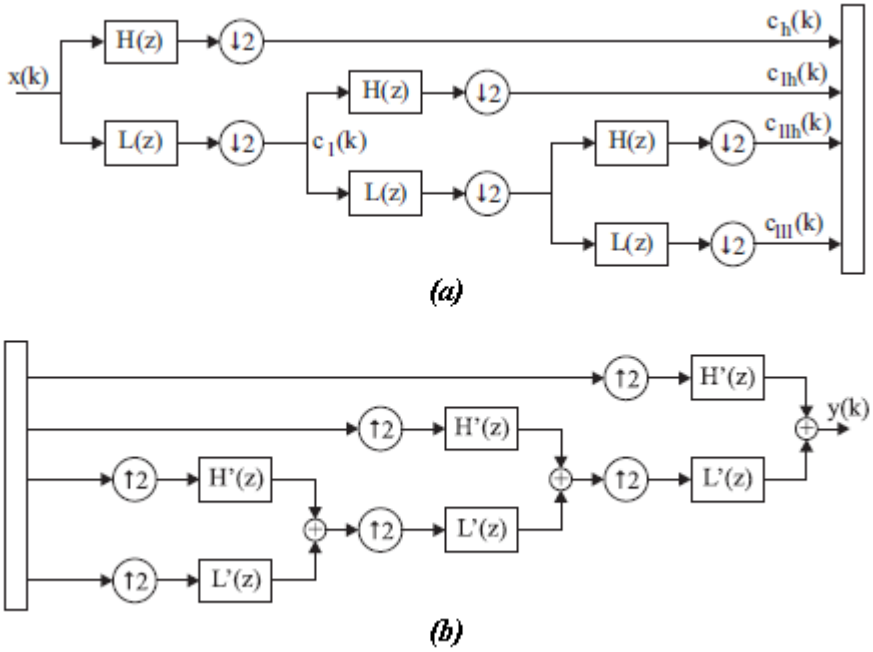


Figure 2.4 Tree level filter bank: *(a)* analysis bank *(b)* synthesis bank (Merry, 2005)

As it can be seen, the filter bank can be designed depending on the desired resolution. $c_l(k)$ are the coefficients that represent the lowest half of the frequency content of the frequencies in $x[k]$ and $c_h(k)$ coefficients are vice versa. It should not

be forgotten that the downsampling operation by factor 2 is performed after each filter.

After each level, highest and lowest frequency components are represented by the outputs of high-pass and low-pass filter outputs. As mentioned before, the level of filtering can be increased or decreased arbitrarily depending on the desired resolution. For a special set of filters $L(z)$ and $H(z)$, this structure is called as DWT and the filters are named as wavelet filters (Merry, 2005).

2.2.2.4 Vanishing moments

The vanishing moment represents how a function decays toward infinity (Chun-Lin, 2010). For example, the function $\cos t / t^2$ decays at a rate of $1/t^2$ as t approaches to infinity. The estimation of rate of decay is performed by the integration,

$$\int_{-\infty}^{\infty} t^k f(t)dt \qquad (2.17)$$

The parameter k in the integration shows the rate of decay. It is said that the wavelet function $\psi(t)$ has p vanishing moments if

$$\int_{-\infty}^{\infty} t^k \psi(t)dt = 0 \text{ for } 0 \le k \le p \qquad (2.18)$$

2.2.2.5 The Fundamental Wavelet Families

Wavelet transforms contain an infinite set of several wavelet types. Selection of different wavelets exists different characteristics such as how smooth they are and whether they provide a good representation in time / frequency domain (Graps, 1995).

Daubechies Wavelets are the wavelets which have been designed for a given vanishing moment p and minimum size discrete filter. In these types of wavelets, if it is asked to use a wavelet function with p vanishing moments, the minimum filter size will be length of 2p (Chun-Lin, 2010).

Within each family of wavelets (such as the Daubechies family), wavelet subclasses are defined by the number of coefficients and by the level of iteration. Number of vanishing moments are also essential in terms of classification of wavelets within a family. For example, the wavelets within the Daubechies wavelet family are divided into subclasses according to number of vanishing moments (Graps, 1995). Some examples of the wavelet family members are shown in Fig. 2.5. The number of next to the wavelet name represents the number of vanishing moments in the figure.

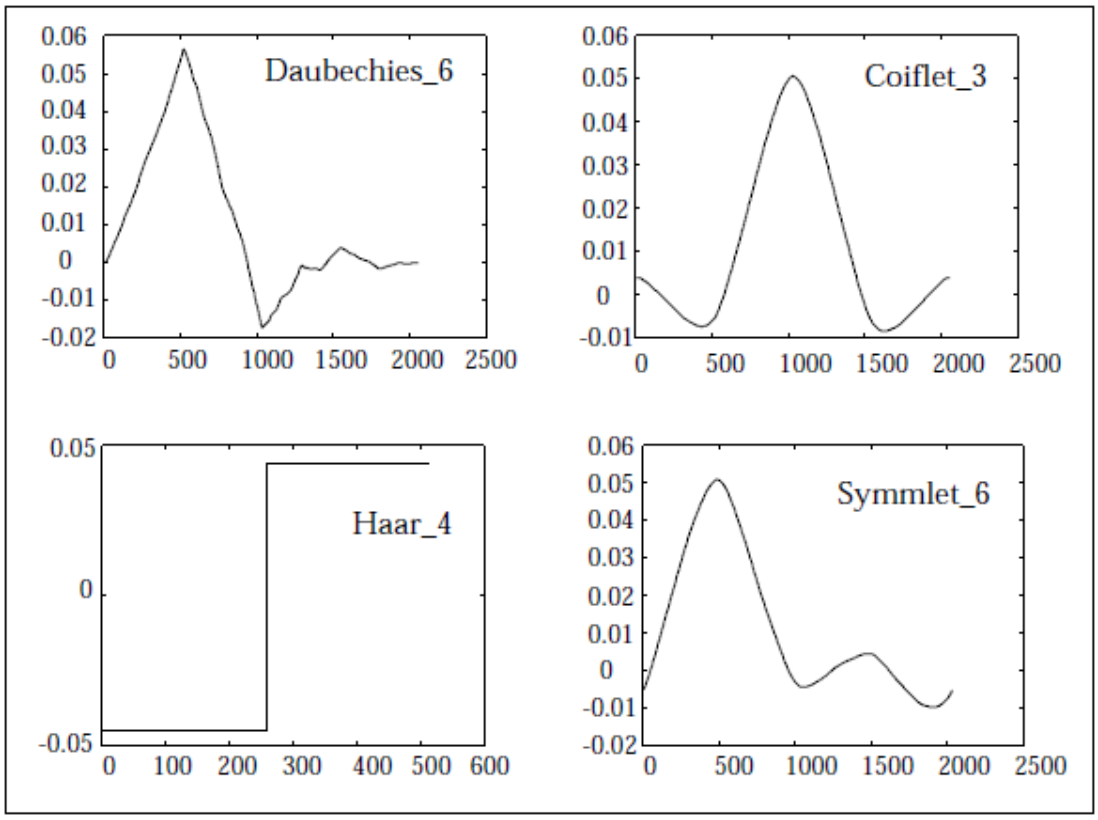
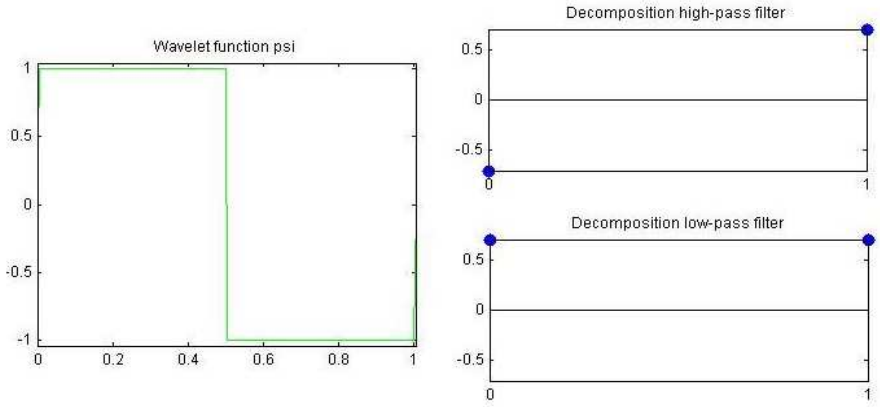
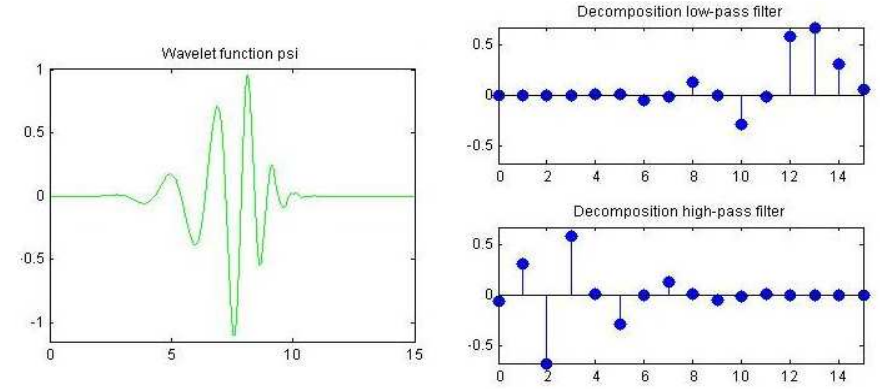


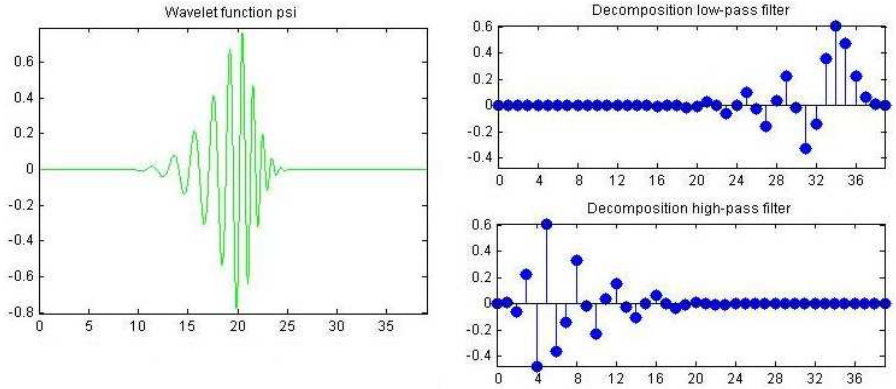
Figure 2.5 Several different families of wavelets (Graps, 1995).



(a)



(b)



(c)

Figure 2.6 The wavelet functions with low pass and high pass filter coefficients for (a) Haar, (b) Daubechies8 and (c) Daubechies20

2.3 Wavelet Transform Based Energy Features

In study of Didiot & other. (2010), it has been talked about the energy based features which are calculated using wavelet transform. According to study, the energy distribution in each frequency band is a very relevant acoustic cue and energy, calculated from DWT, can be used as a speech/music discrimination feature. In our study, these energy based parameters have also been used in order to make comparison among different feature extraction methods.

2.3.1 Instantaneous Energy

This is a feature which gives the energy distribution in each band and given as:

$$f_j^E = \log_{10}\left(\frac{1}{N_j}\sum_{r=1}^{N_j}(w_j(r))^2\right) \qquad (2.19)$$

where $w_j(r)$ is the wavelet coefficient at time position r and frequency band j and N is the length of the analysis window.

2.3.2 Teager Energy

Teager Energy has been recently applied for speech recognition and given as:

$$f_j^{T_E} = \log_{10}\left(\left|\frac{1}{N_j}\sum_{r=1}^{N_j\text{-}1}\left(w_j(r)\right)^2 - \left(w_j(r\text{-}1)*w_j(r+1)\right)\right|\right) \qquad (2.20)$$

It is said that the discrete Teager Energy Operator (TEO), allows modulation energy tracking and gives a better representation of the formant information in the feature vector compared to MFCC in (Didiot & other., 2010). It is also pointed out that the Teager energy is a noise robust parameter for speech recognition because the effect of additive noise is attenuated.

2.4 Complex Wavelet Transform

2.4.1 Introduction

In previous section, a detail explanation has been presented about DWT and important points of DWT based feature extraction has been mentioned. One of the properties which makes DWT so essential is getting information which cannot be provided by Fourier Transform. DWT allows to expression of signals without losing information about location in time domain and it provides an optimal representation for signals including sudden transitions like jumps and spikes. In this way, DWT is often used in applications such as image processing, speech processing, statistical signal processing for noise removing, signal modeling and compression. However, although all these advantages of DWT, it has some shortcomings which makes complex wavelet transform superior than DWT. In these section, the shortcomings of DWT based analysis and how *C*WT overcomes these problems will be examined.

2.4.1.1 Oscillations

As previously mentioned, since wavelets are band-pass and time-limited functions, they exhibit oscillatory behaviour around singularities. This behaviour makes difficult to extract singularities and analysis with wavelet based modeling. Wavelet coefficients take high values in parts containing singularities.

2.4.1.2 Shift Variance

One of the disadvantages of DWT is its sensivity to a small shift of the signal in time domain. This situation leads to problems in DWT based analysis. The designed algorithm must be capable of coping with high valued DWT coefficients caused by shifted singularities.

2.4.1.3 Aliasing

DWT coefficients are obtained with dowsampling operations between non-ideal low pass and high pass filters and this process cause aliasing problems. Although the inverse DWT can eliminate this problem, wavelet and scaling coefficients should not be changed in order to do this elimination and in addition, artifacts in reconstructed signal cause loss of balance between forward and inverse DWT transforms.

