DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

SPEECH PROTECTED ACTIVE NOISE CONTROL SYSTEM

by

Özge CANLI

January, 2015

İZMİR

SPEECH PROTECTED ACTIVE NOISE CONTROL SYSTEM

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

by
Özge CANLI

January, 2015
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"SPEECH PROTECTED ACTIVE NOISE CONTROL SYSTEM"** completed by **ÖZGE CANLI** under supervision of **ASST. PROF. DR. HATİCE DOĞAN** 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. Hatice DOĞAN

Supervisor

Asst. Prof. Dr. Güleser K. Demir

(Jury Member)

Asst. Prof. Dr. Deniz Ekin Akyol

(Jury Member)

Prof. Dr. Ayşe OKUR
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

I would like to thank my supervisor, Hatice DOĞAN, for providing me the opportunity to pursue and finish this degree; her constant encouragement, guidance and mentorship helped me a lot in completing this thesis.

I would like to thank Hakan KAHRAMAN for being a perfect project mate and a good friend during the thesis program.

I am also grateful to Şerif USTA for his support, either mentally and physchogically.

I would also like to thank my family and friends for all their support and inspirations over the years.

Özge CANLI

SPEECH PROTECTED ACTIVE NOISE CONTROL SYSTEM

ABSTRACT

Noise can be described as the unwanted sounds. This unwanted sound causes permanent hearing damage when it is too loud. Therefore, the usage of hearing equipment is needed in noisy environments. The usual noise Control systems eliminate both unwanted sounds and desired sounds. In this thesis, an active noise control system which preserves the speech while eliminating the noise in noisy environments is proposed. With the help of this system a worker will be able to hear speech in a noisy environment. The software of the proposed system which is the combination of Voice Activity Detection (VAD), Independent Component Analysis (ICA) and Active Noise Control (ANC) blocks is developed. As a first step, VAD block determines that the speech is contained in the environment or not. At the second step, ICA block separates the signals only if VAD block makes a decision that the speech exists there. From the output of the ICA block two signals are obtained. To decide which one is the noise, variance of autocorrelation which is a simple feature is used. Then the noise is eliminated by the ANC block. ANC block is designed with the Fx-LMS algorithm. The system performance is tested on a dataset which is generated by taking the several samples from different noise types. These sounds are mixed with the speech by using a matrix which is nonsingular symmetric at five different SNR levels. The dataset contains 11600 samples of which length is 12 seconds. Simulations show that the performance of the proposed system highly depends on the VAD block. If this block performs well, ANC protects the speech and eliminates the noise. To obtain better performance a new feature based on Mel frequency is proposed and the performance of this feature has better results considering the standard ones. The whole system performance is evaluated by comparing the input SNR level with the output SNR level. The result shows that the proposed system performance is great, even for the low SNR levels.

Keywords: Active noise control, independent component analysis, voice activity detection

KONUŞMA KORUMALI AKTİF GÜRÜLTÜ KONTROL SİSTEMİ

ÖZ

Gürültü istenmeyen sesler olarak tanımlanabilir. Bu istenmeyen ses yüksek olduğunda kalıcı duyma hasarlarına yol açmaktadır. Gürültü kontrol sistemleri istenen sesler de dâhil olmak üzere ortamdaki tüm sesleri yok eder. Bu tezde gürültüyü yok ederek, gürültünün yok edilerek konuşmanın gürültü ortamda bozulmadan korunmasını sağlayan bir gürültü kontrol sistemi önerilmiştir. Bu sistemin yardımıyla, işçi gürültülü ortamda konuşmayı rahatlıkla duyabilecektir. Bu tezde önerilen sistemin, ses etki tespiti, bağımsız bileşenler analizi ve aktif gürültü sistemi bloklarının da içerisinde bulunduğu yazılımı geliştirilmiştir. Kullanılan veri setleri, farklı türde gürültü kaynaklarından pek çok örnek alınarak oluşturulmuştur. Bu sesler konuşma örnekleri ile birleştirilmiştir. Veri setinin son hali zamansal uzunluğu 12 saniye olan 11600 sesi içermektedir. İlk aşamada, ses etki tespiti bloğu ortamda konuşmanın olup olmadığını belirler. Eğer ses etki sistemi bloğu ortamda konuşmanın olduğu kararını verirse ikinci aşama olarak bağımsız bileşenler analizi bloğu sinyalleri birbirinden ayırır. Bağımsız bileşenler analizi bloğunun çıkışında iki farklı sinyal elde edilir. Bunlardan hangisinin gürültü olduğunu anlamak için basit bir sinyal karakteristik özelliği olan özilinti varyansı kullanılır. Daha sonra, gürültü olduğu belirlenen sinyal aktif gürültü sistemi bloğu tarafından elenir. Bu aktif gürültü sistemi bloğu, en küçük ortalama kareler algoritması kullanılarak geliştirilmiştir. Simülasyon sonuçları, öne sunulan sistem performansının büyük oranda ses etki tespiti bloğuna bağlı olduğunu göstermiştir. Eğer bu ses etki tespit bloğu verimli ise, aktif gürültü sistemi konuşma sinyalini korur ve gürültü sinyalini eler. Daha iyi bir performans elde etmek için, Mel frekans tabanlı olan yeni bir öznitelik önerilmiştir ve bu konuşma tespitinin başarımı yazında önerilen diğer özniteliklerden daha iyidir. Tüm sistem performansı giriş sinyal gürültü oranı ve çıkış sinyal gürültü oranı karşılaştırılarak değerlendirilmiştir. Sonuçta görülmüştür ki, önerilen sistemin performansı düşük seviyeli sinyal gürültü oranlarında bile çok iyidir.

Anahtar kelimeler: Aktif gürültü sistemi, bağımsız bileşenler analizi, ses etki tespiti

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Noise can be described as the unwanted and unpleasant sounds. The consequence of noise exposure is a vital problem for people. This unwanted sound causes permanent hearing damage when it is too loud. For example, if a person is exposed to noise above 85db during 8 hours in a day, it causes noise induced hearing loss (Noise Control: A practical approach to controlling noise in the workplace (n.d.)). Therefore, the usage of hearing equipment is needed in noisy environments.

The workers have to use ear plugs or ear muffs in the industrial places where the noise level is harmful for the human health. In recent years, noise cancelling headphones have also become popular in industry and workplaces. Although a worker wears hearing protection equipment and prevents himself from the effect of noise, his communication with the other people cuts off. This situation is undesirable because it is known that speech communication is one of the fundamental requirements for industrial workplaces.

So, a hearing protection equipment which suppresses the noise but prevents the speech is necessary. In this thesis, a software of such an equipment is proposed. In this system, noise will be suppressed by using active noise control (ANC) methods and speech will be separated from the noise by using independent component analysis (ICA) after it is detected by using voice activity detection (VAD). By the implementation of these methods, a worker will be able to hear speech in a noisy environment.

So the proposed system is the combination of VAD, ICA and ANC methods. If the speech is not contained in the environment, only the noise elimination step is applied. On the contrary, if the speech is contained in the environment, it is separated by independent component analysis and it is transferred into the headphone. Thus, the noise elimination is only applied to noise signals. The algorithm of the software is shown Figure 1.1.

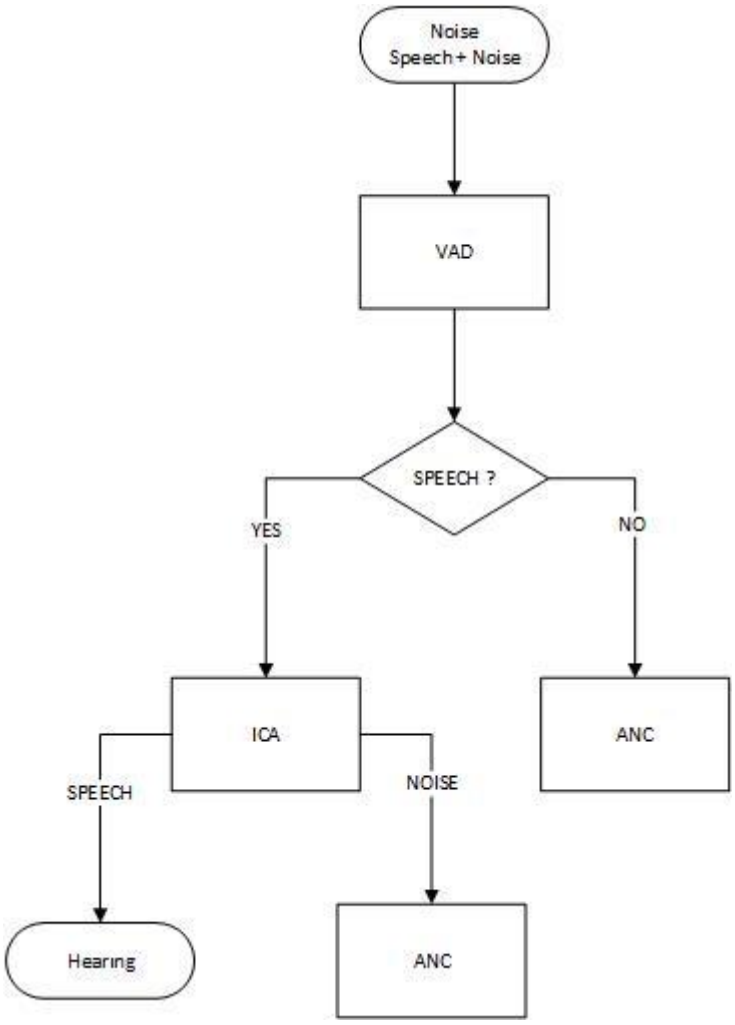

Figure 1.1 Structure of the proposed system

There are two similar applications in the literature. One of them is a patent application for a noise cancelling headset that evaluates the content of an external audio signal and decides whether this signal should be suppressed or preserved (Trajkovic, Gutta & Cohen, 2002). This system suppresses the other audio signals but transfers the desired signals such as an emergency announcement or alarm to the user (Trajkovic, Gutta & Cohen, 2002). Since this is a patent application, technical specifications are not given in detail.

Another application is a system proposition for noise cancelation which can selectively cancel a particular noise signal from a mixture (Sohn & Lee, 2000). In this paper, blind source separation (BSS) algorithm separates the noise and the desired sound, afterwards ANC cancels the noise. While standard ANC algorithms

cancel both a desired sound and a noise signal, the constructed selective attention system is able to cancel only noise signal (Sohn & Lee, 2000). As clearly stated in (Sohn & Lee, 2000), it is assumed that two signals always exist in the environment. Therefore, BSS algorithm and ANC algorithm are used all the time without activating any VAD algorithms. However, there is another disadvantage of this application. BSS algorithm can separate the independent source signals by using the observation signals. In this application it is assumed that the separated speech signal is always given to the first output channel but in real life, this information is unknown. That means the noise can also be given to the first channel by BSS algorithm. In our proposed system, the output channels are checked whether they contain noise or speech. Thus, the other signal, noise, is eliminated. Using a VAD system is the main difference between our proposed system and the constructed selective attention system. The advantage of the system that is proposed in this thesis is the fact that there is no need to run BSS algorithm if the speech is not contained in the environment. In this case, only the noise elimination is carried out. Since the performance of the proposed system highly depends on the VAD block, a new feature which can be called as "modified Mel-spectrum" is also proposed in this thesis.

1.1 Voice Activity Detection

VAD or speech detection is a mechanism that determines whether a speech signal is contained in an audio signal. It has been widely used in speech processing areas and audio applications including speech coding (Enqing, Heming & YongLi, 2002), noise reduction for hearing aids (Itoh & Mizushima 1997), and speech transmission on the Net (Prasad, Sangwan, Jamadagni, Chiranth & Sah, 2002). Furthermore, it has long been known that VAD is commonly used in speech recognition, speech enhancement and audio indexing applications (Moattar & Homayounpour, 2009).

There are several proposed algorithms in the literature for voice activity detection. In these algorithms, different feature extraction and classification methods have been used. It is easy to detect the speech signal when Signal to Noise Ratio (SNR) is high

and the noise is stationary. In this case, VAD algorithms perform well by using even the short-time energy and zero-crossing rate which are known easily extracted features (Moattar & Homayounpour, 2009). However, it is hard to distinguish the speech signal when the SNR is low and noise is non-stationary. Although many different features and classifiers are used, the performance of VAD falls significantly in these cases. As a result, in recent years, researchers focus on the design of VAD algorithms which perform well under different noise types especially on the cases where the noise is non-stationary and SNR is low.

The main common requirements in a VAD system are robustness, real-time processing, accuracy, simplicity and having no knowledge about the noise (Moattar & Homayounpour, 2009). To achieve the real-time processing, the short-time features are generally used. On the other hand, long-time features have better performance and are more robust than the short time features. Robustness and real-time implementation are the most crucial features of VAD system.