2.5.1.4 Lack of Directionality

This problem emerges particularly in image processing applications. It makes difficult to process edges and corners in 2 or higher dimensional signals.

In (Selesnick & other. , 2005), it is said that Fourier transform can overcome these problems and it can be given as a solution. It is possible to see a smooth representation has been provided and there aren't positive and negative oscillations in frequency domain when the amplitude of Fourier transform is concerned. The amplitude of FT is not affected from any shifts in the signal as well and also, FT does not experienced with aliasing and lack of directionality problems. The biggest difference between FT and DWT can be seen by looking at decomposition methods of these two transforms. FT decompose the signals into complex valued sinusoids differently from DWT's real valued wavelets.

$$e^{jwt} = \cos(\Omega t) + j.\sin(\Omega t) \qquad (2.21)$$

Since there is a phase difference of 90° between cos and sin, these two elements form a Hilbert Transform pair by coming together. The analytical signal formed by this pair provides a one-sided spectrum in frequency domain.

Complex Wavelet Transform (*C*WT) has been proposed inspiring by the Fourier Transform which does not suffer from these types of problems. *C*WT is defined with a complex-valued scaling function and complex-valued wavelet

$$\Psi_c(t) = \Psi_r(t) + j\Psi_i(t) \qquad (2.22)$$

where $\Psi_r(t)$ and $\Psi_i(t)$ are real and imaginary parts of the complex wavelet $\Psi_c(t)$. If these functions are $90°$ out of phase with each other, that is, if they form a Hilbert Transform pair, then $\Psi_c(t)$ becomes analytic signal and it has a one-sided spectrum. Projecting the signal onto $2^j\psi_c(2^j t - n)$, the complex wavelet coefficients are obtained as

$$d_c(j,n) = d_r(j,n) + jd_i(j,n) \qquad (2.23)$$

Complex Wavelet Transform can be performed in two class. In first one, a complex wavelet $\Psi_c(t)$ that forms an orthonormal or biorthogonal basis is searched. The second method seeks a redundant representation and it searches $\Psi_r(t)$ and $\Psi_i(t)$ that provide orthonormal and biorthogonal bases individually. Resulting *C*WT has 2x redundancy in 1-D and has power to overcome the shortcomings of DWT. In this thesis, the dual-tree approach for performing complex wavelet transform which is a natural approach to second, redundant type has been preferred.

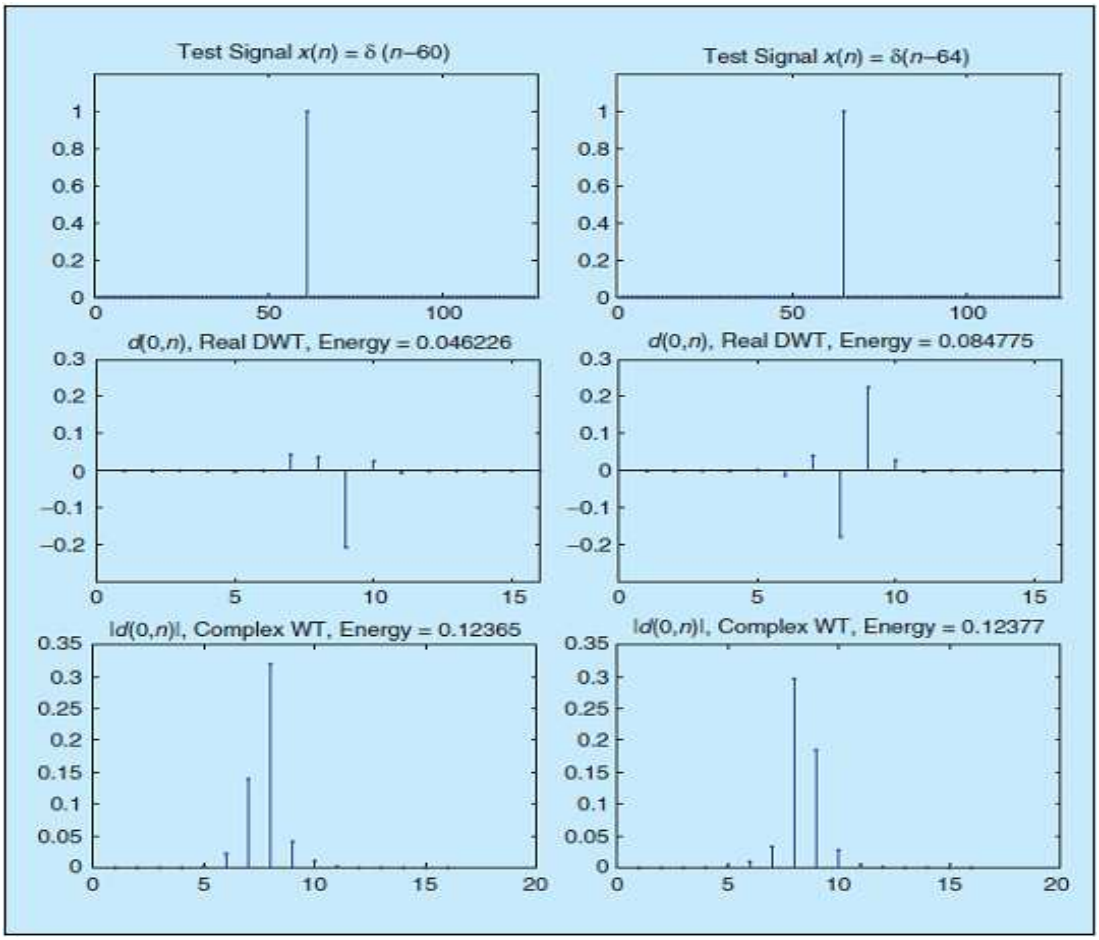


Figure 2.7 Sensitivity of DWT and *C*WT coefficients to shiftings in time domain (Selesnick & other. , 2005)

In Figure 2.7, it is possible to see that DWT coefficients are very sensitive to any shift in time domain while CWT coefficients are not. For two impulse signals $x(n) = \delta(n - 60)$ and $x(n) = \delta(n - 64)$, the real coefficients of conventional real discrete wavelet transform (with Daubechies length-14 filters) and magnitude of the complex coefficients of the dual-tree complex wavelet transform are shown in the figure.

2.4.2 Dual-Tree Complex Wavelet Transform (DT-CWT)

Dual-Tree Complex Wavelet Transform was first introduced by Kingsbury in 1998 (Kingsbury, 1998). The dual tree implements an analytic wavelet transform by using two real discrete wavelet transform with two filterbank trees; the first DWT gives the real and the second one gives the the imaginary part of the *C*WT. Analysis and synthesis filter banks can be illustrated as in the Figure 2.8 where $h_0(n)$ and $h_1(n)$ denote the lowpass/ high-pass filter pair for the upper filterbank which implements WT for real part. In the same way, $g_0(n)$ and $g_1(n)$ denote the low-pass / high-pass filter pair for the lower filterbank for imaginary part. In this approach, the key challenge is joint design of two filterbanks to get complex wavelet and scaling function as close as possible to analytic (Selesnick & other. , 2005).

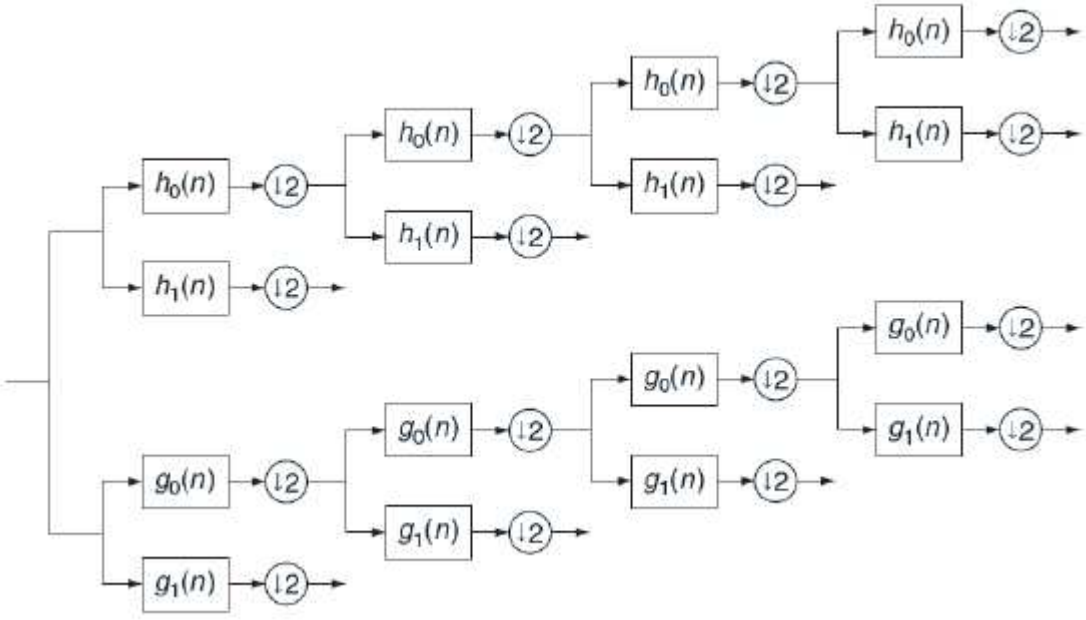


Figure 2. 8 Analysis filter bank for the dual tree *C*WT (Selesnick & other. , 2005)

The filters used for real and imaginary parts of the transform must satisfy the perfect reconstruction condition given as

$$\sum_n h_0(n)h_0(n+2k) = \delta(k)$$
$$h_1(n) = (-1)^n h_0(M-n) \qquad (2.24)$$

Two low pass filters of dual tree $h_0(n)$ and $g_0(n)$ satisfying a very simple property makes corresponding wavelets to form an approximate Hilbert Transform pair: One of them must be approximately a half- sample shift of the other (Selesnick, 2001)

$$g_0(n) = h_0(n-0.5) \Rightarrow \psi_g(t) = \mathrm{H}\{\psi_h(t)\} \qquad (2.25)$$

Since $h_0(n)$ and $g_0(n)$ are defined only on integers, it will be useful to rewrite the half-sample delay condition in terms of magnitude and phase functions separately in frequency domain to make the statement rigorous:

$$\begin{aligned} &\left|G_0(e^{jw})\right| = \left|H_0(e^{jw})\right| \\ &\angle G_0(e^{jw}) = \angle H_0(e^{jw} - 0.5w) \end{aligned} \qquad (2.26)$$

There are two popular methods for design of filters for DT-CWT (Selesnick & other., 2005):

2.4.2.1 Q-Shift Solution

According to q-Shift solution, $g_0(n)$ must be selected as

$$g_0(n) = h_0(N-1-n) \qquad (2.27)$$

where N is the length of filter $h_0(n)$ and is even. In this case the magnitude condition in *2.25* is satisfied but not the phase condition.

$$\begin{aligned} &\left|G_0(e^{jw})\right| = \left|H_0(e^{jw})\right| \\ &\angle G_0(e^{jw}) \neq \angle H_0(e^{jw} - 0.5w) \end{aligned} \qquad (2.28)$$

The quarter-shift (q-shift) solution has an interesting property that causes to take its name: When you ask that $g_0(n)$ and $h_0(n)$ be related as $g_0(n) = h_0(N-1-n)$ and

also that they approximately satisfy $\angle G_0(e^{jw}) = \angle H_0(e^{jw} - 0.5w)$, then it turns out that the frequency response of $h_0(n)$ has approximately linear phase. This is verified by writing $g_0(n) = h_0(N-1-n)$ in terms of Fourier transforms

$$G_0(e^{jw}) = H_0^*(e^{jw})e^{-j(N-1)w} \qquad (2.29)$$

where the * represents complex conjugation. This implies that the phases satisfy

$$\angle G_0(e^{jw}) = -\angle H_0(e^{jw}) - (N-1)w \qquad (2.30)$$

If the two filters satisfy the phase condition approximately, it can be written that

$$\angle H_0(e^{jw}) - 0.5w = -\angle H_0(e^{jw}) - (N-1)w \qquad (2.31)$$

And we have the equation,

$$\angle H_0(e^{jw}) \approx -0.5(N-1)w + 0.25w \qquad (2.32)$$

As it can be seen, $h_0(n)$ is an approximately linear-phase filter. This means that $h_0(n)$ is approximately symmetric around the point $n = 0.5\ (N-1) - 0.25$. This is one quarter away from the natural point of symmetry and solutions of this kind were introduced as q-shift dual-tree filters for this reason (Selesnick & other., 2005).

2.4.2.2 Common Factor Solution

Another method for filter design stage named as Common Factor Solution (CFS) can be used to design both orthonormal and biorthogonal solutions for the Dual Tree *C*WT (Selesnick, 2001).

In this approach, the filters, h_0 and g_0 are set as,

$$h_0(n) = f(n) * d(n) \qquad (2.33)$$

$$g_0(n) = f(n) * d(L-n) \qquad (2.34)$$

where $d(n)$ is supported on $0 \le n \le L$ and * represents the discrete time convolution. In terms of Z-transform, we have

$$H_0(z) = F(z)D(z) \qquad (2.35)$$

$$G_0(z) = F(z)z^{-L}D(1/z) \qquad (2.36)$$

In this kind of solution, the magnitude part of half - sample delay condition is satisfied; however, the phase part is not exactly satisfied as in q-shift solution (Selesnick & other., 2005).

$$\left|G_0(e^{jw})\right| = \left|H_0(e^{jw})\right| \qquad (2.37)$$

$$\angle G_0(e^{jw}) \neq \angle H_0(e^{jw}) - 0.5w \qquad (2.38)$$

So, we must design the filters so that the phase condition is approximately satisfied. Using the equations,

$$H_0(z) = F(z)D(z) \qquad (2.39)$$

$$G_0(z) = F(z)z^{-L}D(1/z) \qquad (2.40)$$

we can say,

$$G_0(z) = H_0(z)A(z) \qquad (2.41)$$

where

$$A(z) = \frac{z^{-L} D(1/z)}{D(z)} \qquad (2.42)$$

$A(z)$ is an all-pass transfer function; the magnitude of $A(z)$ is $\left|A(e^{jw})\right| = 1$. Then, from the equation

$$G_0(z) = H_0(z)A(z) \qquad (2.43)$$

we have

$$\left|G_0(e^{jw})\right| = \left|H_0(e^{jw})\right| \qquad (2.44)$$

and

$$\angle G_0(e^{jw}) = \angle H_0(e^{jw}) + \angle A(e^{jw}) \qquad (2.45)$$

As it can be seen easily, for satisfaction of phase property, the $D(z)$ must be chosen so that

$$\angle A(e^{jw}) \approx -0,5w \qquad (2.46)$$

With this result, it can be said that $A(z)$ should be a fractional delay all-pass system (Selesnick, 2001).

$D(z)$ can be defined by adapting Thiran's formula for maximally flat delay allpole filter (Thiran, 1971) to maximally flat delay all pass filter.

$$D(z) = 1 + \sum_{n=1}^{L} d(n) z^{-n} \qquad (2.47)$$

with

$$d(n) = (-1)^n \binom{L}{n} \frac{(\tau - L)_n}{(\tau + 1)_n} \qquad (2.48)$$

where $(x)_n$ represents the rising factorial

$$(x)_n := (x)(x+1)(x+2)....(x+n+1) \qquad (2.49)$$

With this $D(z)$, we have the approximation

$$A(z) \approx z^{-\tau} \text{ around } z = 1 \qquad (2.50)$$

or equivalently,

$$A(w) \approx e^{-jw\tau} \text{ around } w = 0 \qquad (2.51)$$

The coefficients of $d(n)$ can be computed easily using the ratio (Selesnick, 2001)

$$\frac{d(n+1)}{d(n)} = -\frac{\binom{L}{n+1}}{\binom{L}{n}} \frac{(\tau - L)_{n+1}}{(\tau - L)_n} \frac{(\tau + 1)_n}{(\tau + 1)_{n+1}} = \frac{(L-n)(L-n-\tau)}{(n+1)(n+1+\tau)} \qquad (2.52)$$

Using this ratio, the filter $d(n)$ can be generated as follows:

$$d(0) = 1$$
$$d(n+1) = d(n) \frac{(L-n)(L-n-\tau)}{(n+1)(n+1+\tau)}, \quad 0 \le n \le L-1 \qquad (2.53)$$

The second step, finding $F(z)$ so that $h_0(n)$ and $g_0(n)$ satisfy the PR conditions, requires only a solution to a linear systems of equations and a spectral factorization.

To obtain wavelet bases with K vanishing moments, we let

$$F(z) = Q(z)(1+z^{-1})^K \qquad (2.54)$$

So,

$$H_0(z) = Q(z)(1+z^{-1})^K D(z) \qquad (2.55)$$

$$G_0(z) = Q(z)(1+z^{-1})^K z^{-L} D(1/z) \qquad (2.56)$$

$Q(z)$ of minimal degree is obtained using a spectral factorization approach. The procedure consists of two steps (Selesnick, 2001).

1) $r(n)$ is found with minimal length such that

   a) $r(n) = r(-n)$

   b) $R(z)(z+2+z^{-1})^K D(z)D(1/z)$ is halfband.

2) $Q(z)$ is set to be a spectral factor of $R(z)$

$$R(z) = Q(z)Q(1/z) \qquad (2.57)$$

The first step can be carried out by solving only a system of linear equations. By defining

$$S(z) := (z+2+z^{-1})^K D(z)D(1/z) \qquad (2.58)$$

the half band condition can be written as

$$\delta(n) = \left[\downarrow 2\right](s*r)(n) = \sum_k s(2n-k)r(k) \qquad (2.59)$$

The second step assumes $R(z)$ permits spectral factorization.

With $Q(z)$ obtained in this way, the filters $H_0(z)$ and $G_0(z)$ satisfy the PR conditions and have desired half-sample delay.

Using this design procedure, the filters $h_0(n)$ and $g_0(n)$ of (minimal length) $2(L+K)$ are defined. K and L are the number of zeros at $z=-1$ and degree of fractional delay, respectively. (Selesnick, 2001)

As it can be seen, the design procedure allows for an arbitrary number of vanishing wavelet moments to be specified. In Figure 2.9, filter coefficients obtained by common factor solution is shown . It can be seen from the figure that the complex wavelet defined by real and imaginer components has an approximately one-sided spectrum as referring to it is an approximately analytical signal.

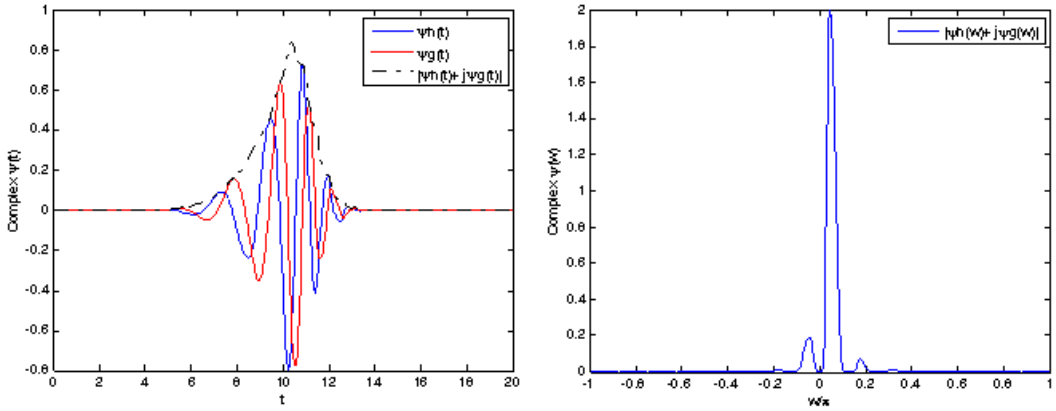


Figure 2. 9 Aproximate Hilbert Transform Pair of orthonormal wavelet bases with N = 20, K = 5, L = 5 (Selesnick, 2001).

CHAPTER THREE
ARTIFICIAL NEURAL NETWORKS AND PRINCIPAL COMPONENT ANALYSIS

3.1 Artificial Neural Networks

An Artificial Neural Network (ANN) is a tool that aims to solve problems by imitating the mental calculations which are specific to human brains. A human brain contains small computing units named as “neurons” that can perform very simple calculations. Neurons have the ability of building networks that can operate in paralel to solve more difficult problems (Roy, 2000). These networks allow to paralel implementations for nonlinear static or dynamic systems. Also they have a very important feature such that their adaptive nature replaced programming with learning by example to solve complex problems. This feature makes these networks very attractive in application domains where one has little or incomplete understanding of the problem to be solved but where training data is readily available. The most widely used learning algorithm in ANNs is the Backpropagation Algorithm (Jha, 2003). There are various types of ANNs which use this algorithm such as Multilayered Perceptron, Radial Basis Function and Kohonen Networks.

ANNs have been used for a wide variety of applications where statistical methods such as discriminant analysis, logistic regression, Bayes analysis, multiple regression and ARIMA time-series model are traiditionally employed (Jha, 2003). It has been mentioned by Haykin (1999) that there are several benefits of ANNs including nonlinearity, input-output mapping, adaptivity, evidental response, fault tolerance and so on. In this regard, ANNs are considered a powerful tool for data analysis and classification.

3.1.1 Architecture of an artificial neuron

The most simple procedure performed by a neuron can be expressed in the form of $y_i=f(z_i)$ in general. Here y_i, z_i and f represent the output of i^{th} neuron, input of the i^{th} neuron and a non-linear function, respectively. The nonlinear function f, also called a node function, takes different forms in different models of the neuron; a typical choice for the node function is a step function or a sigmoid function (Roy, 2000). The neurons get their input signals from other neurons or from external sources such as various organs of the body like the eyes, the ears and the nose. The output signal from a neuron may be sent to other neurons.

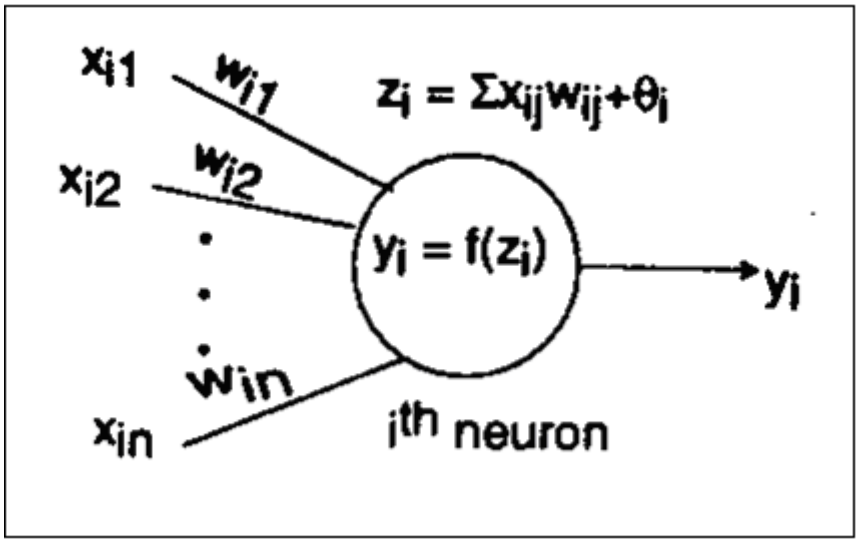


Figure 3.1 Architecture of an artificial neuron (Roy, 2000)

In Figure 3.1, an artificial neuron structure is given with its inputs, weights and output. It is possible to see in the figure that how an artifical neuron defines its output.

3.1.2 Multilayered Artificial Neural Networks

The neurons can form large scale ANN architectures by coming together and connecting among themselves. The basic architecture includes three types of neuron layers: input, hidden, and output layers. In feed-forward networks, the signal flow is from input to output units, strictly in a feed-forward direction (Abraham, 2005). In

Figure 3.2, a feed-forward multilayered network is shown with three basic layers such as input, hidden and output. The weights are represented as connections between each layer in the figure.

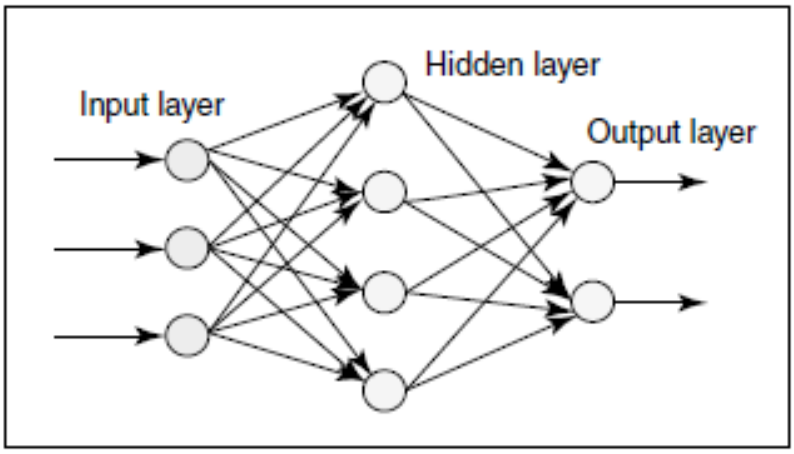


Figure 3.2 Multilayered Artificial Neural Network (Abraham, 2005)

3.1.3 Learning Algorithms for Neural Networks

Before classification process, an ANN must be configured in order to produce desired outputs for a given set of inputs. There are several methods to strength the connections of the weights such as setting the them explicitly or training the neural network by feeding it with training patterns and changing its weights according to some learning rule. The learning methods in neural networks can be classified as three types. These are given as supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, an input vector is presented at the inputs together with a set of desired responses, one for each node, at the output layer. A forward pass is performed, and the errors between the desired and actual response for each node in the output layer are found. The weight changes are determined using these errors according to learning rule in use. The desired outputs of distinct output nodes are provided by an external teacher in supervised learning method. The backpropagation algorithm, the delta rule, and the perceptron rule are the best-known examples of this technique (Abraham, 2005).

A perceptron is a single layer neural network and its weights and biases could be trained to produce a correct target vector when presented with the corresponding input vector. The training technique used is named as the *perceptron learning rule*. Perceptrons are especially suited for simple problems in pattern classification. Training procedure of a perceptron contains four essential points. According to this learning rule, initially random weights are used for connections and training samples have been applied to perceptron. The output of the network is obtained with these existing weights and if the output of the network does not match with the desired output, the weights are updated according to rule given as

$$w_{ij}(t) = w_{ij}(t-1) + \Delta w_{ij}(t) \qquad (3.1)$$

where

$$\Delta w_{ij}(t) = \eta(d_k - y_k)x_i \qquad (3.2)$$

η is the learning rate, d_k is desired output, y_k is output of the perceptron and x_i is the input of the network in the equation.

Perceptron learning rule is similar to Hebbian learning. The only difference is that when the network responds correctly, no connection weights are modified. On the other hand, Hebbian learning continually strengthens its weights without bound (Abraham, 2005).

In backpropagation algorithm, the weights are updated by taking the partial derivative of the error of the network with respect to each weight. The learning rule for backpropagation algorithm is given as

$$\Delta w_{ij}(t) = -\eta \frac{\partial E}{\partial w_{ij}(t)} + \alpha \Delta w_{ij}(t-1) \qquad (3.3)$$

where η and α are the learning rate and momentum respectively.