Up to date, there are different features which have been proposed for VAD systems. VAD features can be classified into groups. Energy based features which had been used in earlier VAD systems are simple to extract but ineffective at low SNR. Spectral-domain features which give clues about frequency content of a signal are more robust than energy based features. In contrast to short-time features, long-term features which are proposed in recent years have the best performance (Ghosh, Tsiartas & Narayanan, 2011); (Ramirez, Segura, Benitez, Torre & Rubio, 2004). Furthermore, different features are represented in literature, and different classifiers have been used such as Gaussian Mixture Model (GMM), Artificial Neural Network (ANN), Support Vector Machine (SVM), Hidden Markov Model (HMM) and the nearest neighbor algorithms for classification of features (Kos, Kačič & Vlaj, 2013). Some earlier VAD systems have become standards in literature such as G.729 (Benyassine, Shlomot, Huan-yu & Massaloux, 1997), and ETSI AMR (Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels, 1998). In this thesis G.729 VAD standard is used for VAD. For example, G.729 VAD standard computes four different measures which are full band and low band

energies, line of spectral frequencies and zero-crossing rate, then it takes decisions for each frame based on the initial frame values (ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, 2012).

1.2 Independent Component Analysis

ICA is a source separation algorithm that separates the independent sources by observing the mixture of signals. Only prior information needed for this algorithm is the number of sources. It is developed by Hyvärinen and based on the fact that independent source signals could be estimated from mixture signals. These signals are defined as linear combination of source signals. It has various applications including audio processing, medical analysis, financial data analysis and mobile communication (Hyvärinen & Oja, 2000).

1.3 Active Noise Cancellation

By ANC algorithm, the noise is eliminated by generating an anti-noise which has same amplitude but has 180 degree phase shift when comparing with the noise (Elliott & Nelson, 1993). The summation of the noise and anti-noise is residual signal which can be barely heard. If a signal is sinusoid or pure tone signal, the performance of ANC is high. However, it is difficult to cancel multiple tones, higher frequency signals and impulsive sounds. Therefore, it is important to know the frequency, amplitude and phase of the noise for the elimination (Thom, Peters & Winters, 2005).

1.4 Thesis Organization

This thesis is organized as follows. Chapter two describes the voice activity detection including various features, and decision rules which are proposed in the literature for VAD system. While the features have the categories such as energy and frequency based, long-term and harmonicity based, the decision rules are represented as thresholding, statistical model approaches and machine learning approaches. Their

advantages and disadvantages are discussed in this chapter. The standard ITU-T G.729 VAD system is used in this thesis and its detail information is also given in this chapter. The different features proposed in the literature are documented. Short-term features and long-term features are explained in a detailed way. Also, modified Mel-spectrum feature is briefly represented.

In chapter tree, separation of mixtures is given in details. Independent component analysis and Fast-ICA algorithms are examined, and the definitions, properties and equations are defined. The different mixing matrices are evaluated and their graphical representations are also demonstrated in this chapter.

The chapter four includes the theory of ANC. ANC terms and algorithms are explained such as Least Mean Square (LMS) and Filtered Input Least Mean Square (Fx-LMS). Also, in the fourth chapter, the selective system using new active noise controller is explained and given in details as an example of Blind Source Separation (BSS) and ANC system.

Results are given in the fifth chapter. This chapter contains the performances of the VAD, ICA and ANC system. Database and data preparation stages are introduced in this chapter. Performances of the methods are calculated for the various sound types at different SNR levels. The modified Mel-spectrum feature for VAD is proposed and its performance is compared with different features proposed in the literature. The non-singular symmetric mixing matrix is performed for ICA system. The performance of ICA is computed. The graphical results of the ANC system are also added in this chapter. For the performance evaluation of the whole system output SNR level is calculated and compared with the input SNR level. The performance of the proposed system is presented at different types of noise with different noise levels. The comparison of input and output SNR levels are also mentioned in this chapter.

Finally, chapter six is the conclusion part and it summarizes and concludes the thesis.

CHAPTER TWO
VOICE ACTIVITY DETECTION

Voice activity detection is a process that discriminates speech from the silence or the background noise. It has two critical steps. The first step is the feature extraction which is a crucial step to distinguish speech and non-speech. The more discriminative feature brings more robust classifier for VAD. The other step is to make speech/non-speech decision by using classifiers. Decision part has also significant importance while performing VAD.

The first requirement for VAD system is the extraction of good features. These features must have distinctive properties when comparing speech and noise (robustness). Moreover, low-complexity is important when implementing real-world applications. There are many features proposed in the literature, and their properties are mentioned in the following part.

2.1 Feature Selection for Voice Activity Detection

It has long been known that speech/non-speech discrimination is not an easy task in noisy environments. Up to now, many features have been proposed to deal with this problem. In addition to the short time features such as Spectral Flux (SF), Spectral Roll-off (SR), Spectral Centroid (SC), Spectral Entropy (SE), Zero Crossing Rate (ZCR), Root Mean Square (RMS), Low Energy Frames (LEF), the long term features like Long Term Spectral variability (LTSV) are also quite efficient. These features which are commonly used in literature are explained in a detailed way.

2.1.1 Short-Time Features

As shown in Figure 2.1, a speech signal has non-stationary characteristic which means that its statistical properties change with time. However, when the speech signal is analyzed with short-time windows, it is seen that the speech signal can be considered as stationary. Short-time analysis is well-known and widely used

technique when investigating the speech signal. On the other hand, the useful spectral information can be obtained by means of Fourier analysis. Fourier analysis is an approach to express the signals as the summation of sinusoidal signals weighted by Fourier coefficients. In addition to this, Fourier analysis provides a good spectral representation and it gives clues about frequency content of the signal.

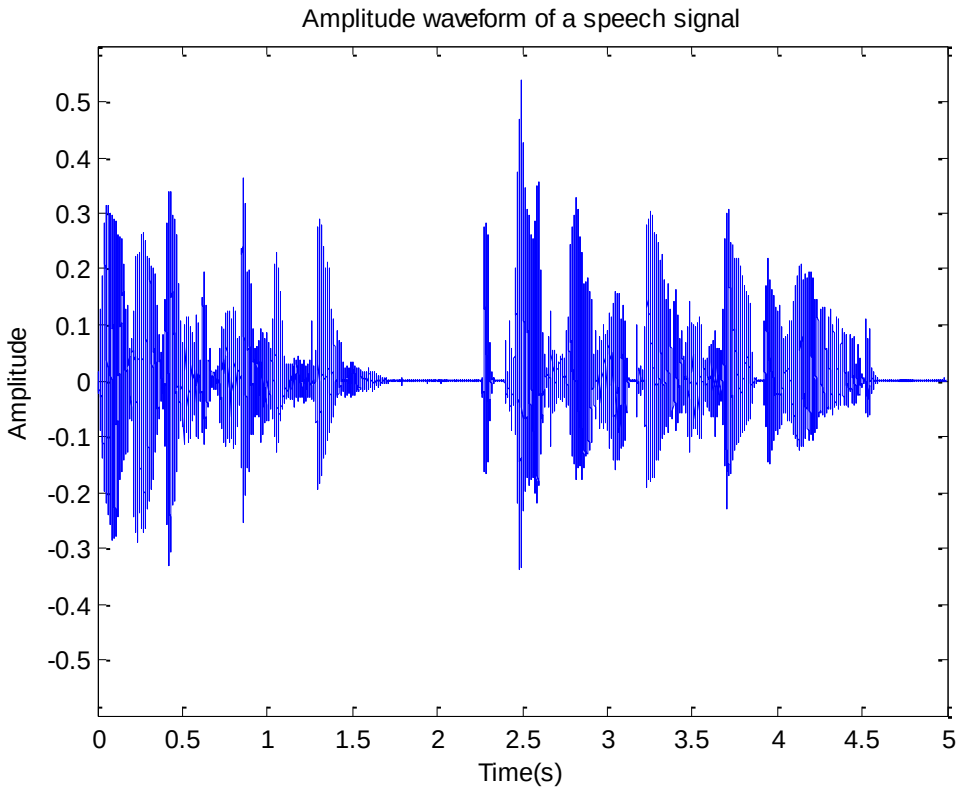

Figure 2.1 Amplitude waveform of speech at 44.1 kHz sampling frequency.

Since Fourier analysis is insufficient to focus on non-stationary signals; the Fourier transform of short time window is calculated. Short-Time Fourier Transform is efficient to analyze the content of the local window. Initially, the signal is divided into small parts (to be considered stationary) and then Fourier transform is applied to each small part. The STFT is given as time series:

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega m} \tag{2.1}$$

where ω symbolizes frequency, n is the frame number, $w[n]$ is analysis window which is assumed to be non-zero only in the interval *[0,N_w-1]* (Quatieri, 2002). Also, the discrete STFT is computed by sampling $X(n,\omega)$ over the unit circle.

$$X(n,k) = X(n,\omega) = x[m]w[n-m]e^{-j\frac{2\pi}{N}km} \tag{2.2}$$

where N is the frequency sampling factor and $\omega = 2\pi/N$ is the multiples of frequency sampling interval (Quatieri, 2002). Short time features can be classified into two groups: a) energy and frequency based features, b) harmoniticity and autocorrelation based features.

2.1.1.1 Energy and Frequency Based Features

Energy of a speech frame is calculated by taking the root mean square of the amplitude value (RMS). It has shown a lot differences due to the nature of speech and has been used in early VAD algorithms (Prasad, Sangwan, Jamadagni, Chiranth & Sah, 2002). It is assumed that speech signal has a higher energy value than noise signal when SNR level is high. However, this feature is not discriminative at the low SNR conditions.

Another short-time feature is the zero-crossing rate that is used to determine the number of zero-crossings in a frame (Lu, Pan, Lane, Choudhury& Campbell, 2009). Other features can be used in VAD system such as spectral line frequencies (Benyassine, Shlomot, Huan-yu&Massaloux, 1997) and percentage of low energy frame ratio (Kos, Grašič&Kačič, 2009). All these features are similar to energy measure; but they are bad at distinguishing speech from the noise when SNR level is low.

Frequency based features or spectral features are computed by means of STFT and are calculated for each frame (usually taken between 20ms and 40ms). They are frequently used in speech applications because they distinctively represent speech's spectrum. For instance, spectral flux, spectral centroid and spectral roll-off features

characterize the shape of signal's spectrum in the speech/non-speech detection mechanism. Therefore, (Kos, Kačič & Vlaj, 2013) have been used them in the acoustic classification. Entropy of speech frames is different from that of the noise signals (Tüske, Mihajlik, Tobler[+] & Fegyó[+], 2005). Entropy of magnitude spectrum indicates that speech frames are more organized than the noise frames. According to this idea, a robust voice activity detector can be implemented by using the short time spectral entropy feature.

2.1.1.1.1 Zero Crossing Rate (ZCR). Zero Crossing Rate is a time-based feature that determines the number of zero-crossings in a frame (Lu, Pan, Lane, Choudhury & Campbell, 2009). At first, a frame which has a constant length is defined and then how many sign changes in this frame is counted. According to the definition of ZCR, it demonstrates that the number of ZCR for music and noise signal has a high ratio but its variance has small scale. Also, the number of ZCR for speech signal has low ratio while its variance has notably large scale. In summary, ZCR focuses on the frequency content of a signal.

$$ZCR = \frac{1}{2N}\sum_{n=1}^{N}|sgn[x(n)] - sgn[x(n-1)]| \tag{2.3}$$

where *N* is number of samples in a frame and *sgn* is the signum function defined as follows (2.4).

$$sgn(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases} \tag{2.4}$$

2.1.1.1.2 Spectral Flux (SF). It is known that spectral flux represents the change of signal's spectrum magnitude between the current frame and the previous frame ratio (Kos, Grašič & Kačič, 2009). In other words, it is described as Euclidian norm of adjacent frames. Spectral flux is generally used in speech/music discrimination because speech signals have lower ratio when they are compared to music signals. Due to similar reasons (to have different characteristics between speech and noise), it is also a good feature while analyzing speech and noise signals.

$$SF = \frac{1}{M}\sum_{k=1}^{M}\sqrt{\left(X_i(k) - X_{i-1}(k)\right)^2} \tag{2.5}$$

where *M* is the total number of frequency bins, *i* is the frame index, *k* is the index of frequency bin, $X_i(k)$ is the current frame's magnitude of Fourier transform and $X_{i-1}(k)$ is the previous frame's magnitude of Fourier transform, respectively.