The momentum term determines the effect of past weight changes on the current direction of movement in the weight space. A good choice of both η and α are required for the training success and the speed of the neural network learning.

The simple perceptron is usable only for linearly separable or linearly independent problems. However, backpropagation learning with sufficient hidden layers can approximate any nonlinear function to arbitrary accuracy. This makes backpropagation learning neural network a good candidate for signal processing and modeling.

Backpropagation (BP) may stuck at a local minimum mainly because of the random initialized weights. For some initial weight settings, BP may not be able to reach a global minimum of weight space, while for other initializations the same network is able to reach an optimal minimum. A long recognized bane of analysis of the error surface and the performance of training algorithms is the presence of multiple stationary points, including multiple minima. Empirical experience with training algorithms show that different initialization of weights yield different resulting networks. Hence, multiple minima not only exist, but there may be huge numbers of them. In practice, there are four types of optimization algorithms that are used to optimize the weights (Abraham, 2005). The first three methods, gradient descent, conjugate gradients, and quasi-Newton use minimization of a quadratic error function to perform optimization. The fourth method of Levenberg and Marquardt uses minimization of an error function that is based on squared error criterion. A common feature for these training algorithms is given as the requirement of efficient calculation of gradients.

3.2 Principal Component Analysis

Principal component analysis (PCA) has been performed in order to reduce the dimension of feature vectors and to provide more compact representation of the speech and music samples. PCA is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in case of high dimension, PCA is a powerful tool for analysing data. (Lindsay & Smith, 2002)

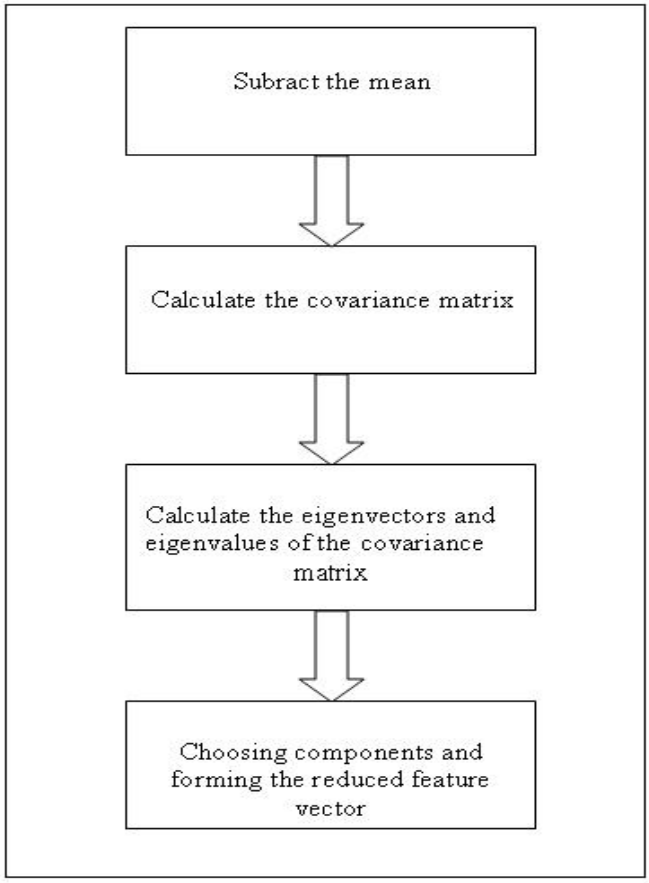


Figure 3.3 PCA process

PCA with variance maximization contains four essential steps as given in Figure 3.3. For PCA to work properly, you have to subtract the mean from each of the data sets. This produces a data set whose mean is zero. After calculation of covariance matrix, eigenvectors and eigenvalues of this matrix are obtained. Eigenvectors show

the direction of the axes with maximum variance and eigenvalues represent the significance of the corresponding axis.

Variance maximization method uses a linear combination (Hyvarinen, Karhunen, & Oja, 2001)

$$y_1 = \sum_{k=1}^{n} w_{k1} x_k = w_1^T x \qquad (3.4)$$

of the elements $x_1, ..., x_n$ of the vector x to perform PCA. w_{k1} are scalar coefficients or weights and they are elements of an n-dimensional vector w_1, and w_1^T denotes the transpose of w_1.

In the equation, the factor y_1 is named as first principal component of x where variance of y_1 is maximally large. To perform PCA process, a weight vector w_1 maximizing the PCA criterion is searched.

$$J_1^{PCA}(w_1) = E\{y_1^2\} = E\{(w_1^T x)^2\} = w_1^T E\{xx^T\} w_1 = w_1 C_X w_1 \qquad (3.5)$$

so that $\|w_1\| = 1$.

The matrix C_X is the covariance matrix of x with size of nxn and given for the zero-mean vector x by the correlation matrix

$$C_X = E\{xx^T\} \qquad (3.6)$$

Solution to the PCA problem is given in terms of the unit-length eigenvectors $e_1, ..., e_n$ of the the matrix C_X. Eigenvectors are ordered in a way such that the corresponding eigenvalues $d_1, ..., d_n$ satisfy $d_1 \geq d_2 \geq ... \geq d_n$.

The solution maximizing *(3.5)* is given by

$$w_1 = e_1 \qquad (3.7)$$

Thus the first principal component of x is $y_1 = e_1^T x$

It is possible to generalize the criterion J_1^{PCA} in eq. *(3.5)* to m principal components with m is any number between 1 and n. The m-th ($1 \le m \le n$) principal component is denoted as $y_m = w_m^T x$ where w_m is the corresponding unit norm weight vector. The variance of y_m is now maximized under the constraint that y_m is uncorrelated with all the previously found principal components:

$$E\{y_m y_k\} = 0, k < m \qquad (3.8)$$

This condition yields:

$$E\{y_m y_k\} = E\{(w_m^T x)(w_k^T x)\} = w_m^T C_X w_k = 0 \qquad (3.9)$$

We already know that $w_1 = e_1$ and for the second principal component, we have the condition that $w_2^T C w_1 = d_1 w_2^T e_1 = 0$. It must be searched that maximal variance $E\{y_2^2\} = E\{(w_2^T x)^2\}$ in the subspace orthogonal to the first eigenvector of C_X .
The solution is given as

$$w_2 = e_2 \qquad (3.10)$$

In general representation, w_k is given as

$$w_k = e_k \qquad (3.11)$$

In this way, the k^{th} principal component is given as $y_k = e_k^T x$.

As an example, the pca process for a two-dimensional vector is given in Figure 3.4. In the figure, the first principal component z_1 is the combination of variables that explains the greatest amount of variation. The second principal component z_2 defines the next largest amount of variation and is independent to the first principal component.

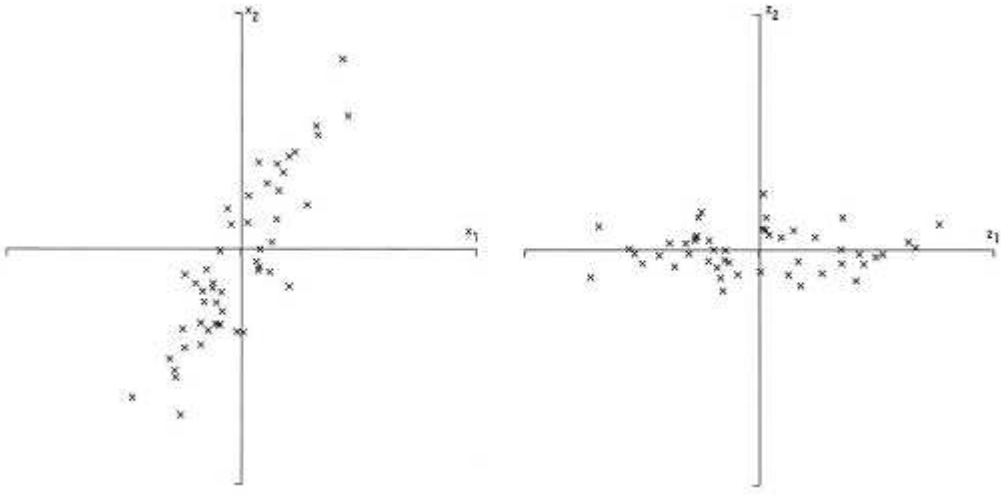


Figure 3.4 The principal component analysis representation for a two dimensional feature vector. (Jolliffe, 2002).

CHAPTER FOUR
RESULTS

In this chapter of the thesis, the results of SMD using time/frequency domain based and wavelet based features will be given.

4.1 Dataset and Preprocessing

The two different data sets have been utilized in the thesis and the features have been extracted separately for these two different datasets. In the first dataset, TIMIT database has been used for speech and several CD recordings with various musical genres have been used for music database. To obtain second dataset, radio broadcasts were recorded containing music and speech. The sampling frequency was set as 44100 Hz in every stage of thesis. However, since the data taken from TIMIT database is sampled with 16000 Hz, they have been interpolated in the pre-processing stage in order to set sampling frequency to 44100 Hz. The segmentation has been performed for a frame of 4196 samples with 512 samples overlapping which corresponds to a frame length of 95 ms since use of shorter window lengths may limit the discriminative characteristics of window.

Both datasets used in the thesis contain samples with length of 0.5 sec. While the first dataset includes 4290 music and 4620 speech samples, the second dataset contains 2190 music and 2624 speech samples entirely derived from radio broadcasts in contrary to first dataset. In rest of the context, first and second data sets will be named as Dataset1 and Dataset2, respectively. For the performance evaluation, the data sets have been divided into two groups as training and test sets. A detailed representation for dataset1 and dataset2 is given in Table 4.1.

Table 4.1 Content of datasets used in the thesis.

	Overall Database		**Train Set**		**Test Set**	
	Speech	**Music**	**Speech**	**Music**	**Speech**	**Music**
Dataset1	4620	4290	3080	2860	1540	1430
Dataset2	2624	2190	1749	1460	875	730

Before classification stage, the features that are highly correlated with the other features have been eliminated using principal component analysis (PCA) to reduce length of feature vectors. The principal components that contribute less than 0.05% to the total variation in the data set have been eliminated. Table 4.2 shows the length of the feature vectors before and after PCA. According to figure, it can be said that there is a reduction rate about 50% in terms of feature vector lengths after PCA process.

Table 4.2 Lengths of the feature vectors before and after PCA

Dimension	**T/F based feature vector**	**DWT based feature vector**					**DWT based energy feature vector**		**CWT based features**			
									CFS		**Q_Shift**	
		Haar	**Db2**	**Db8**	**Db15**	**Db20**	**Db8**	**Coif1**	**5-Band**	**7-Band**	**5-Band**	**7-Band**
Original	21	38	38	38	38	38	10	10	25	35	25	35
PCA	20	19	19	22	21	21	5	3	11	15	11	14

The used methods show different complexity behaviors in feature space. As an example, the PCA analyses for methods with highest and lowest accuracy are given in the Figure 4.1 and Figure 4.2.

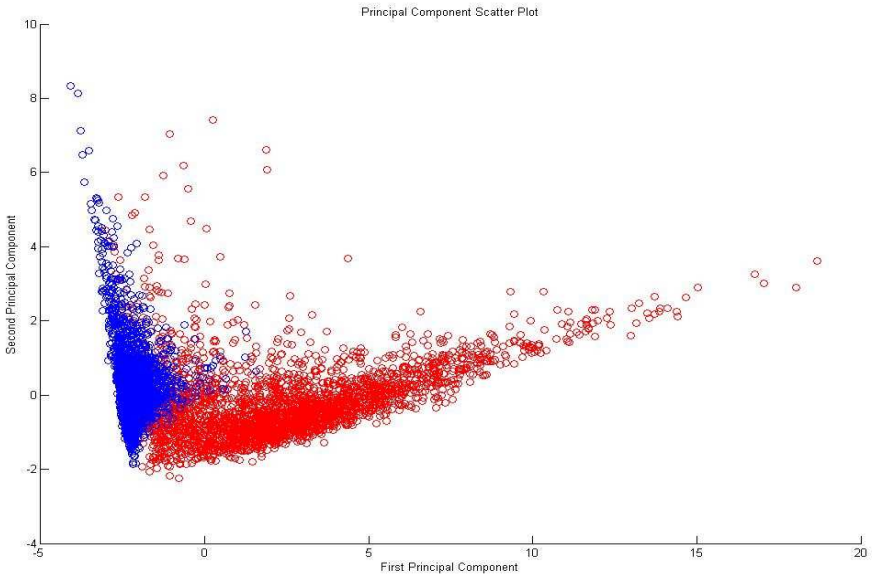


Figure 4.1 Principal component analysis of DWT (Daubechies8) based feature vector extracted for speech and music samples.

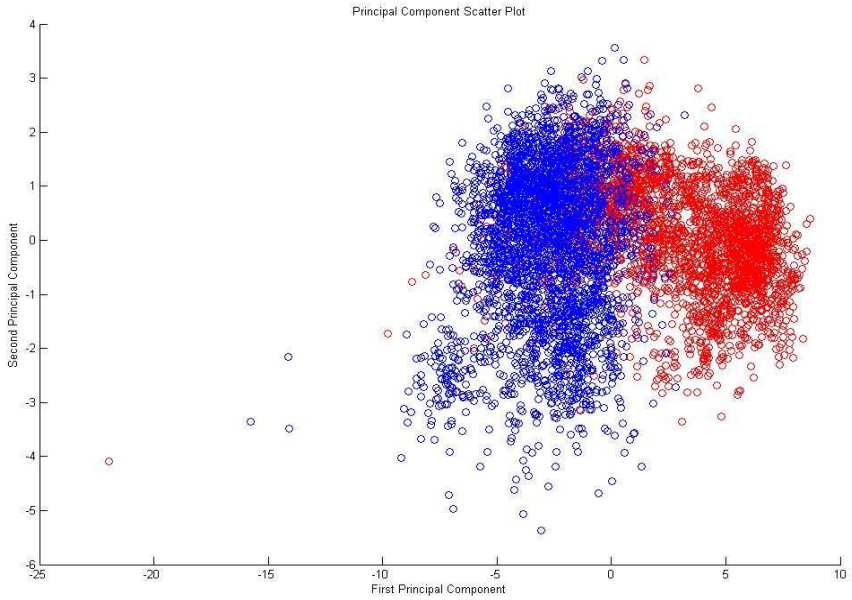


Figure 4.2 Principal component analysis of DWT based energy feature vector extracted for speech and music samples.

As it can be seen from the figures, Db8 based DWT features have higher discrimination capability than DWT based energy features. In Figure 4.1, the first and second principal components provide a good projection and representation, however, the samples are interwined in the feature space for DWT based energy features as given in Figure 4.2.

In classification stage of thesis, the feedforward artificial neural networks with the scaled conjugate gradient (SCG) backpropagation algorithm in MATLAB's Neural Networks Toolbox which belongs to class of the conjugate gradient algorithms have been used. SCG algorithm uses step size scaling instead of line-search per learning iteration and this makes it faster than other second order algorithms (Charalambous, 1992). This algorithm performs well for networks with a large number of weights where it is as fast as the Levenberg-Marquardt and resilient backpropagation algorithms, its performance does not degrade quickly. Also, the conjugate gradient algorithms have relatively modest memory requirements. The number of hidden neurons has been preferred as 40 and the target mean square error has been defined as 0.001, heuristically.

All codes and programs in the thesis were written in MATLAB. The codes for time/frequency based features, DWT based statistical and energy features were written by the author of thesis. For DWT based analysis, Wavelet Toolbox of MATLAB has been used. For CWT based analysis, the codes are taken from the study of I. Selesnick (Selesnick, I.W., 2001) for common factor solution based filter design and the programs written by two students under supervision of I.Selesnick (Cai, S. & Li, K, 2002) have been used for Q-Shift filter based analysis.

In the following section, the classification results will be given for four types of feature vectors. The performance has been given as the accuracy of the classification which can be formulated as

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (4.1)$$

where TP, TN, FP and FN represent number of speech samples labeled as speech, number of music samples labeled as music, number of music samples labeled as speech and number of speech samples labeled as music, respectively.

4.2 Classification Performance

4.2.1 Performance for Time / Frequency Based Features

The feature vector of this method includes mean and variances of the parameters obtained with time/frequency domain features and Mel Frequency Cepstral Coefficients. For this method, the length of the feature vector is 21. The parameters such as mean of spectral centroid, variance of spectral centroid, mean of spectral roll off, variance of spectral roll off, mean of spectral flux, variance of spectral flux, mean of zero crossings, variance of zero crossings, low energy ratio and Mel frequency cepstral coefficients are represented in the Table 1 as Cm, Cvar, RLfm, RLfvar, FLUXm, FLUXvar, zerocrosm, zerocrosv, low_energy and melcepst, respectively.

Table 4.3 The content of feature vector for time/frequency based features

Features	**Cm**	**Cvar**	**RLfm**	**RLfvar**	**FLUXm**	**FLUXvar**	**zerocrosm**	**zerocrosv**	**low_energ**	**melcepst**
Length of feature vector (21)	1	1	1	1	1	1	1	1	1	12

Table 4.4 Classification results for for time/frequency based features.

Performance (%)	*Dataset1*	*Dataset2*
TFPa	99.72	94.27

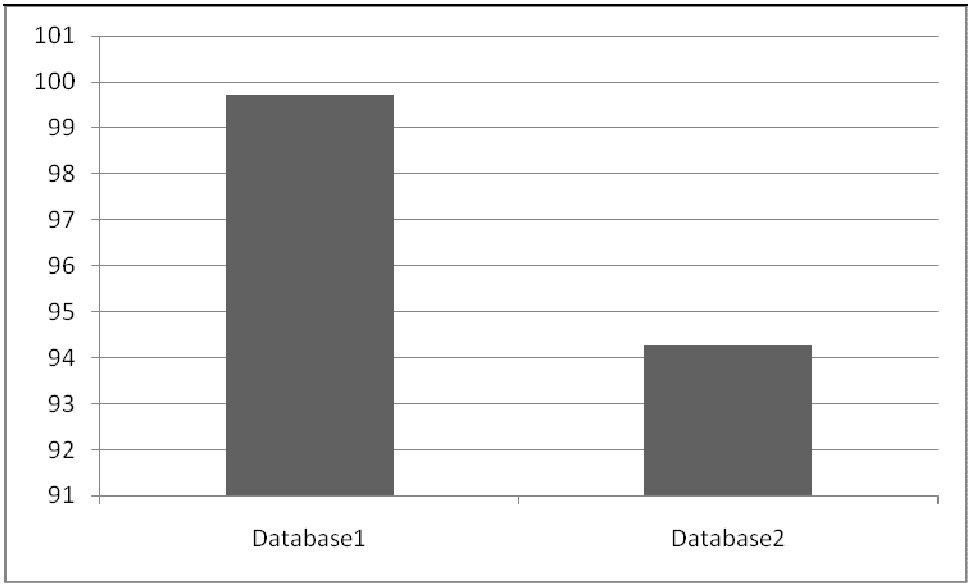


Figure 4.3 Performance of time/frequency based features for dataset1 and dataset2

4.2.2 Performance for DWT Based Features

In the second method, feature extraction has been performed for several wavelets such as Haar (db1), db2, db8, db15 and db20. The filter length is 2N for a Daubechies wavelet which has N vanishing moments. 12-level decomposition has been considered in feature set which covers the analyzed frequency range in detail, therefore 1 approximation and 12 detail signals are obtained for each frame.

In Figure 4.4 and Figure 4.5, 12-level discrete wavelet decomposition using db8 wavelet for music and speech signals are presented. The speech and music signals show different characteristics particularly for different frequency bands as it can be seen from the figures.

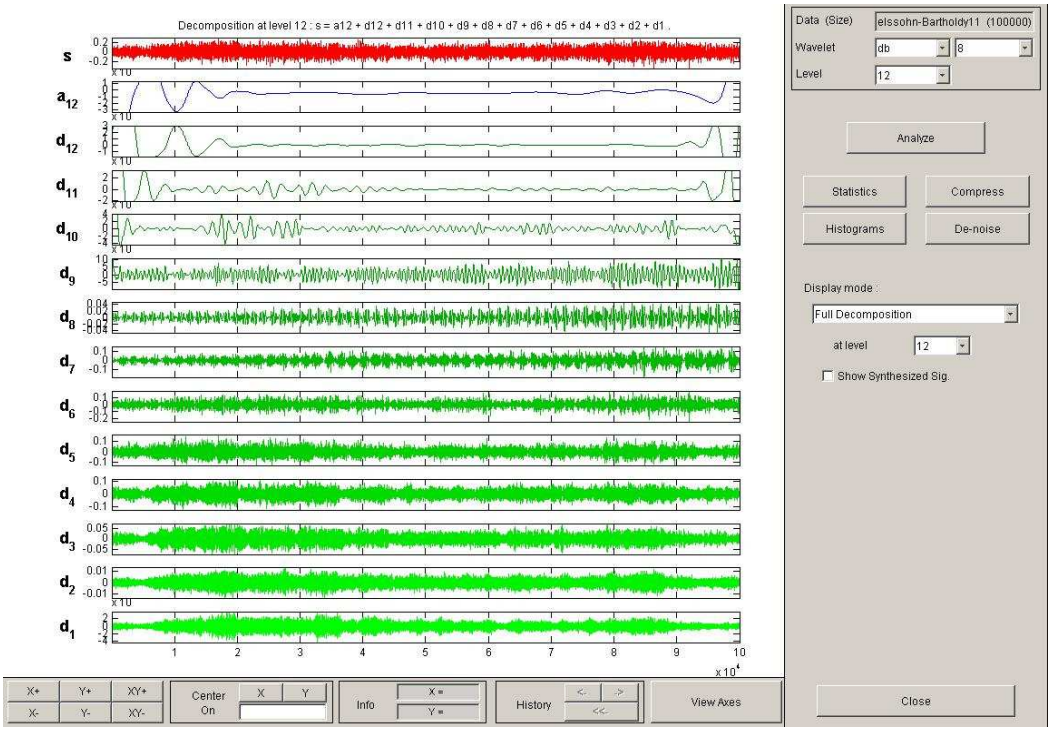


Figure 4.4 The 12-band decomposition with db8 wavelet for a music signal used in the thesis

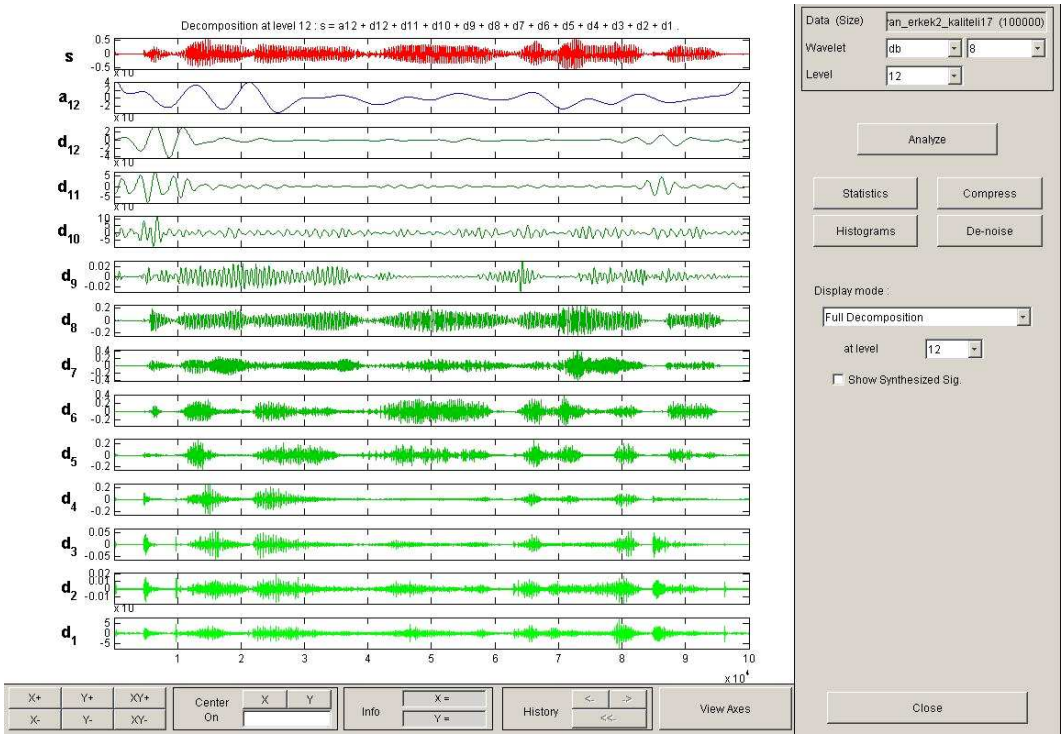


Figure 4.5 The 12-band decomposition with db8 wavelet for a speech signal used in the thesis

The length of feature vector which is constructed from the statistical measures including mean, standard deviation and ratios of the decomposed signals is 38 as it is shown in Table 4.5.

Table 4.5 The length of feature vector for DWT based features is 38.

Features	**Mean of detail and approx. coefs (12-band)**	**Std. of detail and approx. coefs (12-band)**	**Ratios of detail and approx. coefs. (12-band)**
Length of feature vector (38)	13	13	12

The classification results are given in Table 4.6.

Table 4.6 Classification results for DWT based features.

Performance (%)	**Dataset1**	**Dataset2**
Haar	99.9	96.51
db2	99.93	97.69
db8	**99.97**	**99.19**
db15	99.83	98.63
db20	99.9	98.69

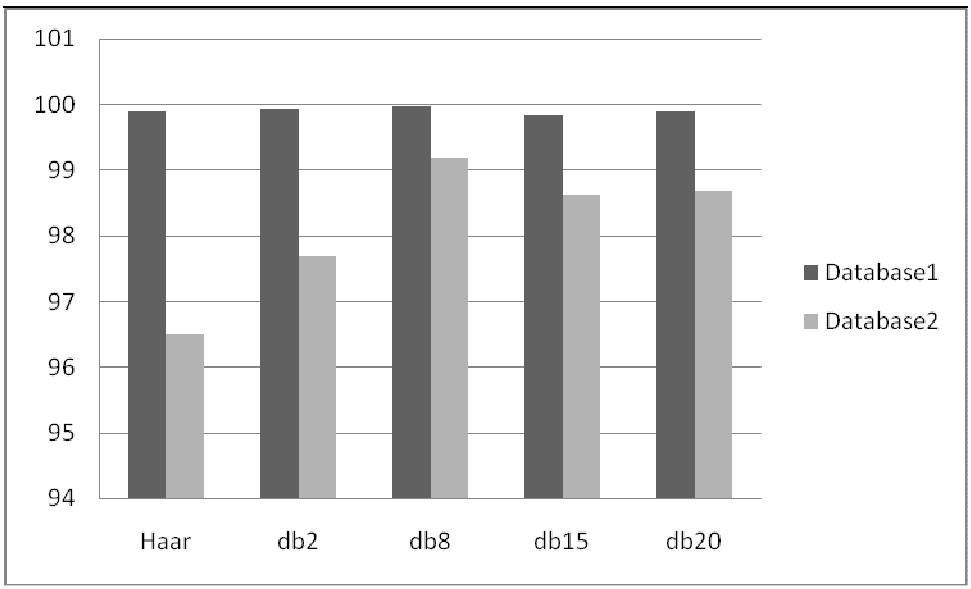


Figure 4.6 Accuracy results of DWT based features for dataset1 and dataset2 with different wavelets

Table 4.6 shows that the DWT has the ability of discrimination of speech and music signals with high accuracy. As in the first method mentioned in section 4.2.1, the performance of the Database 1 is also higher for this method since the signals are more separable. The accuracy changes slightly depending on the used wavelet. It can be said that db8 is the most successful wavelet in terms of classification of speech and music with the accuracy rates of 99.97% for Dataset1 and 99.19% for Dataset2.