2.1.1.1.3 Spectral Roll-off (SR). It is crucial to have knowledge of the frequency spectrum. Spectral Roll-off is defined as the frequency bin below which 95% of the magnitude spectrum is concentrated (Kos, Grašič & Kačič, 2009). For example, speech has lower energy band than noise so this feature is another way to represent signal's spectrum distinctively.

$$\sum_{k=1}^{R} X(k) = 0.95 \sum_{k=1}^{M} X(k) \tag{2.6}$$

where *M* is the total number of frequency bins, *k* is the index of frequency bin, *R* is spectral roll-off number and the magnitude of Fourier transform is $X(k)$ respectively.

2.1.1.1.4 Spectral Centroid (SC). Spectral Centroid is another feature that characterizes the shape of signal's spectrum. It gives tips about the center of the magnitude of Fourier transform (magnitude spectrum of Fourier transform), (Kos, Grašič & Kačič, 2009). If a signal contains high frequency components, it means

spectral centroid has higher value rather than low frequency signals. In short, spectral centroid is a measure of where power distribution of signal focuses on.

$$SC = \frac{\sum_{k=1}^{M} kX(k)}{\sum_{k=1}^{M} X(k)} \tag{2.7}$$

where *M* is the total number of frequency bins, *k* is the index of frequency bin and the magnitude of Fourier transform is $X(k)$, respectively.

2.1.1.1.5 Spectral Entropy (SE). Entropy is commonly known as a measure of disorder. In signal analysis, spectral entropy is generally used and measures disorganization of signals. While spectral entropy is a distinctive feature when a signal changes slowly, it's not distinguishable for complex noises (Tüske, Mihajlik, Tobler^{+}&Fegyó$^{+}$, 2005). In short-time spectrum analysis, the idea comes from a white noise signal has the highest entropy because it represents an equilibrium point in signals (Ekštein & Pavelka, 2004). While the noise frame has the highest entropy, speech frame has lower entropy than noise frame. In other words, the speech frame needs an extra energy to equilibrate this point.

To obtain spectral entropy, the power spectrum of signal is calculated with the help of Fourier transform, it is normalized so it can be treated as a probability density function (PDF) (Tüske, Mihajlik, Tobler^{+}&Fegyó$^{+}$, 2005). Finally, the entropy formula is applied to each PDF, this gives the spectral entropy of short-time frame.

$$H(s) = -\sum_{i=1}^{N}(p_i)\, log_2(p_i) \tag{2.8}$$

where p_i is probability density function of normalized power spectral density, between 0 and 1 respectively.

2.1.1.1.6 Voice2White (V2W). The idea is to measure ratio of the energy of the typical speech band and the whole energy of signal (Alexendre, Rosa, Cuadra & Gil-

Pita, 2006). The typical speech band takes part between 100Hz and 4kHz. Thus, speech signal has the higher value when comparing to noise signal.

$$V2W = 10\,log\frac{\sum_{100Hz}^{4kHz}|X_t[k]|^2}{\sum_{\forall k}|X_t[k]|^2} \tag{2.9}$$

where k is the index of frequency bin , t is the frame index, $X_t[k]$ is the current frame's magnitude of the Fourier transform respectively.

2.1.1.2 Harmonicity and Autocorrelation Based Features

Although frequency based features performs well for stationary signals, they can fail for non-stationary signals. In order to deal with non stationary signals, the periodic measurement and autocorrelation based features have been used in recent years. One of the supposed ideas is to use the harmonics of the signal. Harmonics of speech are sustained over a certain span of time and vary in the frequency domain while harmonics of the noise have specific peaks and remain stable in the frequency domain (Sonnleitner, Niedermayer, Widmer&Schlüter, 2012). The cross-correlation of two adjacent frames is calculated during computation of the feature. Thus, it gives clues about harmonic patterns in the speech detection algorithms. Also, another idea is derived from the inherent characteristics of speech. Voiced sounds in speech signal have dominant periodic property while unvoiced sound and noises have minor peaks in the frequency domain. This periodic property is discriminative measure and can be combined with the autocorrelation function in the feature extraction part (Wu & Wang^{+}, 2006).

2.1.1.2.1 The Correlation Gain. It is expected that harmonic analysis of signals (speech, noise and musical instruments) is a discriminative measure (Sonnleitner, Niedermayer, Widmer&Schlüter, 2012). Music or noise signal which has horizontal tone in the spectrum and speech signal which has peaks in the spectrum due to its inherent characteristics are significantly different while determining speech/music or speech/noise detection. Therefore, this feature calculates cross-correlation between

X_t and $X_{t+offset}$ frames and then determines the correlation gain which is the difference between maximum cross-correlation lag and zero-lag cross correlation. Whereas the correlation gain is zero for music or noise signal, speech signal has a positive gain by means of the harmonic patterns.

$$R_{X_t, X_{t+offset}}(l) = \sum_t X_t \, X_{t+offset} \tag{2.10}$$

X_t and $X_{t+offset}$ are two given vectors of length *N*, the cross correlation for all lags $l \in [-N, N]$ so series of length *2N+1*.

$$r_{xcorr}(X_t, X_{t+offset}) = \max_l R_t, R_{t+offset}(l) \tag{2.11}$$

$$r(X_t, X_{t+offset}) = R_t, R_{t+offset}(0) \tag{2.12}$$

$$\text{The correlation Gain} = r_{xcorr} - r \tag{2.13}$$

2.1.2 Long Term Features

Long term features have become popular in recent years. For example, long term spectral divergence (LTSD) feature measures spectral divergence in the long frames and calculates the ratio of the long term spectral divergence to average noise spectrum (Ramirez, Segura, Benitez, Torre & Rubio 2004). Although this feature works well in different noise conditions, it depends on the average noise spectrum knowledge. Hence, it is not a preferable situation in real world scenarios.

The second example of the long term features is the long term spectral variability (LTSV). It analyses speech in 150-250 ms window, this makes speech detection activity more robust in comparison to short-term features (Ghosh, Tsiartas & Narayanan, 2011). It measures the non-stationary degree or variability of signals. It has known that the analysis frame of speech is taken short windows in order to

provide stationary in early researches. However, by taking long-term windows, the speech signal becomes non-stationary.

2.1.2.1 Long Term Spectral Variability (LTSV)

Long Term Spectral Variability analyses speech in 150-250ms window hence it make speech-detection activity more robust in comparison short-term features (Ghosh, Tsiartas, & Narayanan, 2011). The idea comes from the long-term non-stationary degree or long term spectral variability of speech signals has higher value than noise values. As a result, this property becomes distinctive measure in the speech detection.

$$\mathcal{L}_x(m) \triangleq \frac{1}{K}\sum_{k=1}^{K}\left(\xi_k^x - \overline{\xi^x(m)}\right)^2 \tag{2.14}$$

where $$\overline{\xi^x(m)} = \frac{1}{K}\sum_{k=1}^{K}\xi_k^x\,(m) \tag{2.15}$$

and

$$\xi_k^x(m) \triangleq -\sum_{n=m-R+1}^{m}\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m}S_x(l,\omega_k)} \times log\left(\frac{S_x(n,\omega_k)}{\sum_{l=m-R+1}^{m}S_x(l,\omega_k)}\right) \tag{2.16}$$

R is the number of consecutive frames. The short-time spectrum at ω_k frequency is symbolized $S_x(n,\omega_k)$. It is calculated by $\mathrm{S_x(n,\omega_k)} = |\mathrm{X(n,\omega_k)}|^2$, where

$$X(n,\omega_k) = \sum_{l=(n-1)N_{sh}+1}^{N_\omega+(n-1)N_{sh}}\omega(l-(n-1)N_{sh}-1)x(l)e^{-j\omega_k l} \tag{2.17}$$

$X(n,\omega_k)$ is the short-time Fourier transform (STFT) coefficient at frequency ω_k, n represents frame number, N_ω is the frame length, N_{sh} is the overlapping length, $\omega(i)$ is the short-time window $0 \le i < N_\omega$.

From (2.14) to (2.17) formulas are associated with spectral entropy feature. However, the idea comes from using long-term spectral variability in signals so that the feature value is extracted from the last R adjacent frames by computing their entropy in (2.14), ending at m^{th} frame (Ghosh, Tsiartas, & Narayanan, 2011). The performance of the feature discrimination is dependent on choice of R value. Another criteria when extracting features is the selection of K which is a range of frequency value. With the help of K value, the sample variance of entropies is computed in the range of K frequency, named as $\mathcal{L}_x(m)$ (2.14). LTSV feature gives different results due to the selection of K, R and noise type. In generally, if noise is stationary, speech signal has the lowest entropy. The entropy of speech plus noise signal has greater than only speech case but smaller than only noise signal. Moreover, the entropy of only noise case has the highest value, the range of speech plus noise case is dependent on the SNR selection (Ghosh, Tsiartas, & Narayanan, 2011).

2.1.2.2 Modified Mel-spectrum Feature

A spectrogram is a representation of amplitude of signal at any given frequency and time. When the speech spectrogram is analyzed, local peaks can be easily seen (especially SNR is 10db case) between 100 Hz and 4 kHz. In this work, the local peaks in speech spectrogram are analyzed, after that a new feature which is based on Mel frequency is proposed.

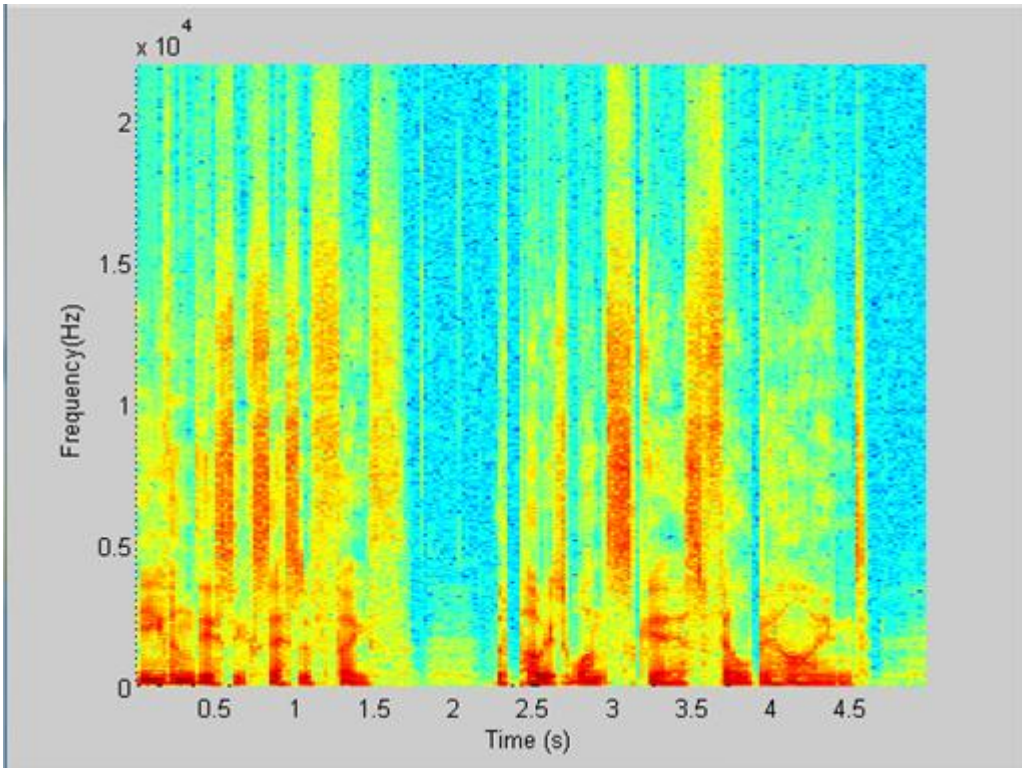

Figure 2.2 Spectrogram of a speech signal

The frame length is chosen 300ms for this feature. First of all, it is crucial to get information from spectrogram. As a result, the square of the current frame's magnitude of Fourier transform is computed so that the power of signal spectrum is calculated. It indicates power of signal at any given frequency and time relating to spectrogram definition. Another way to illustrate the spectrogram is to convert it to mel-frequency domain instead of frequency domain. Mel-frequency transformation is frequently used in speech detection, speech recognition and music information retrieval areas. It is based on the perceptual human-ear modeling. It is known that human hear captures sounds in linear scale between 0 Hz and 1 kHz, but above 1 kHz capturing sounds becomes logarithmic scale. The Mel-frequency formula is improved by this assumption.

The general formula conversion from Hertz to Mel-frequency is:

$$m = 2595 \times log_{10}(1 + \frac{f}{700}) \tag{2.18}$$

The mel-frequency spectrum is calculated by multiplying power spectrogram by each of the Mel Weighting filters (Zhang, 2003).

$$\check{P}[l] = \sum_{k=0}^{N/2} P[k] M_l[k] \tag{2.19}$$

where $P[k]$ is the frequency domain power spectrogram and $\check{P}[l]$ is the mel-frequency domain power spectrogram, N is the length of Fourier transform, L is the total number of Mel weighting filter $l = 0,1,2,..,L-1$.