When the DWT based feature extraction is performed for shorter samples such as with length of lower than 0.5 sec., it has been observed that the classification performance tends to decrease, since the wavelets cannot represent the segments that have such a short length.

In order to see the contribution of ratio parameters to the discrimination performance, a classification has been also performed using the feature vectors with length of 26 where the feature set does not contain ratios of frequency sub-band coefficients. The PCA process is also applied to features which do not contain ratio parameters and results are given in Table 4.7. At the end of this experiment, it has been observed that ratio parameters provide a contribution to overall performance

about 1-1.5% for DWT based parameters.

Table 4.7 The length of feature vector of DWT based parameters with and without ratio parameters after PCA process.

Length of feature vector	**DWT based feature vector**									
	Haar		**Db2**		**Db8**		**Db15**		**Db20**	
	With ratio	**Without ratio**	**With ratio**	**Without ratio**	**With ratio**	**Without ratio**	**With ratio**	**Without ratio**	**With ratio**	**Without ratio**
Original	38	26	38	26	38	26	38	26	38	26
After PCA	19	13	19	13	22	16	21	16	21	16

Table 4.8 Comparison of the classification results between DWT based feature sets with ratio parameters and without ratio parameters.

Wavelets	**Performance with ratio parameters**	**Performance without ratio parameters**
Haar	96.51	94.58
db2	97.69	95.95
db8	**99.19**	97.69
db15	98.63	97.88
db20	98.69	**98.32**

In Table 4.8, it can be seen that db8 shows the best result if the feature vector contains ratio parameters. In absence of ratio parameters, db20 shows higher performance among other wavelet families. The number of vanishing moments is related with the smoothness of the wavelet although there is not a proved correspondence. Since Haar and db2 mother wavelets have a few vanishing moments, they have sharp transitions. Therefore, they cannot represent smooth signals such as music samples. For db8 and wavelets with more vanishing moments,

an acceptable accuracy has been obtained. Db8 has been chosen since using a higher order will result in a more complex computation due to the increasing number of the filter coefficients.

4.2.3 Performance for DWT based energy features

In this method, only detail coefficients have been used at the feature extraction stage. The decomposition has been performed for 5 levels of subbands and two energy parameters such as instantaneous and teager energy have been obtained for each band. In this way, length of the feature vector for each sample is 10 according to this method as shown in Table 4.9.

Table 4.9 The length of feature vector for DWT based energy features is 10.

Features	**Instantaneous Energy (5-band)**	**Teager Energy (5-band)**
Length of feature vector (10)	5	5

According to classification results given in Table 4.10, this method has not provided high performances comparing to other methods.

Table 4.10 Classification results for DWT based energy features.

Performance (%)	**Dataset1**	**Dataset2**
Db8	89.02	91.21
Coif1	82.93	77.45

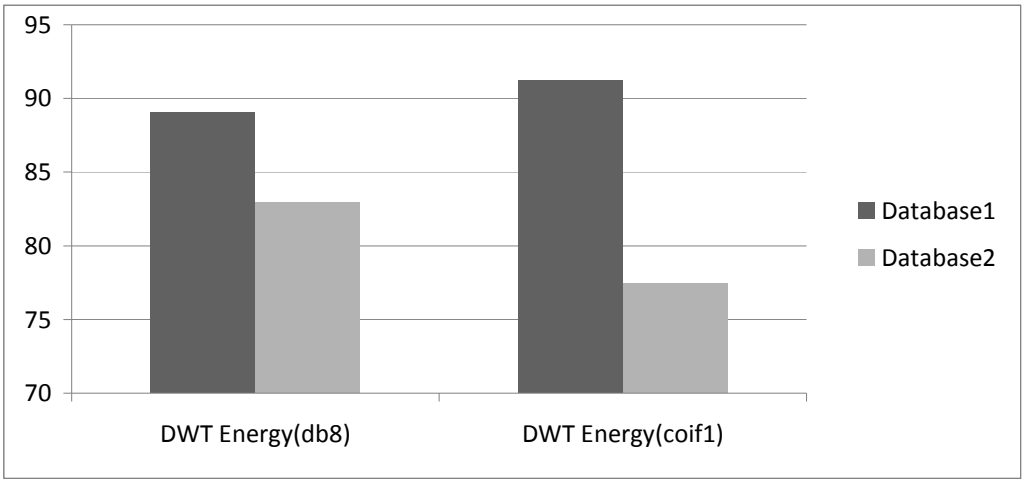


Figure 4.7 Accuracy results for DWT based energy features dataset1 and dataset2 with different wavelets.

4.2.4 Performance for CWT Based Features

The complex wavelet transform (*C*WT) has been accomplished by using two filter design methods introduced in Chapter 3. The decomposition has been made for 5 and 7 levels in order to avoid increasing length of the feature vector. The feature vectors have been constructed from mean, variance and median of the magnitudes of complex wavelet coefficients at each band instead of using all coefficients to avoid increasing in length of feature vector. In this way, the length of feature vectors are defined for each sample is given as 25 or 35 for 5-level and 7-level decomposition, respectively. The content of feature vector is given in Table 4.11.

Table 4.11 The length of feature vector for *C*WT based features is 25 for 5-level decomposition and 35 for 7-level decomposition.

Features	**3.moment for each band**	**4.moment for each band**	**Mean of each band**	**Standard deviation of each band**	**Median of each band**
Length (25 or 35)	5 or 7	5 or 7	5 or 7	5 or 7	5 or 7

Table 4.12 Classification results for *C*WT based features with different filter design methods and different numbers of subbands.

Performance (%)	Dataset1	Dataset2
CFS (5 Levels)	99.12	**98.13**
Q_Shift (5 Levels)	**99.93**	97.95
CFS (7 Levels)	99.87	97.82
Q_Shift (7 Levels)	**99.93**	97.57

In Table 4.12, it can be seen that Dataset1 can be easily discriminated with the proposed features with slightly more accurate results with Q-shift parameters. According to these results, an increment in the number of frequency bands does not contribute to the classification. CFS solution based CWT coefficients have higher accuracy rate than Q-Shift solution for Dataset2. It is possible to say that any remarkable enhancement in classification results has not been observed since complexity in feature space increase when high-level decomposition is performed.

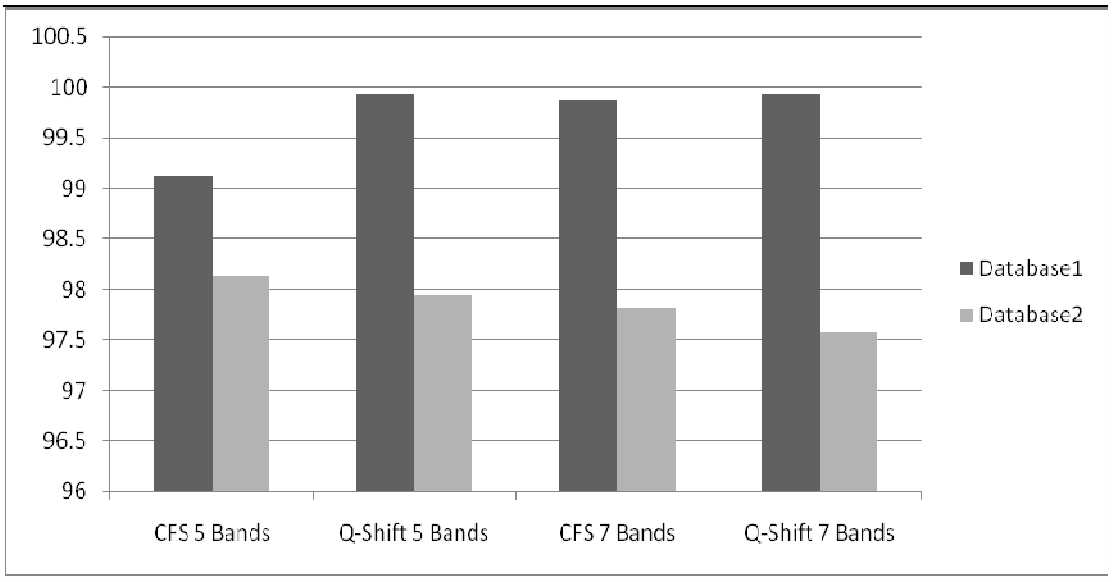


Figure 4.8 Accuracy results of CWT based features for dataset1 and dataset2

Classification performance of samples shorter than 0.5 sec. stated in previous section is also valid for CWT based features. It has been observed for CWT based parameterization that it is not very effective in terms of classification of signals with

such a short length as in DWT.

It was mentioned that ratio parameters of adjacent subbands provided a contribution about 1-1.5% to classification performance for DWT based parameters. As in other methods, the PCA process has been applied to *C*WT based features which do not contain ratio parameters before classification stage and dimensions of new feature vectors are given in Table 4.13. It has been also investigated for *C*WT based parameters if there is an increase in terms of classification performance. The results are given in Table 4.14

Table 4.13 The length of feature vector of *C*WT based parameters with and without ratio parameters after PCA process

Length of feature vector	**CWT based feature vector**							
	Q_Shift				**CFS**			
	5-Level		**7-Level**		**5-Level**		**7-Level**	
	With ratio	**Without ratio**	**With ratio**	**Without ratio**	**With ratio**	**Without ratio**	**With ratio**	**Without ratio**
Original	29	25	41	35	29	25	41	35
PCA	13	11	18	14	13	11	18	15

Table 4.14 Comparison of classification performances between *C*WT based parameters with ratio parameters and without ratio parameters.

Filter Design Method & Number of subbands	**Performance with ratio parameters (%)**	**Performance without ratio parameters (%)**
CFS & 5-Level	97.88	98.13
CFS & 7-Level	**98.50**	97.82
Q_SHIFT & 5-Level	97.76	97.95
Q_SHIFT & 7-Level	**98.50**	97.95

According to results in Table 4.14, addition of ratio parameters to *C*WT based feature vector does not make a remarkable contribution to classification performance. However, it should be noted that in this method, 5-band and 7-band decompositions have been made differently from DWT based feature extraction method. Hence, it can be thought as the effect of ratio parameters in CWT is less than in DWT because of the difference between decomposition levels of two methods.

4.2.5 General Performance

General performance is given as in the Table 4.15.

Table 4.15 General classification results

Performance (%)	**Dataset1**	**Dataset2**
Haar	99.9	96.51
db2	99.93	97.69
db8	**99.97**	**99.19**
db15	99.83	98.63
db20	99.9	98.69
TFPa	99.72	94.27
CFS (5 Levels)	99.12	98.13
Q_Shift (5 Levels)	99.93	97.95
CFS (7 Levels)	99.87	97.82
Q_Shift (7 Levels)	99.93	97.57
DWT_Energy (db8)	89.02	91.21
DWT_Energy (coif1)	82.93	77.45

When Table 4.15 is taken into consideration, it can be seen that wavelet based parameters have higher classification results than traditional time / frequency based methods. In general, all methods are successful in classification of samples in Dataset1, which indicates that the TIMIT speech data and CD recordings are separable. However, it is not possible to say same thing for Dataset2 since the samples in Database 2 reflects a more realistic case where samples are recorded from

radio broadcast. The best performance has been obtained with db8 wavelet. The complex wavelet based features performs better than time / frequency based methods and wavelets with fewer vanishing moments. However, they are not as successful as the db8. The similarity of the mother wavelet with the analyzed waveforms is an important criterion for the wavelet analysis which may be the cause of this performance difference. Therefore, the accuracy for different databases may differ drastically.

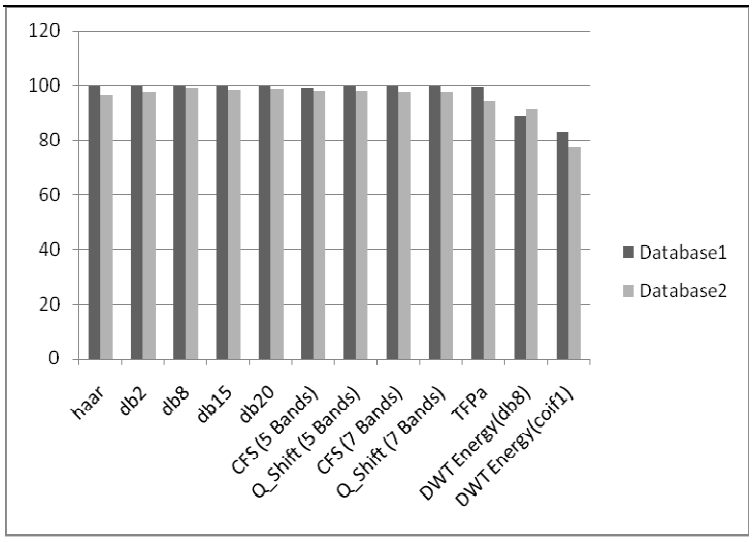


Figure 4.9 General Accuracy

When these feature extraction methods are considered in terms of their calculation times, DWT based energy features emerge as the fastest algorithm in terms of feature extraction since it contains only ten parameters in feature vector. On the other hand, DWT based energy features have the lowest classification performance among the considered methods according to results. In Table 4.16, the computation times for feature extraction stage for all methods used in the thesis are given.

Table 4.16. Average computation times for feature extraction methods used in the thesis

	Speech (msn.)	**Music (msn.)**
Time/freq. based features	0.2768	0.2745
Haar	0.0357	0.0382
Daubechies2	0.0401	0.04
Daubechies8	0.0485	0.0462
Daubechies15	0.1035	0.1034
Daubechies20	0.155	0.1547
Daubechies8 based energy features	0.0216	0.0217
Coiflet1 based energy features	0.0176	0.0176
Q-shift based CWT features	0.0298	0.0296
CFS based CWT features	0.0301	0.03

According to average computation times in Table 4.16, a sorting among the feature extraction methods can be made as:

$$t_{TF} > t_{DWT} > t_{CWT} > t_{DWTE}$$

where t_{TF} , t_{DWT} , t_{CWT} and t_{DWTE} show the computation time for the methods based on time/frequency, DWT, CWT and DWT based energy features.

The calculation time for DWT based statistical feature extraction shows differences according to used wavelet in the analysis. Wavelet families including high number of vanishing moments such as db15 and db20 spend more time for computation comparing to other wavelet families since they have longer filter lengths. It is encountered that the db8 families as the optimum wavelet for DWT based analysis since it shows highest performance in classification of speech and music and it has acceptable calcularion time.

*C*WT based method is faster than DWT based analysis and it shows performance results close to DWT. In this perspective, *C*WT based features can be used for online

implementation as well.

Time/frequency based analysis is the slowest method since it performs many computations in time and frequency domain and it has a long feature vector with length of 38. It should be noted that the silence parts of samples could be determined more quickly than the feature extraction methods during online implementation since only a threshold value is considered to give decision if the segment is silence or not.

4.3 Graphical User Interface (GUI) Design for Speech / Music Discrimination

A graphical user interface has been designed as well in order to perform speech music discrimination visually. An online labelling module has been also embedded to the interface and observation of performance for real time classification has become possible with this tool.

4.3.1 Main Module

Main module can be used to see classification results obtained by methods used in the thesis.In Figure 4.8, the GUI designed for main module is shown. Using “Load File” button, the file to be analyzed is selected and the “play” button plots and plays the signal at the same time. Since time/ frequency based features are also used at classification stage, it is important to see the general structure of spectrogram. For this aim, there is a button named as “Spectrogram of signal” in the module to plot time/frequency properties. In the DWT based features part of module, 12-level decomposition is performed using selected wavelet from pop-up menu. It is also possible to see shape of wavelet and wavelet coefficients using “Show Wavelet” and “Plot Wavelet Coefficients” buttons, respectively. For *C*WT based features there is also a pop-up menu which you can select the filter design method for analysis. It can be seen the existing complex wavelet with its real and imaginer parts using “Plot Filter Corfficients” button in CWT based feature extraction part of the module. For time/frequency based feature extraction, there is also a button and the values of

parameters such as spectral centroid, spectral rolloff, spectral flux, number of zero crossings and low energy ratio of loaded file exist in the blanks when this button is pushed. It is possible to get the classification results of four feature extraction methods simultaneously using “Classification Results” button. If “Online Labelling Module” button is pushed, the online labelling module will appear in a new window. This module will be introduced in the next section.

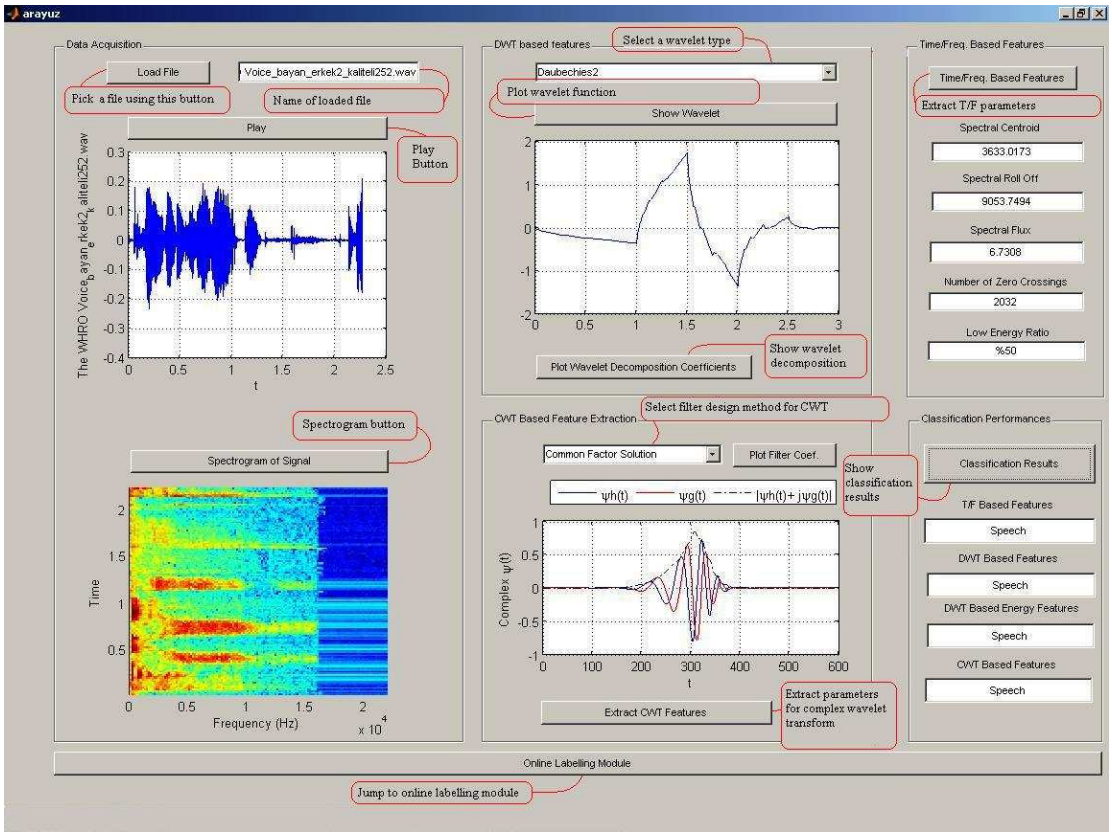


Figure 4.8. Graphical user interface for main module

4.3.2 Online Labelling Module

This module has been designed to observe speech / music classification performance for online implementations. In the given module in Figure 4.9, a pre-recorded sample is fetching using “Open” button and the online labelling process is started using “Start” button. “Pause” button makes it possible to stop the process temporarily and using “Continue” button the labeling process can be continued from

where it is stopped. The “Stop” button interrupts the program and ends the label assignment process.

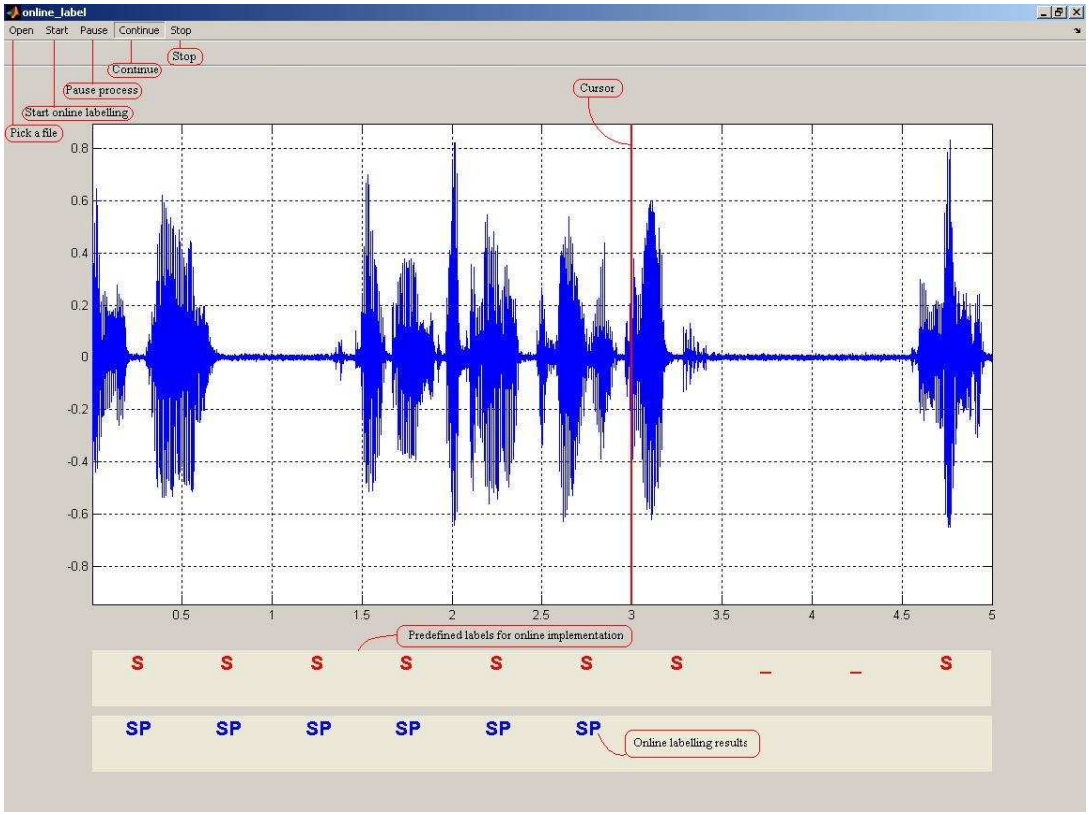


Figure 4.9 Graphical user interface for online labelling module

In the module, the red letters under the signal graph shows the pre-assigned labels for data and “S”, “M” and “_” are used to indicate speech, music and silence parts of data, respectively. Online classification results are shown with blue color under pre-assigned labels as it can be seen from Figure 4.9. These labels are assigned for segments which have the length of 0.5 sec. In online label assigning, features are extracted using 12- level DWT with db8 wavelet for each sample since it has given the most accurate results in experiments and a previously trained artificial neural network is used to determine the labels.

CHAPTER FIVE
CONCLUSION

5.1 Summary

The discrimination of music and speech have been an important task in multimedia signal processing with the increasing role of the multimedia sources in our life. The music/speech discrimination systems can be used in several applications as a preproecssing stage such as in the development of the efficient coding algorithms for audio decoders, in automatic speech recognition when the recordings include music such as radio broadcasts, in content based multimedia retrieval and in automatic channel selector design problem for radios. In addition, there are other emerging applications with a growing interest for music/speech discriminators.

In this thesis, classification of speech and music signals has been investigated in many aspects. The feature extraction has been performed with four different methods and artificial neural networks have been used as a classification tool. Two different databases have been used and feature extraction has been made individually for these databases. The first method has a parameter vector which contains time/frequency based features and mel-cepstrum coefficients with length of 21. Second and third method use DWT based features. In second method, using several types of mother wavelets, 12-level decomposition has been performed to cover the analysis frequency range in detail. The length of feature vector constructed from the statistical measures of the coefficients and ratios between the adjacent subbands is 38. The third method contains DWT based energy parameters named as Teager and Instantaneous energy differently from second method. The length of feature vector for third method is 10. The last method is based on Complex Wavelet Transform (*C*WT) and two different filter design strategies including Common Factor Solution and Q_Shift solution have been used at feature extraction stage. *C*WT has been performed for 5-level and 7-level to avoid further increase in the length of the feature vector which results in feature vectors with length of 25 and 35 for 5 and 7 bands, respectively. It has been observed that time / frequency based features are not very effective in discrimination

of speech / music samples when they are used alone. However, if they are used together, the accuracy tends to increase conspicuously.

The methods except the energy based ones, shown higher performance for Database 1 than the results for Database 2. Because, the second database consists of the recordings from radio broadcasts which reflects a more realistic case.

The selection of the analysis window length which specifies the content of the nonstationary signal and the speed of implementation is an important choice for SMD. The selection of a short window order of miliseconds as in literature will not give the necessary information on time varying frequency content, since the signal can be assumed as stationary in this interval. On the other hand, the usage of long windows order of seconds which is reported as successful limits the online application of the algorithms. In this study, it has been observed that the 0.5 sec analysis window length is effective in terms of performance.

*C*WT and DWT based features have shown a high success comparing to time / frequency based features according to classification results. Different accuracy rates have been encountered for different mother wavelets belonging to Daubechies wavelet family. Daubechies8 demonstrated the highest classification performance among the others. The *C*WT based classification has shown results as 99.93% for Database1 and 98.13% for Database2. When all features are concerned, we see that Daubechies8 based parameters have superior discrimination features in terms of classification of speech and music.

In the thesis, the contribution of the ratio parameters to the discrimination performance have also been examined for DWT and *C*WT based features. It has been observed that ratio parameters provide a contribution about 1-1.5% to the overall performance for DWT based parameters. However, the results were inconclusive for *C*WT.

Classification performance of DWT based feature vector in method 2 varies depending on mother wavelet used in feature extraction stage. When the number of vanishing moments is increased, the wavelet becomes smoother. These smooth wavelets produce large coefficients for slowly changing signals like music, while it produces relatively small coefficients for speech signals. This can be used as a discriminative property for SMD. The Haar and db2 wavelets have a few vanishing moment, this may cause to prevent the good representation of signal in frequency domain. In contrary, db15 and db20 have much more filter coefficients and vanishing moments, but this increases the complexity in the feature space and also requires longer computations. In this way, db8 has emerged as the most ideal wavelet type of wavelets used in the thesis.