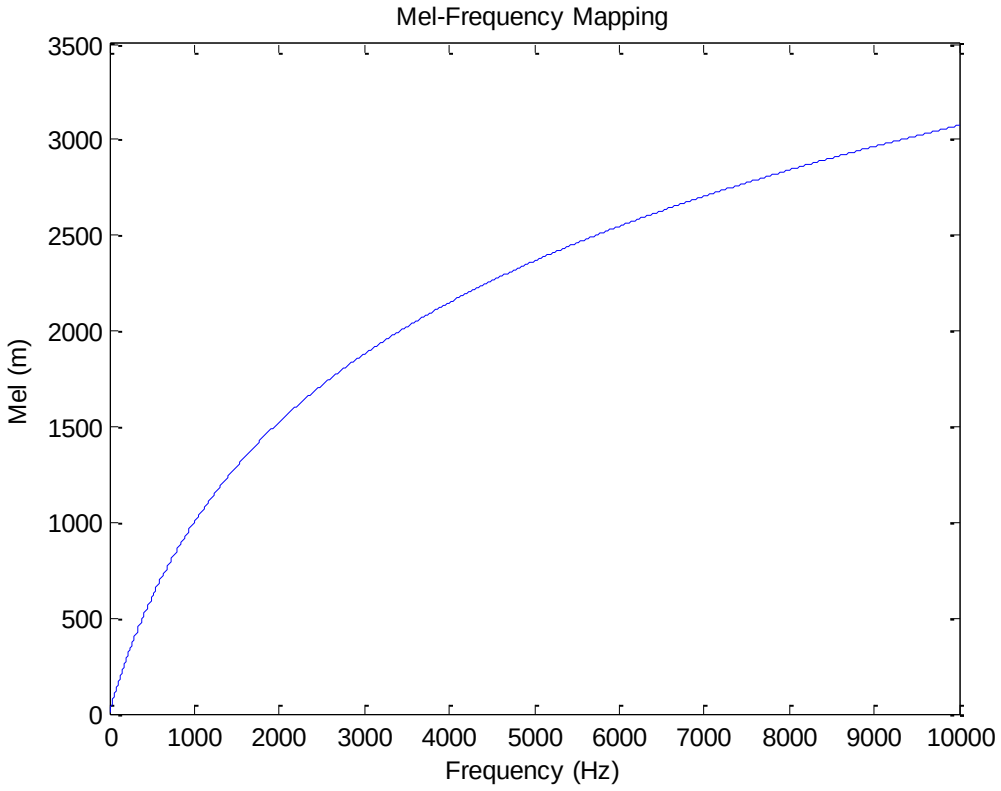

Figure 2.3 Mel-frequency mapping

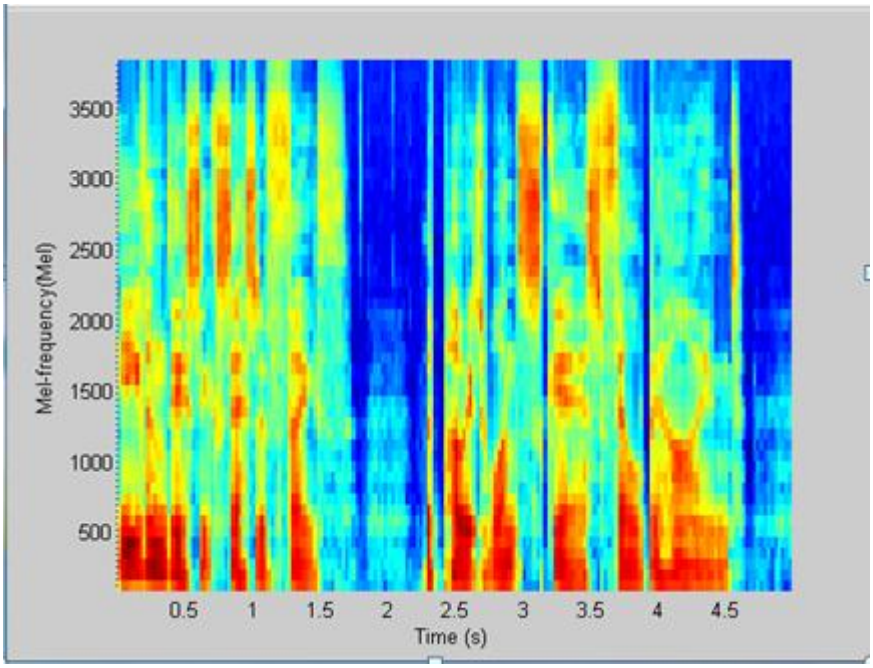

Figure 2.4 Mel-frequency spectrogram of a speech signal

After the computation of mel-frequency spectrogram and mel-frequency power at any given bins, the specific mel-bins are selected for the discrimination of speech parts. The bins for this feature are selected between 4 and 13 for L=50 bins. These bin centers represent 307mels and 1000mels. As it is known, during the conversion from frequency domain to mel-frequency domain, the filters are uniformly spaced on the mel-frequency axis (Prahallad, 2008). However, the number of filters in low frequency band is more frequent than number of filters in high frequency band in frequency domain.

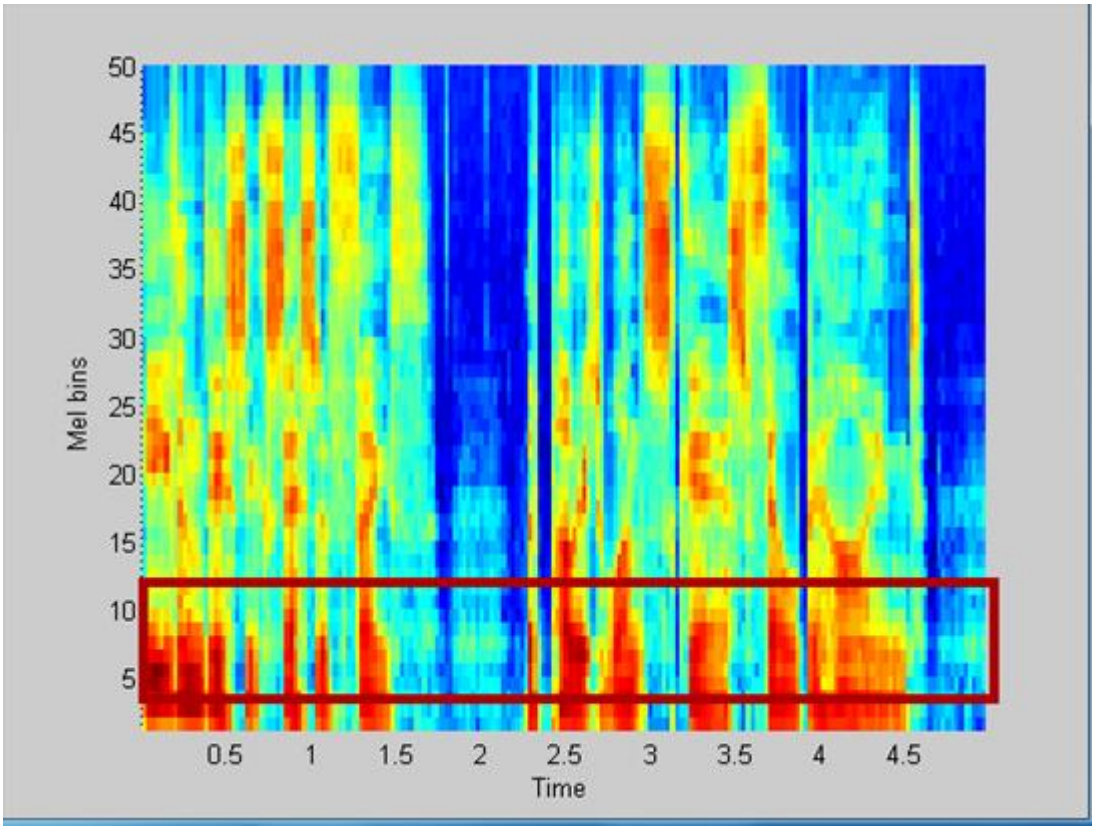

Figure 2.5 Selected bins of mel-frequency spectrogram

The power of selected mel-bins is calculated and then its average value is computed. To make more robust this feature even in low SNRs, the standard deviation of average value is computed. This value represents the modified Mel-spectrum feature.

In the literature, discrete–cosine transform (DCT) is also computed after transforming mel-frequency domain. When DCT is computed, the spectral shape of signal is analyzed. It can be clearly seen that the properties of mel-spectrum magnitudes would be good as the spectral shape of signal. Therefore, the power of mel-bins at specific frequencies is also important in mel-frequency analysis.

In summary, it has seen that there are various features to deal with speech/non-speech decision. The main issues are the robustness and low-complexity at the selection of features. By categorizing them, their advantages and disadvantages are discussed. Energy based features are easy to compute but they are not sufficient in

low SNR conditions. Frequency based features perform well in the speech detection, but they are sometimes insufficient in low SNRs. When the noise has non-stationary characteristics, the performance of these features decreases quickly. Finally, long term features are quite efficient in various noise types and different SNR conditions (Ghosh, Tsiartas, & Narayanan, 2011). However, it can be hard to implement in real world case scenarios.

2.2 Decision for Voice Activity Detection

The first step for VAD system is the extraction of features from the signal. After doing that, the second part of VAD system makes a final decision when determining whether speech or non-speech labels. The decision system uses the feature information, and its task is to assign the data (feature information) to one of the two classes (Alpaydın, 2010). In other words, the decision system selects the best separation boundaries between speech and noise.

Thresholding is the simplest way when determining the decision region. When selecting a threshold in the feature space, a line or lines are drawn which separate the data into classes. The threshold values can be fixed that is generally decided in the training stage (Haigh& Mason, 1993) or it can be adaptive which is updated in each frame according to the newest speech or non-speech feature values (Ghosh, Tsiartas, & Narayanan, 2011), (Ramirez, Segura, Benitez, Torre, & Rubio, 2004), (Verteletskaya & Sakhnov, 2010).

Thresholding is an easy way to separate classes. However, in the case of low SNRs, it is insufficient to give an outstanding performance. Therefore, non-linear surfaces and complex boundaries are required. To find complex boundaries, machine learning and statistical model based approaches have been proposed.

Statistical model approaches enhance the performance of VAD system by using a statistical model. In this statistical model decision rule is derived from the likelihood ratio test (LRT) by estimating unknown parameters using the maximum likelihood (ML) criterion (Sohn, Kim, & Sung, 1999). It is generally assumed that speech and

noise spectra are defined as a Gaussian random process. Hence, their discrete-Fourier transform (DFT) coefficients can be represented as Gaussian random variables, and these variables are described in terms of probability density functions such as the Gamma and Laplacian distributions (Chang, Kim, & Mitra, 2006). To sum up, by using Gamma and Laplacian distributions, speech can be modeled well in the frequency domain.

Machine learning approaches have become popular due to the need for the robust classifier. Many classifiers have been designed such as Gaussian Mixture Model (GMM) (Ying, Yan, Dang, & Soong, 2011), Artificial Neural Network (ANN) (Pham, Tang, & Statdtschitzer, 2009), Support Vector Machine (SVM) (Baig, Masud, & Awais, 2006), and the nearest neighbor algorithms (Kos, Kačič, & Vlaj, 2013). Even though the machine learning approaches are more robust than thresholding, they are more complex and can work well in specific noise and SNR levels. It should be taken into consideration that types of noise and noise levels could change the performance of VAD system.

One of these approaches is neural networks that are inspired by the brain and the brain processes. While the brain uses the neurons to solve problems in real world, the neural networks take advantage of the artificial neurons in a specific application, such as pattern recognition or classification. Brain involves synaptic connections in neurons (Stergiou & Siganos, n.d). Similarly, learning ways include neuron connections via weighting coefficients in ANN. As shown in Figure 2.6 a typical neuron consists of dendrites, axon, and cell body. The neural networks are modeled by using the essential features of neurons. When modeling the neuron, there are also interconnections, inputs and outputs similar like components of neuron.