In classification stage, artificial neural networks have been used as classification tool. The number of hidden neurons has been preferred as 40 and the target mean square error has been defined as 0.001, heuristically. Conjugate gradient algorithms have been selected as learning algorithm since they have advantages according to other methods. Also, principal component analysis has been performed before classification stage to represent signals more efficiently and to decrease the dimension of feature vector.

5.2 Advantages

In this thesis, the speech and music samples with length of 0.5 sec. have been used at feature extraction and classification stages. Although longer segments are used in the literature generally, it has been shown in the thesis that 0.5 sec. length is enough to get high performance in classification of speech and music. The proposed algorithm used in the thesis is computationally efficient (average running time for proposed method is <50 msn) and this allows the use of method for online implementation. As mentioned before, a fast running speech / music discrimination system with high accuracy can be designed by using suggested method as a preprocessing stage for several applications.

5.3 Disadvantages

The observed SMD performance of *C*WT based features were less than the DWT based ones. A feature set which reflects the most powerful properties of *C*WT must be constructed. The filter structure used in *C*WT based parameterization has the possibility of presence of unsuitable charactestics in terms of speech/music discrimination and as a result, the accuracy is observed as lower than performance of DWT based features. In this manner, adaptive filter design is required to get more succesful results.

5.4 Future Studies

Since the SMD is an hot topic for multimedia applications, the studies can be extended in several directions. One of them might be the research on adaptive filter design methods to reveal more advantages of CWT on DWT in speech / music discrimination. Therefore, the parameters fort he SMD tasks can be determined automatically according to the problem at hand.

The dataset can be expanded to include mixed speech-music samples. In this way, a multiclass classification can be performed instead of binary classification for future studies.

The performance of CWT based features can further examined to construct a more discriminated feature space

An hardware implementation can be done using digital signal processors to have a faster SMD system.

REFERENCES

Abraham, A. (2005). Artificial Neural Networks. Sydenham P. H., & Thorn R., (Eds.), *Handbook of Measuring System Design*. (901-908). London: John Wiley & Sons, Ltd.

Burred, J.J., Lerch, A. (2003). A Hierarchical Approach To Automatic Musical Genre Classification. *Proceedings of the 6th Int. Conference on Digital Audio Effects,* 1-4. Retrieved April 27, 2010, from CiteSeerX database.

Cai, S., Li, K. (2002). *Matlab Implementation of Wavelet Transforms*. Retrieved December 3, 2009, from http://taco.poly.edu/WaveletSoftware/

Carey, M.J., Parris, E.S., Lloyd-Thomas, H. (1999). A comparison of features for speech, music discrimination. *IEEE international conference on acoustics, speech and signal processing (ICASSP), 1,* 1432–1435. Retrieved May 3, 2010, from IEEE Xplore database.

Charalambous, C., (1992). Conjugate gradient algorithm for efficient training of artificial neural networks. *IEEE Proceedings-G on Circuit Devices and System, 139(3),* 301-310. Retrieved May 4, 2010, from IEEE Xplore database.

Choi, M. Y., Song, H. J., Kim, H. S. (2007). Speech/Music Discrimination for Robust Speech Recognition in Robots. *IEEE International Conference on Robot, & Human Interactive Communication,* 118-121. Retrieved June 30, 2009, from IEEE Xplore database.

Chun-Lin, L. (2010). *A Tutorial of the Wavelet Transform.* Retrieved May 28, 2010, from disp.ee.ntu.edu.tw/tutorial/WaveletTutorial.pdf

Didiot, E., Illina, I., Fohr, D., Mella, O. (2010). A wavelet-based parameterization for speech/music discrimination. *Computer Speech and Language, 24(2),*341–357. Retrieved May 4, 2010, from ScienceDirect database.

El-Maleh, K., Klein, M., Petrucci, G., Kabal, P. (2000). Speech/music discrimination for multimedia applications. *IEEE international conference on acoustics, speech and signal processing (ICASSP), 6,* 2445–2448. Retrieved May 3, 2010, from IEEE Xplore database.

Exposito, J.E.M., Galan, S.G., Reyes, N.R., Candeas, P.V. (2007). Audio coding improvement using evolutionary speech/music discrimination. *IEEE international conference on fuzzy systems (FUZZ-IEEE),* 1–6. Retrieved May 3, 2010, from IEEE Xplore database.

Ezzaidi, H., Rouat, J. (2007). Comparison of the statistical and information theory measures: application to automatic musical genre classification. *IEEE Workshop on Machine Learning for Signal Processing*, 241–246. Retrieved May 3, 2010, from IEEE Xplore database.

Gong, C., Xiong-wei Z. (2006). The application of speech/music automatic discrimination based on gray correlation analysis. *IEEE international conference on cognitive informatics (ICCI), 1,* 68–72. Retrieved May 4, 2010, from IEEE Xplore database.

Graps, A. (1995). An Introduction to Wavelets. *IEEE Computational Science & Engineering, 2(2),* 50-61. Retrieved May 3, 2010, from http://www.amara.com/ftpstuff/IEEEwavelet.pdf

Harb, H., Chen, L. (2003). Robust speech music discrimination using spectrum's first order statistics and neural networks. *IEEE International Symposium on Signal Processing and its Applications, 2,* 125–128. Retrieved May 3, 2010, from IEEE Xplore database.

Haykin, S. (1999). *Neural Networks A Comprehensive Foundation* (Second Edition). New Jersey: Prentice Hall International

Heil, C.E., Walnut D.F., (1989). Continous and Discrete Wavelet Transforms. *Society for Industrial and Applied Mathematics Review, 31(4),* 628-666. Retrieved May 28, 2010, from CiteSeerx database.

Hyvarinen, A., Karhunen, J., Oja E. (2001). *Independent Component Analysis.* Canada: John Wiley & Sons, Inc.

Jha, G. K. (2003). *Artificial Neural Networks.* Retrieved May 3, 2010, from http://www.iasri.res.in/ebook/EB_SMAR/e-book_pdffiles/Manual IV/3-ANN.pdf

Jolliffe, I. T. (2002). *Principal Component Analysis,*(Second Edition). USA: Springer

Karneback S. (2001). Discrimination between speech and music based on a low frequency modulation feature. *European conference on Speech Communications and Technology,* 1891–1894. Retrieved May 2, 2010, from CiteSeerX database.

Khan M. K. S., Al-Khatib, W. G. (2006). Machine-learning based classification of speech and music. *Multimedia Systems, 12(1),* 55–67. Retrieved May 3, 2010, from SpringerLink database.

Kingsbury, N. (1998). The dual-tree complex wavelet transform: a new technique for shift invariance and directional filters. *Proceedings of the IEEE Digital Signal Processing Workshop*, 1-4. Retrieved from Retrieved February 19, 2010, from CiteSeerx database.

Lindsay, I., (2002). *A tutorial on Principals Component Analysis.*Retrieved ay 3, 2010, from http://www.cs.toronto.edu/~fleet/courses/D11/fall09/Handouts/pca smithTutorial.pdf.

Lu, L., Zhang, H., Jiang, H. (2002). Content analysis for audio classification and segmentation. *IEEE Transactions on Speech and Audio Processing*, *10(7),* 504–516. Retrieved May 3, 2010, from IEEE Xplore database.

Matsunaga, S., Mizuno, O., Ohtsuki, K., Hayashi, Y. (2004). Audio source segmentation using spectral correlation features for automatic indexing of broadcast news. *XII. European Signal Processing Conference,* 2104–2106. Retrieved May 3, 2010, from http://www.eurasip.org/Proceedings/Eusipco/Eusipco2004/defevent/papers/cr1404.pdf

Merry, R.J.E., (2005). *Wavelet Theory and Applications A literature study.* Retrieved May 28, 2010, from http://alexandria.tue.nl/repository/books/612762.pdf

Minami, K., Akutsu, A., Hamada, H., Tonomura, Y. (1998). Video handling with music and speech detection. *IEEE Multimedia, 5(3),* 17–25. Retrieved May 3, 2010, from IEEE Xplore database.

Ntalampiras, S., Fakotakis, N. (2008). Speech/Music Discrimination Based on Discrete Wavelet Transform. *Proceedings of the 5th Hellenic conference on Artificial Intelligence: Theories, Models and Applications*, *5138*, 205-211. Retrieved May 3, 2010, from SpringerLink database.

Panagiotakis, C., Tziritas, G. (2005). A speech/music discriminator based on RMS and zero–crossings. *IEEE Transactions on Multimedia*, *7(1),* 155–166. Retrieved May 3, 2010, from IEEE Xplore database.

Pikrakis, A., Giannakopoulos, T., Theodoridis, S. (2008). A Speech/Music Discriminator of Radio Recordings Based on Dynamic Programming and Bayesian Networks. *IEEE Transactions on Multimedia,10* (5), 846-857. Retrieved June 30, 2009, from IEEE Xplore database

Richard G., Ramona M., Essid S. (2007). Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. *IEEE international conference on acoustics, speech and signal processing (ICASSP), 2,* 461–464. Retrieved May 4, 2010, from IEEE Xplore database

Rong-Yu, Q. (1997). Mixed wideband speech and music coding using a speech/music discriminator. *TENCON- IEEE Region 10 Annual Conference on Speech and Image Technologies for Computing and Telecommunications, 2,* 605–608, Retrieved May 4, 2010, from IEEE Xplore database.

Roy, A. (2000). Artificial Neural Networks-A Science in Trouble. *ACM SIGKDD Explorations Newsletter, 1(2),* 33-38. Retrieved June 30, 2009, from CiteSeerX database.

Ruiz-Reyes, N., Vera-Candeas, P., Muñoz, J.E., García-Galán, S., Cañadas, F. J. (2009). New speech/music discrimination approach based on fundamental frequency estimation. *Multimedia Tools and Applications, 41(2),* 253–286. Retrieved May 2, 2010, from SpringerLink database.

Saad, E.M., El-Adawy, M.I., Abu-El-Wafa, M.E., Wahba, A.A. (2002). A multifeature speech/music discrimination system. *Nineteenth National Radio Science Conference, Alexandria,* 208–213. Retrieved May 3, 2010, from IEEE Xplore database.

Saunders, J. (1996). Real-time discrimination of broadcast speech / music. *IEEE international conference on acoustics, speech and signal processing (ICASSP), 2,* 993-996. Retrieved May 4, 2010, from IEEE Xplore database

Scheirer, E., Slaney, M. (1997). Construction and evaluation of a robust multifeature speech/music discriminator. *IEEE international conference on acoustics, speech and signal processing (ICASSP), 2,* 1331–1334. Retrieved May 4, 2010, from IEEE Xplore database.

Schneiders, M.G.E. (2001). *Wavelets in control engineering. Master's thesis, Eindhoven University of Technology, 1-85.* Retrieved May 28, 2010, http://alexandria.tue.nl/repository/books/551516.pdf

Selesnick, I.W. (2001). The Design of Hilbert Transform Pairs of Wavelet Bases. *IEEE Signal Processing Letters, 8(6),* 170-173. Retrieved June 11, 2009, from IEEE Xplore database.

Selesnick, I. W., Baraniuk, R. G., Kingsbury N. G. (2005). The Dual Tree Complex Wavelet Transform. *IEEE Signal Processing Magazine,22(6),*123-151. Retrieved December 10, 2009, from IEEE Xplore database.

Strang, G. & Nguyen, T. (1997). *Wavelets and Filter Banks (Second Edition).* Massachusetts: Wellesley -Cambridge Pres.

Sumbera J. (2001). *Wavelet Transform using Haar Wavelets.* Retrieved May 28, 2010,from http://www.jikos.cz/~sumbera/vyplody/wavelet/Jiri_Sumbera_Wavelet_Transform_using_Haar_Wavelets.pdf

Tancerel, L., Ragot, S., Ruoppila, V.T., Lefebvre, R. (2000). Combined speech and audio coding by discrimination. *IEEE workshop on speech coding,* 17–20. Retrieved May 4, 2010, from IEEE Xplore database.

Thiran, J. P. (1971). Recursive digital filters with maximally flat group delay. *IEEE Tranactions on Circuit Theory, 18(6),* 659–664. Retrieved from June 11, 2009, from IEEE Xplore database.

Tzanetakis, G., Essl, G., Cook, P. (2001). Audio Analysis using the Discrete Wavelet Transform. *Proceedings of WSES International Conference on Acoustics and Music: Theory and Applications,* 1-6. Retrieved June 30, 2009, from CiteSeerX database.

Wang, W.Q., Gao, W., Ying, D.W. (2003). A fast and robust speech/music discrimination approach. *IEEE. 4th pacific rim conference on multimedia,3.*1325– 1329. Retrieved May 4, 2010, from IEEE Xplore database.

Wang, J., Wu, Q., Deng, H., Yan, Q. (2008). Real-time speech/music classification with a hierarchical oblique decision tree. *IEEE international conference on acoustics, speech and signal Processing (ICASSP),* 2033–2036. Retrieved May 4, 2010, from IEEE Xplore database.

Zhang, T., Kuo, J. (2001). Audio content analysis for online audiovisual data segmentation and classification. *IEEE Transactions on Speech and Audio Processing 9(4),* 441-457. Retrieved April 29, 2010, from http://viola.usc.edu/Publication/PDF/selected/2001_IEEE-TSAP_Zhang.pdf

Zheng, F., Zhang, G.,Song, Z., (2001). Comparison of Different Implemantations of MFCC. *Journal of Computer Science and Technology, 16(6),* 582–589. Retrieved June 30, 2009, from SpringerLink database.