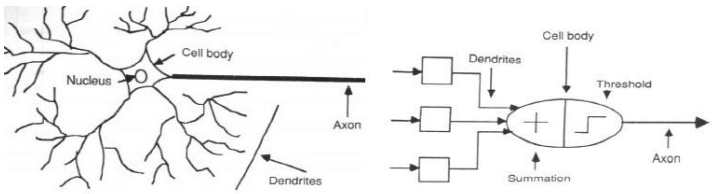

Figure 2.6 Components of neuron and the neuron model (Stergiou & Siganos, n.d)

The major thing is to adjust the appropriate weights of connections. As a result, neural networks find a best separation hyperplane in the feature space. Neural networks are widely used in speech recognition and speech detection areas but it can be complex to find the best separation plane. A multilayer perception neural network model which has one or more layers between the input (features) and output (speech/non-speech decision) is used for the decision stage in this thesis. Training algorithm is chosen Levenberg-Marquardt backpropagation algorithm for the perceptron. This algorithm updates weight values according to Levenberg-Marquardt method. It is used Hessian matrix in this algorithm. Adaption rule is also given in Equation (2.21).

$$H \approx \boldsymbol{J^T J} + \mu \tag{2.20}$$

where J is the Jacobian matrix (it contains first order derivatives of the network errors with respect to the weights), μ is always positive called combination coefficient and I is the identity matrix.

$$w_{k+1} = w_k - (\boldsymbol{J^T J} + \mu\, I)^{-1} J^T e \tag{2.21}$$

where e is the error vector and w represents the weight vector. The Jacobian matrix is computed because it is very complex to calculate the second-order derivatives of the total error function. Since it is the combination of gradient descent and the Gauss-Newton method, it solves the problems existing in both two algorithms (Hao, & Wilamowski, 2010). When it is compared with the first order algorithms, it is fast, stable training method, and this algorithm has powerful search ability.

2.3 ITU-T G.729 VAD System

It has been known that VAD system has been widely used in speech coding areas. The International Telecommunication Union (ITU) developed high-quality speech coding algorithm (G.729) which contains a VAD system in DTX mode. G.729 VAD

system has become standard for speech applications. This speech codec with VAD in DTX is shown Figure 2.7. The VAD system makes a decision for each 10ms and length of frame is associated with the sampling frequency in speech codec algorithm.

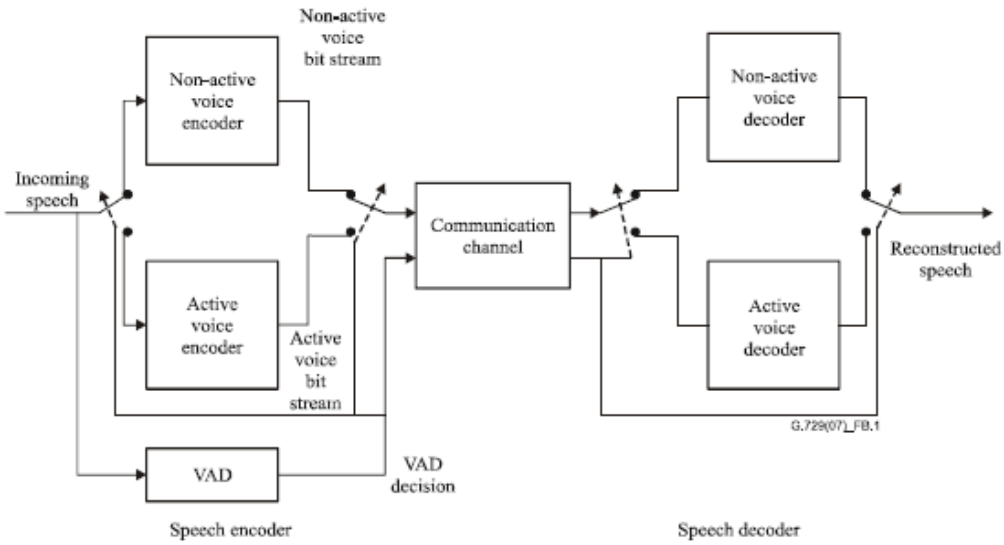

Figure 2.7 Speech coding with VAD in DTX (ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, 2012)

The output of VAD returns either 1 or 0. In other words, this represents the presence or absence of speech (ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, 2012).

VAD system computes various features such as the full band energy, the low band energy, the zero crossing rate and spectral measure. The flowchart of the VAD system is given as Figure 2.8. Firstly, feature extraction part is occurred. If the frame number is less than N_i (usually taken as 30) and the frame energy from the Linear Predictive Coding analysis is above 15db, the output of VAD is forced to 1. But in the case of the frame energy from the Linear Predictive Coding analysis is below 15db, the output of VAD is forced to 0. If the frame number is reached the number of initial frames N_i, an initialization stage for the characteristic energies of the background noise occurs (Benyassine, Shlomot, Huan-yu, & Massaloux, 1997). In other words, initial parameters are computed from these initial thirty frames. At the

next part, the different parameters are calculated for each frame by using four features. The decision is made at the multi-boundary decision stage by extracting an energy difference, a zero-crossing difference, a low band-energy difference and a spectral line difference. When the background noise is reached to energy thresholds, the average background parameters have to be updated. Therefore, an adaptive threshold is used in G.729 VAD system.

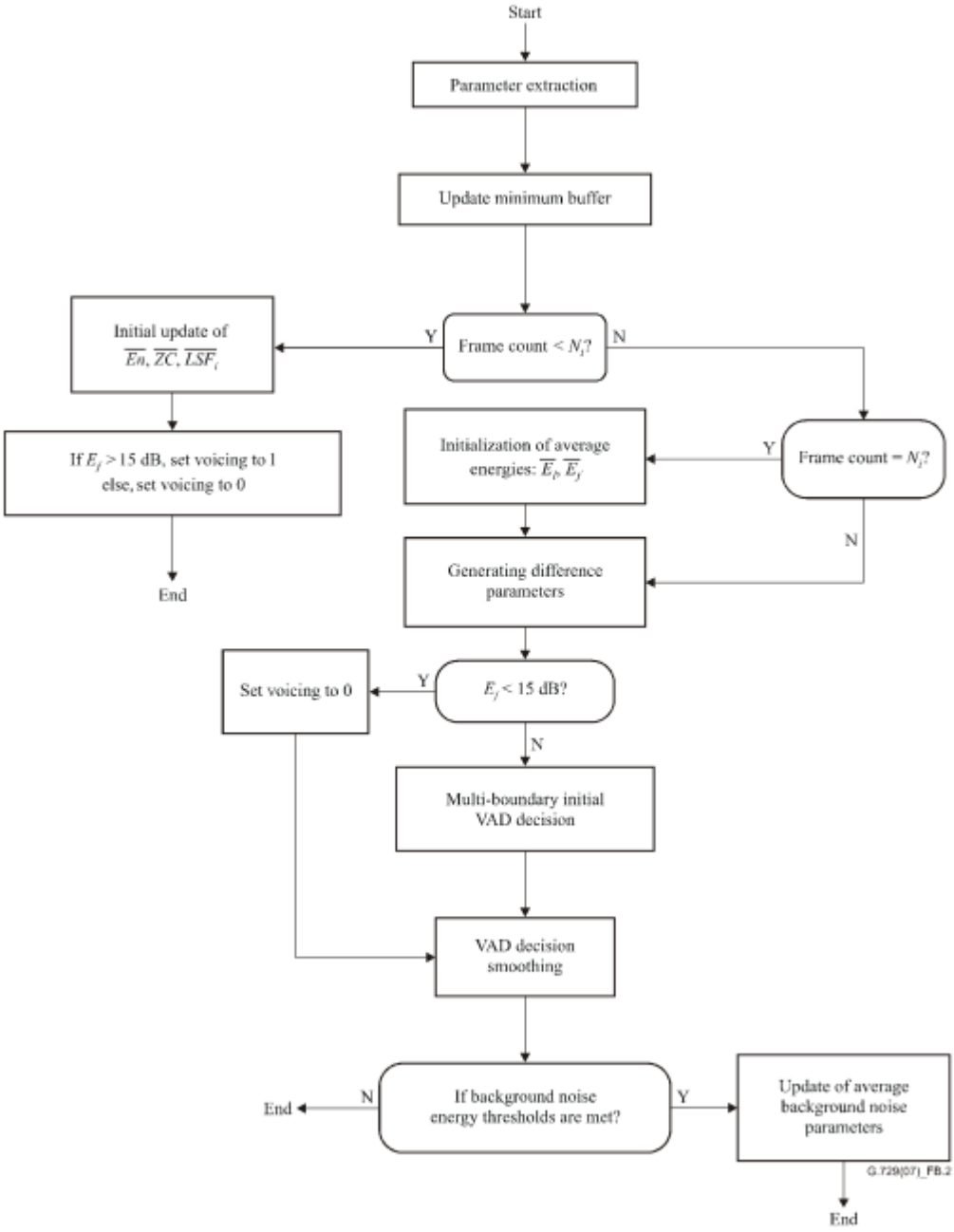

Figure 2.8 VAD flowchart (ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, 2012).

There are four features extracted when making a decision in the G.729 VAD system. Some basic information is given in details in this part. For example, the full band energy E_f computed from the logarithm of the normalized first autocorrelation coefficient $R(0)$ is given as follows:

$$E_f = 10 * log_{10}\left(\frac{1}{10}R(0)\right) \tag{2.22}$$

The second feature, the zero-crossing rate Z_x is a time-based feature that is used to determine the number of zero-crossings in a frame. It is mentioned in Equation (2.3) before. The low-band energy is another feature which is measured on 0 to F_l Hz band:

$$E_l = 10 * log_{10}\left(\frac{1}{N}h^T Rh\right) \tag{2.23}$$

where h is the impulsive response of an FIR filter with the cut-off frequency at F_l Hz, R is the Toeplitz autocorrelation matrix with the autocorrelation coefficients on the each diagonal.

The final feature for the G.729 VAD system is the line spectral frequencies (LSF). A set of $\{LSF_i\}_{i=1}^{p}$ where $p=10$, is derived from the set of linear prediction coefficients (ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, 2012).

CHAPTER THREE
SEPARATION OF MIXTURES

3.1 Independent Component Analysis

In signal analysis, 'the cocktail party problem' is based on that if there are *N* signal sources and at least *N* observation sensors like microphones in a room or environment, the mixed signals can be estimated from the combination of these source signals. It can be called as source separation problem. To solve this source separation problem, many methods have been proposed. Independent component analysis (ICA) is the most commonly used method for separation of mixture signals. The basic idea of the ICA is to find linear projection of the signals that maximize their mutual independence (Hong-yan&Guang-long, 2010).

However, there are some limitations for ICA. One of them, the source signals have to be stationary and independent of each other (Hong-yan, Qing-hua, Guang-long&Bao-jin, 2009). Second one, the number of microphones must be equal or less than the number of original signals (Hong-yan, Qing-hua, Guang-long&Bao-jin, 2009). Finally, it is allowed that only one source may be Gaussian because this algorithm measures non-Gaussianity of signals.

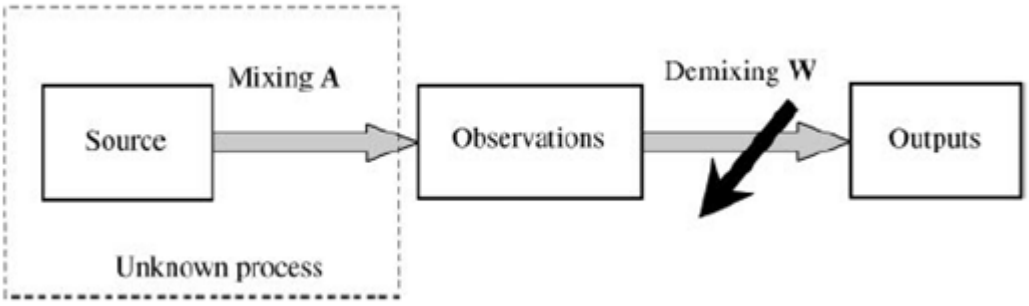

Figure 3.1 The model of ICA (Li, Gu, Ball, Leung & Phipps, 2001)

The model of ICA is given in Figure 3.1. Also, equations of ICA are given as follows:

$$x(t) = \boldsymbol{A}s(t) \tag{3.1}$$

where $\boldsymbol{A}$ is $m \times n$ the unknown mixing matrix, $s_i(t)$ is n source signals where represents $s_1(t), s_2(t), s_3(t), \dots, s_n(t)$, and $x_i(t)$ is m observation signals or mixing signals where represents $x_1(t), x_2(t), x_3(t), \dots, x_n(t)$.

$$\hat{s}(t) = \boldsymbol{W}x(t) \tag{3.2}$$

where $\boldsymbol{W}$ is the de-mixing matrix and $\hat{s}(t)$ is the estimation of n source signals. In this equation only the observation signals are known. The unknown mixing matrix and the source signals ($\boldsymbol{A}$ and $s(t)$) are assumed to be unknown. Both components ($\boldsymbol{A}$ and $s(t)$) must be estimated using the observation signals $x(t)$ (Hong-yan, Qing-hua, Guang-long&Bao-jin, 2009).

3.1.1 Fast-ICA Algorithm

For the solution of the ICA problem, gradient descent based methods were proposed in the literature (Wu & Chiu, 2001). The performance and the speed of these methods highly depend on the step size parameter. Hyvarinen and Oja were developed Fast-ICA method based on the principal of fixed point iteration scheme and independent from the step size parameter. It is fast when comparing to gradient descent methods and is easy to apply (Hyvärinen & Oja, 2000). It finds a non-Gaussianity measurement by using non-linear functions. This algorithm computes the source signals separately or simultaneously (Hyvärinen & Oja, 2000).

At the preprocessing stage, the observation signals are centered by making their mean zero and then whitening transform is applied. After the whitening, the observation signals become uncorrelated and have unit-variance property.

$$\hat{x}(t) = Ux(t) \tag{3.3}$$

where $\hat{x}(t)$ is the whitening signals $E[\hat{x}(t)\hat{x}(t)^T] = I$, U is the whitening matrix. The whitening signals $\hat{x}(t)$ signify that the observation signals $x(t)$ is transformed linearly. To compute this transformation, the eigen-value decomposition of the covariance matrix is used.

$$U = \Lambda^{-\frac{1}{2}} V^T \tag{3.4}$$

where $\Lambda = diag[\lambda(1), \lambda(2), \dots, \lambda(m)]$ is the diagonal matrix of the eigen-values of the data covariance matrix, V is the matrix of the corresponding eigenvectors as its columns. By using (3.1),

$$\hat{x}(t) = UAs(t) \tag{3.5}$$

After the preprocessing step, the matrix $B = UA$ is defined as orthogonal. The estimated source signals can be defined as follows.

$$\hat{s}(t) = W\hat{x}(t) \tag{3.6}$$

If $W = B^{-1}$, the estimated signals $\hat{s}(t)$are same as original source signals $s(t)$.

To better comprehend the Fast-ICA algorithm, its structure and basic terms such as measures of nongaussianity, negentropy must be understood clearly. There are several functions to calculate demixing matrix W by using nonlinearity of the probability density function of source signals (Hong-yan&Guang-long, 2010). It is known that ICA uses the non-Gaussianity of signals. It is assumed that Fast-ICA unit looks like an artificial neuron having a weight vector. The weight vector is updated in each iteration with the help of a learning rule (Hyvärinen&Oja, 2000). The aim of Fast-ICA unit is to find a direction that maximizes nongaussianity of $w^T x$. Non-gaussianity is measured by the approximation of negentropy $J(w^T x)$. The

negentropy, also negative entropy, is defined as the reverse of entropy. In other words, it is based on the information-theoretic quantity of (differential) entropy (Hyvärinen & Oja, 2000). By expanding negentropy term $J(w^T x)$, the approximation of negentropy uses the nonquadradic functions.

The steps of the Fast-ICA begin with choosing an initial weight matrix w. The weight matrix is updated as follows,

$$w(t+1) = E\{xg(\boldsymbol{w^T}x)\} - E\{g'(\boldsymbol{w^T}x)\}\boldsymbol{w} \tag{3.7}$$

where g is a nonquadratic function, the weight matrix w, $w(t+1)$ the new updated weight matrix respectively.

Then, the weight matrix is normalized. If the weight matrix does not converge, Equation (3.7) is computed again. In this algorithm, convergence means that the old and new weight matrixes are in the same direction.

3.1.2 Selection of Mixing Matrix (A)

Since the mixing matrix is unknown, it can be selected according to the some assumptions. In some algorithms, the mixing matrix elements are taken as uniformly distributed random numbers between 0 and 1. For example, if the number of sources is equal to 2, the mixing matrix can be chosen as Equation (3.8)

$$\mathrm{A} = \begin{bmatrix} 0.4456 & 0.7094 \\ 0.6463 & 0.7547 \end{bmatrix} \tag{3.8}$$

In Figure 3.2, the sources and the observations are shown due to matrix of *A* given in (3.8).

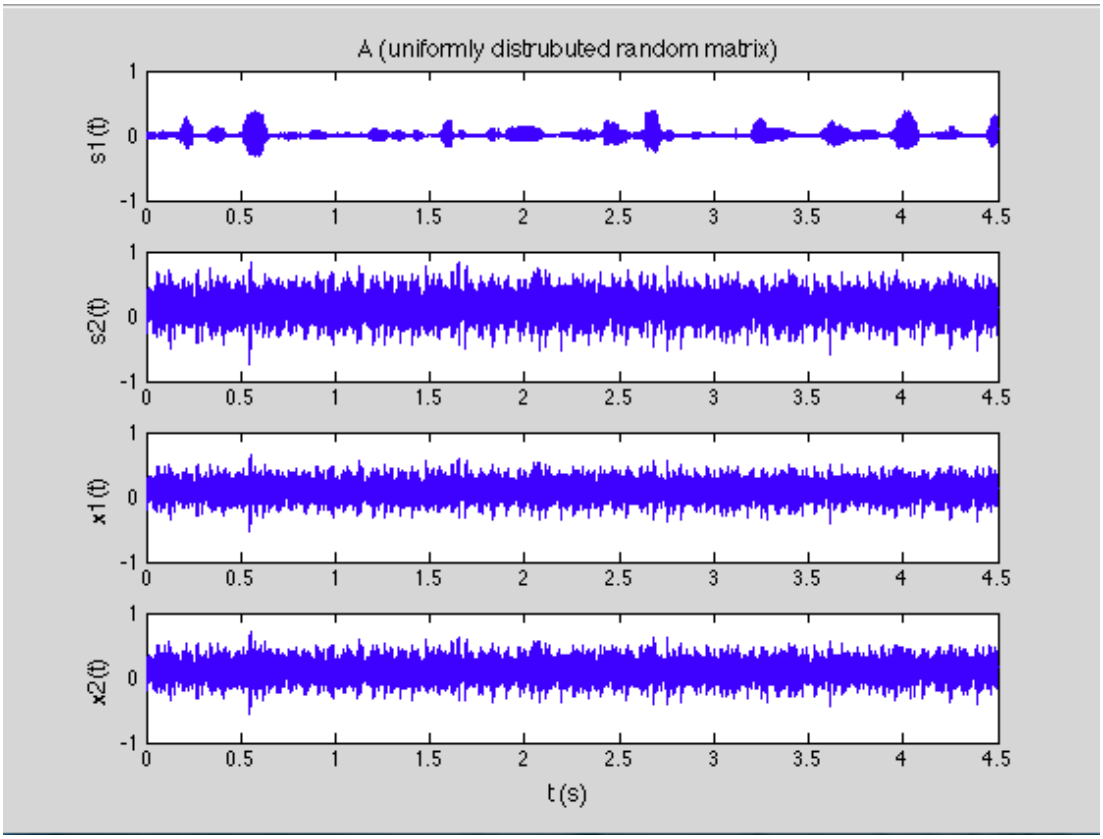

Figure 3.2 The sources and observations according to the uniform random matrix of A

Also, the mixing matrix of A can be selected as unity matrix. It means that when the number of sources is equal 2, one of observation signal $(x1(t))$ is similar to source one $(s1(t))$, the other observation $(x2(t))$ is similar to source one $(s2(t))$. An example mixing matrix for this case is given in Equation (3.9), the source and observation signals according to this matrix are shown in Figure 3.3.

$$A = \begin{bmatrix} 1.0080 & 0.0010 \\ 0.0090 & 1.0090 \end{bmatrix} \tag{3.9}$$

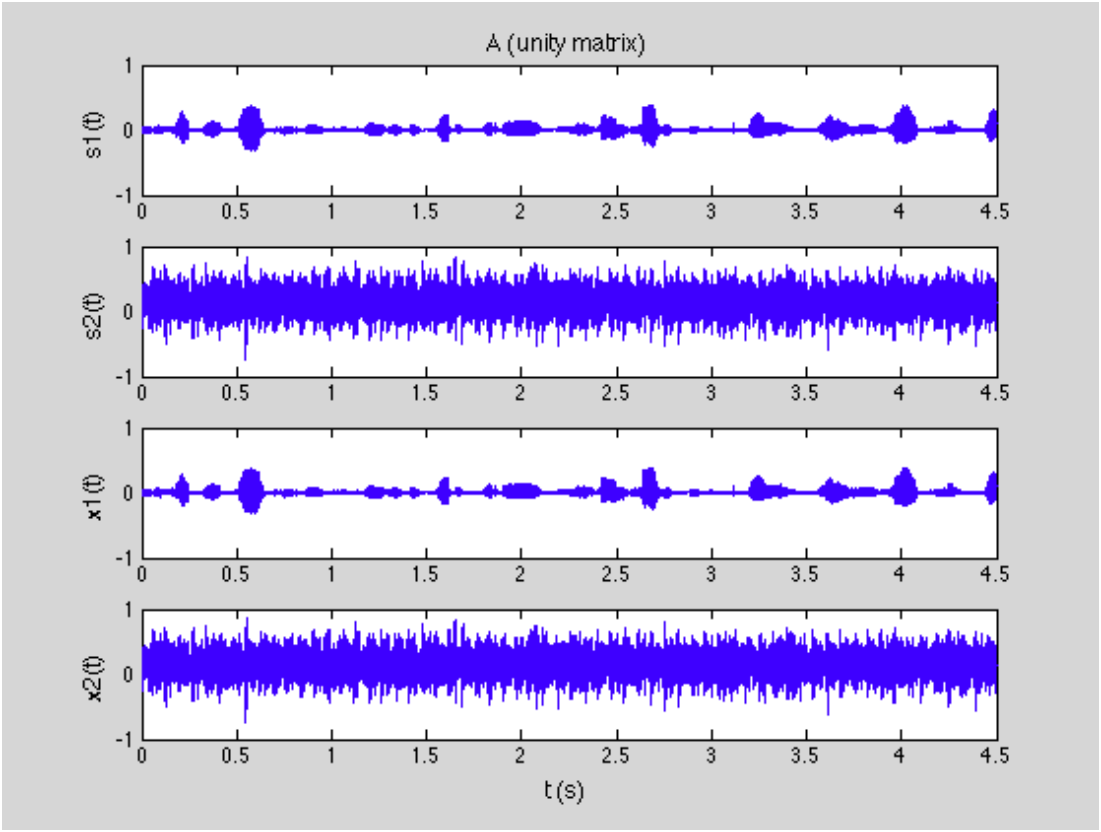

Figure 3.3The sources and observations according to the unity matrix of A

.

The third way for the specification of the mixing matrix is to make its diagonal element one, but non-diagonal elements can also be chosen as one in Equation (3.10). By selecting the mixing matrix as equation (3.10), observation signals are affected on same ratio by both 2 sources (it is assumed that the number of sources is equal 2.).

$$A = \begin{bmatrix} 1.0000 & 0.9500 \\ 0.9500 & 1.0000 \end{bmatrix} \quad (3.10)$$

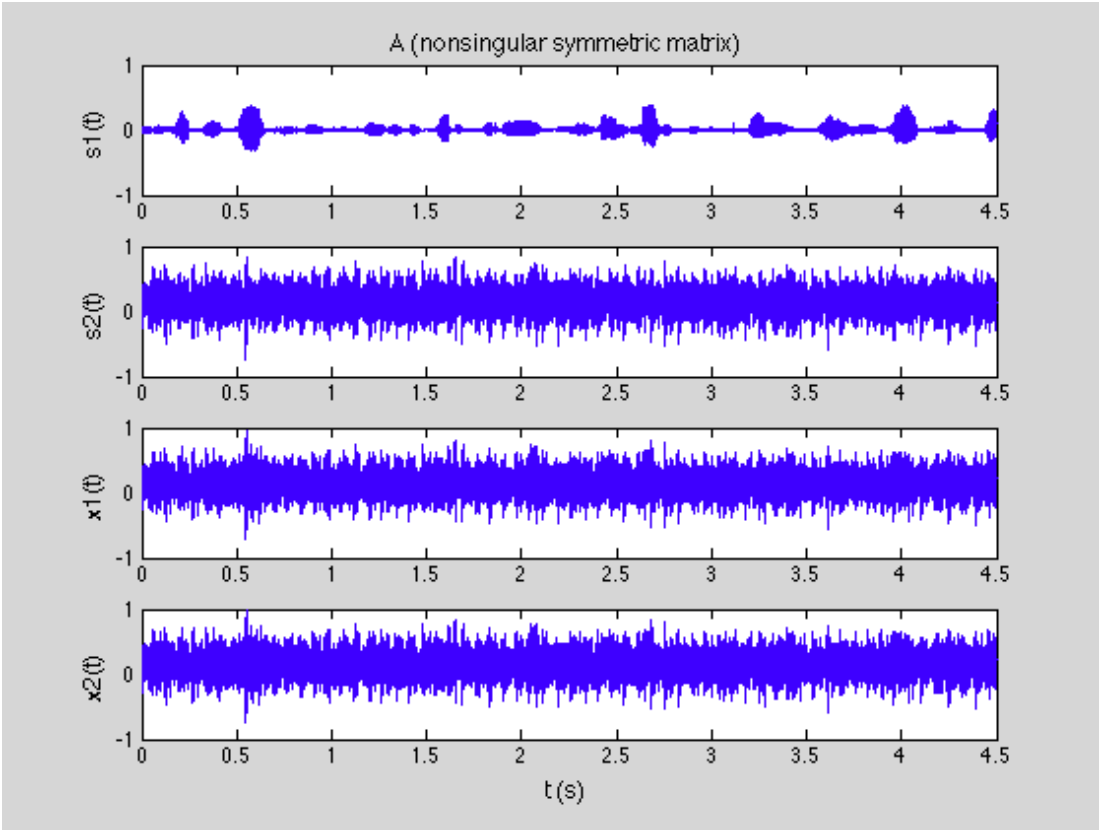

Figure 3.4The sources and observations according to the nonsingular symmetric matrix of A

However, sometimes the mixing matrix chosen with these methods does not sufficient to symbolize a real acoustic room environment. In order to model the real acoustic room environment, a high order FIR filter or a low order IIR filter can be used. It has shown that the observation signals taken by microphones can be represented by a convolution of source signal with a high order of FIR filter (Christensen, 1992). The room acoustic can also be modeled with a low order IIR filter. However, according to Mitianoudis & Davies, IIR filters are less stable and require minimum-phase mixing (Mitianoudis & Davies, 2003). In conclusion, the convolutive mixing matrix takes into consideration the time delays and room acoustics. The formula is given as:

$$x_i[n] = \sum_{j=1}^{N} \sum_{k=1}^{L} \alpha_{jk} s_j[n-k] \qquad i = 1,2,\dots,N \qquad (3.11)$$

where L is the maximum filter length, α_{jk} is the FIR filter coefficients and N is the number of sources. $x_i[n]$, the observation signals are calculated from the convolutions of the N sources with N filter of maximum length L.

The mixing matrix of A has elements α_{jk} where α_{11} and α_{12} are convolved with s_1, and α_{21} and α_{22} are convolved with s_2. The number of filter coefficients (the order of filter) α_{jk} is chosen as twenty. The sources and observations are given in Figure 3.5.

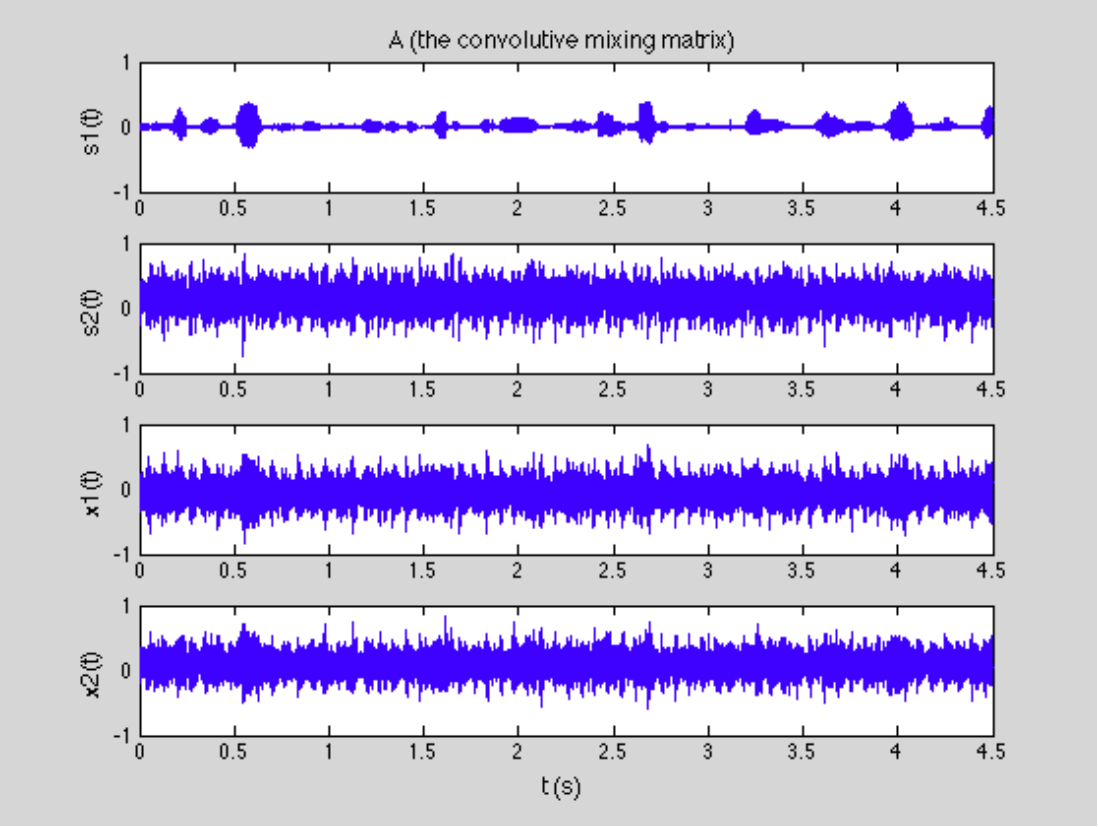

Figure 3.5 The sources and observations according to the convolutive mixing matrix of A

The different types of mixing matrix are investigated in this part. The nonsingular symmetric random matrix Equation (3.10) is used for the ICA block in this thesis.

CHAPTER FOUR
ACTIVE NOISE CONTROL SYSTEM

Active noise control is based on the theory that the unwanted sound can be eliminated by an anti-noise. For example, if this sound is a sinusoid signal, it can be cancelled by the signal which has same amplitude but 180 degree phase shift. While it is easy to eliminate the sinusoid signals or pure tone signals, it is difficult to cancel multiple tones and impulsive sounds. The summation of noise and anti-noise generates residual noise. In the ideal case, the residual noise must be equal to zero. In the real world applications, the aim of active noise controller is to produce an accurate anti-noise so that the residual noise can barely heard. Also, active noise control system must be fast and stable to give an accurate solution (Elliott & Nelson, 1993).

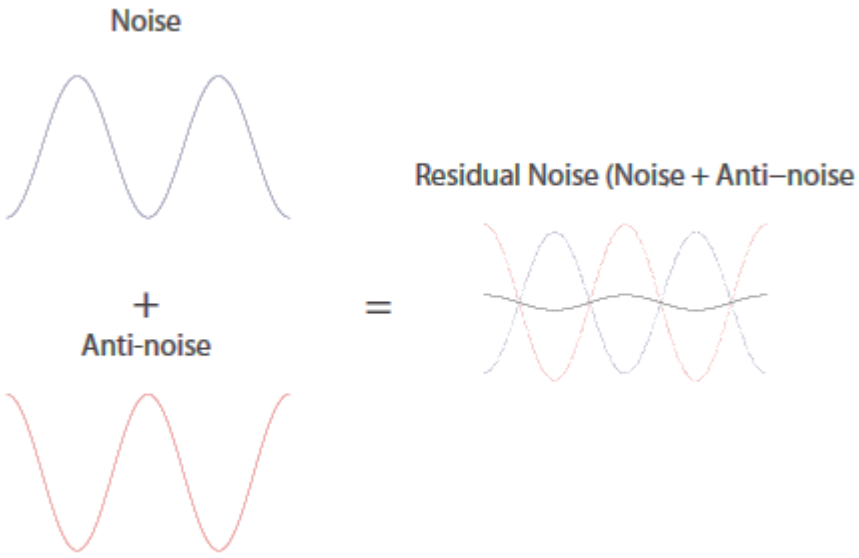

Figure 4.1 Principle of sound cancellation (Forsgren, 2011)

4.1 Algorithms of ANC

Active noise control system is generally designed with the help of adaptive filters. The very common algorithm is Least Mean Square Algorithm (LMS) which minimizes the mean-square value of the error signal.

4.1.1 Least Mean Square

The LMS algorithm is based on the steepest descent algorithm. The adaptation is done recursively using the gradient vector of the squared error in the steepest descent algorithm (Kataja, 2012). The purpose of the LMS algorithm is to find the coefficients of the filter iteratively. The best coefficients mean that the filter minimizes the mean square error.

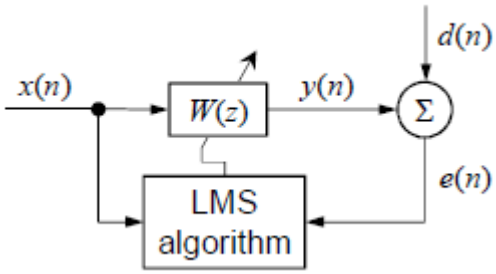

Figure 4.2 LMS algorithm block diagram (Kataja, 2012)

In the LMS algorithm, the error is the difference between the desired signal and the filter output as given in (4.1). For the adaptation, the input signal is needed. Moreover, the filter output equation is given as in Equation (4.2),

$$e(n) = d(n) - y(n) = d(n) - w^T(n)x(n) \tag{4.1}$$

$$y(n) = w(n)^T x(n) \tag{4.2}$$

where $x(n)$, is the input reference signal,$d(n)$ is the primary noise signal, $y(n)$ is the output of the filter, $w(n)$ represents the filter coefficients.

The weights are updated by using Equation (4.3).

$$w(n+1) = w(n) - (\frac{\mu}{2})\nabla_{\boldsymbol{w}} e^2(n) \tag{4.3}$$

Since the equation of weight adaptation is proportional to the gradient of mean square of the error, Equation (4.4) is substituted into Equation (4.3) and final update

rule is obtained as given Equation (4.5). The final updated filter coefficients become as in Equation (4.5).

$$\nabla_{\boldsymbol{w}} e^2(n) = -2e(n)x(n) \tag{4.4}$$

$$w(n+1) = w(n) + \mu e(n)x(n) \tag{4.5}$$

4.1.2 Filtered-x LMS Algorithm

Filtered-x LMS (FxLMS) algorithm is an alternative method of the LMS algorithm that involves a secondary path filter. The secondary path filter symbolizes the transfer paths between digital output and digital input of adaptive filter. It includes D/A Converter, amplifier, loudspeaker, the acoustic path between the loudspeaker and the error microphone, amplifier and A/D converter (Figure 4.3). The model of secondary path transfer function is given as Figure 4.4.

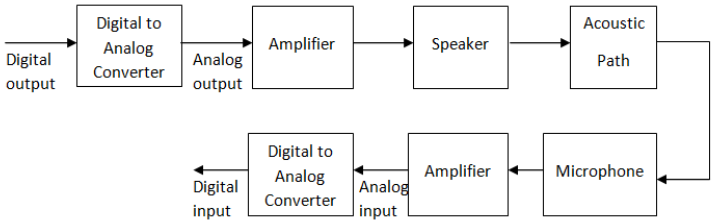

Figure 4.3 Transfer functions of the secondary path

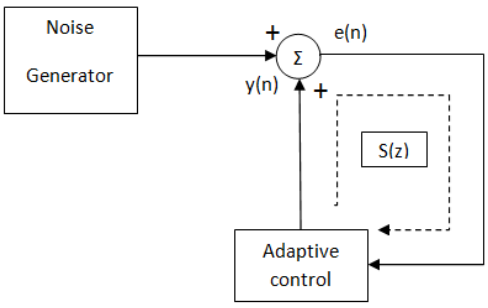

Figure 4.4 Model of transfer functions of the secondary path

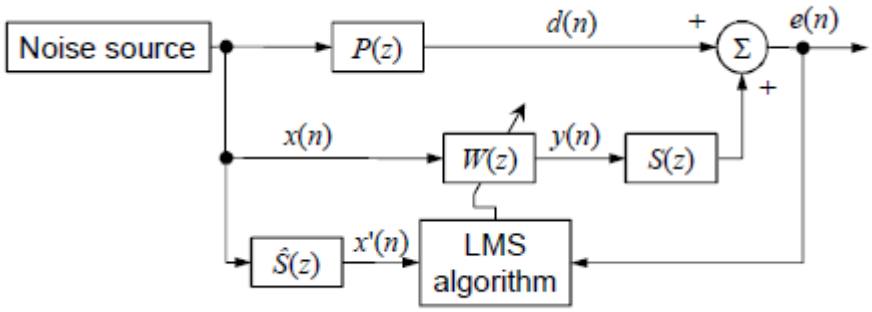

Figure 4.5 Fx-LMS algorithm block diagram (Kataja, 2012)

$$e(n) = d(n) + y'(n) = d(n) + s(n) * y(n) \tag{4.6}$$

$$y(n) = w(n)^T x(n) \tag{4.7}$$

where $x(n)$, is the input reference signal, $P(z)$ is the transfer function of the acoustic path, $d(n)$ is the primary noise signal, $y(n)$ is the output of the filter, $y'(n)$ is the filtered version of $y(n)$ with secondary path. $\hat{S}(z)$ is the estimated transfer function of the secondary path and $x'(n)$ is the filtered input signal with the secondary path.

$$e(n) = d(n) + s(n) * w(n)^T x(n) \tag{4.8}$$

According to the Figure 4.5, following equations can be written

$$x'(n) = \hat{s}(n)x(n) \tag{4.9}$$

$$e(n) = d(n) + w(n)^T x'(n) \tag{4.10}$$

By using Equation (4.3), the weight update rule can be given as follows

$$\nabla e^2(n) = 2e(n)x'(n) \tag{4.11}$$

$$w(n+1) = w(n) - \mu e(n)x'(n) \tag{4.12}$$

4.2 Methods of ANC

According to the presence of reference or error microphone, there are two types of ANC algorithms namely feedforward and feedback ANC.

4.2.1 Feedforward ANC

Feedforward ANC consists of two microphones (reference and error microphone) and one secondary source. The noise is taken by the reference microphone and fed into the adaptive filter that produces cancelling output. Error microphone measures the residual error. Feedforward ANC algorithms are generally used to cancel time-varying noises.

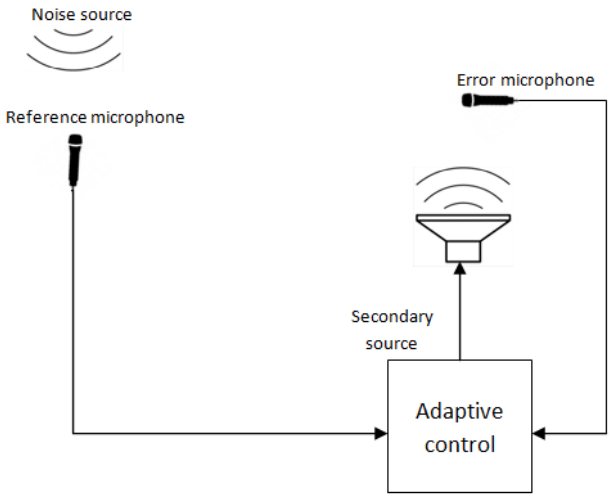

Figure 4.6 Feedforward ANC system

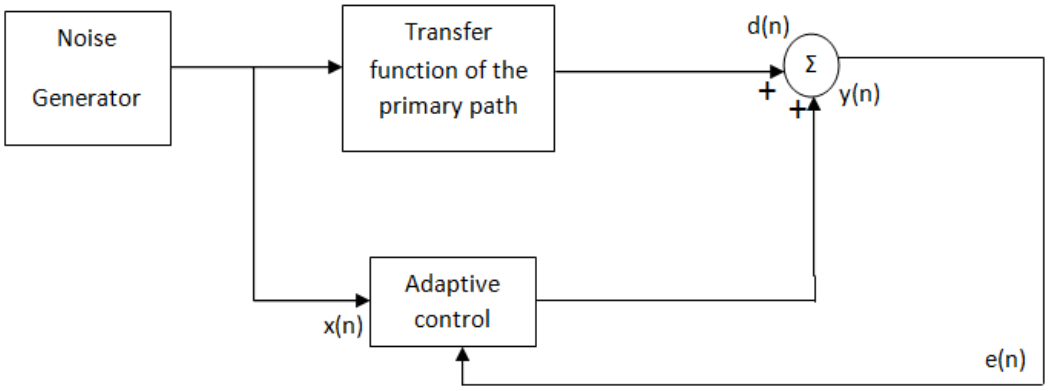

Figure 4.7 Feedforward ANC block diagram

4.2.2 Feedback ANC

There is no reference microphone in the feedback ANC algorithm in contrast to feedforward ANC. The only sensor is the error microphone so the ANC system produces the anti-noise by using error signal and filter output. In the feedback ANC, the secondary noise source produces the anti-noise and the error microphone measures the residual error (Kajikawa& Hirayama, 2010). The adaptive filter coefficients are updated according to this residual error. Since there is no reference microphone in this method, feedback ANC estimates the input signal. As a result, feedback ANC algorithms and applications are generally used for predictable noise and narrow-band noises.

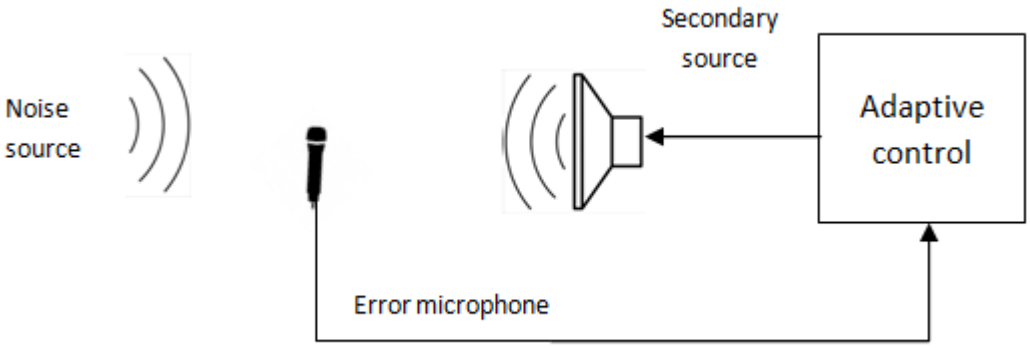

Figure 4.8 Feedback ANC system

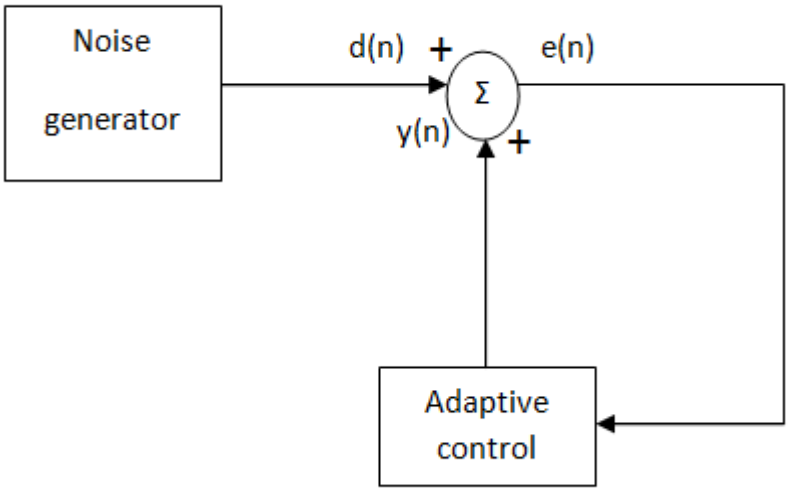

Figure 4.9 Feedback ANC block diagram

4.3 The Selective System Using New Active Noise Controller

Sohn & Lee proposed a noise cancellation system which can selectively cancel a particular noise signal from a mixture. Blind source separation (BSS) algorithm separates the signal sources and one of the separated signals is eliminated with the help of ANC system.

The position of microphones is crucial for the selective attention system. It is assumed that the second microphone and loudspeaker are closed to each other (Figure 4.10). Using this assumption, the observation signals can be expressed as follows:

$$x(t) = H(z)s(t) + [h'(z)\ 1]^T y_a(t) \tag{4.13}$$

where $s(t)$ represent the source signals, $y_a(t)$ is the output of the ANC system $H(z)$ is the mixing matrix and $h'(z)$ is the transfer function between the loudspeaker and the first microphone. According to Figure 4.10, the error function for the ANC system is the square of the difference between the desired signal $y_{b1}(t)$ and the received signal by the second microphone $x_2(t)$.

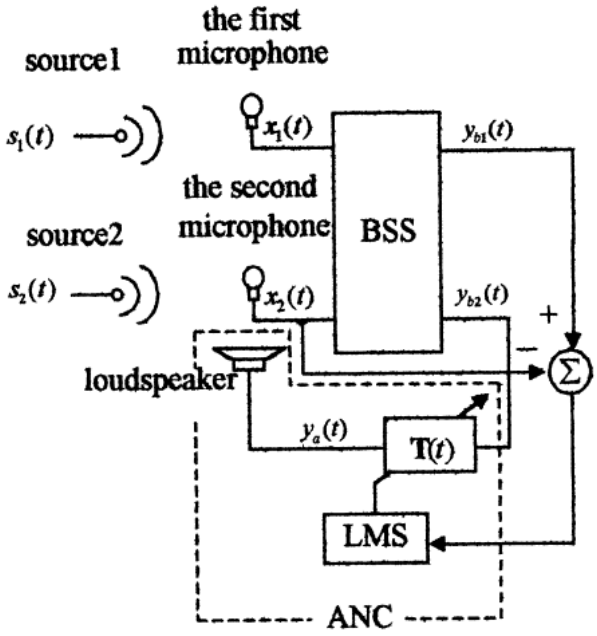

Figure 4.10 Proposed ANC system for selective attention (Sohn & Lee, 2000)

$$e(t) = \varepsilon^2(t) = \{y_{b1}(t) - x_2(t)\}^2 \tag{4.14}$$

The error becomes by using Equation (4.13),

$$e(t) = \{y_{b1}(t) - h_{21}(z)s_1(t) - h_{22}(z)s_2(t) - y_a(t)\}^2 \tag{4.15}$$

The filter coefficients update rule is given in Equations from (4.16) to (4.17). $Y_{b2}(t)$ is the delayed vector of the unwanted signal of the BSS output $y_{b2}(t)$, $y_a(t)$ is the output of the ANC system, and $T(t)$ represents the weight vector of the adaptive filter for the ANC system.

$$T(t+1) = T(t) + 2\mu \frac{\partial e(t)}{dt} \tag{4.16}$$

$$T(t+1) = T(t) + 2\mu e(t) \frac{\partial y_a(t)}{dt} \tag{4.17}$$

$$y_a(t) = Y_{b2}(t) * T(t) \tag{4.18}$$

By taking the gradient of Equation (4.18), the final representation of filter coefficients update rule is given in Equation (4.19).

$$T(t+1) = T(t) + 2\mu e(t) Y_{\boldsymbol{b2}}(t) \tag{4.19}$$

In order to minimize the error, the BSS output $y_{b1}(t)$ is equal to the desired source signal $s_1(t)$ and the output of the adaptive filter $y_a(t)$ converges to $(1 - h_{21}(z))s_1(t) - h_{22}(z)s_2(t)$ (Equation (4.20), (4.21) and (4.22)). By using Equation (4.21), the BSS observation signals are given as follows Equations from (4.23) to (4.25).

$$y_a(t) = y_{b1}(t) - h_{21}(z)s_1(t) - h_{22}(z)s_2(t) \tag{4.20}$$

$$y_a(t) = s_1(t) - h_{21}(z)s_1(t) - h_{22}(z)s_2(t) \tag{4.21}$$

$$y_a(t) = s_1(t)(1 - h_{21}(z)) - h_{22}(z)s_2(t) \tag{4.22}$$

$$x(t) = H(z)s(t) + [h'(z)\ 1]^T((y_a(t))) \tag{4.23}$$

$$= \begin{bmatrix} h_{11}(z)s_1(t) + h_{12}(z)s_2(t) + h'(z)\{(1 - h_{21}(z))s_1(t) - h_{22}(z)s_2(t)\} \\ h_{21}(z)s_1(t) + h_{22}(z)s_2(t) + s_1(t) - h_{21}(z)s_1(t) - h_{22}(z)s_2(t) \end{bmatrix} \tag{4.24}$$

$$= \begin{bmatrix} h_{11}(z)s_1(t) + h_{12}(z)s_2(t) + h'(z)\{(1 - h_{21}(z))s_1(t) - h_{22}(z)s_2(t)\} \\ s_1(t) \end{bmatrix} \tag{4.25}$$

Sohn & Lee assumed that two signals (speech and noise) always exist in the environment. The selective attention system works under this assumption. At the output of the system, there are obviously two signals but when the BSS or the other separation algorithm is used, the order of speech signal is not known. The system also makes another assumption that the order of speech signal always takes first channel; this is not common approach in separation algorithms when they are used in real time applications.

CHAPTER FIVE
RESULTS

This chapter contains the result of the proposed system which is the combination of VAD, ICA and ANC blocks. The system is tested with the different noise types with different SNR ratio. The performance of each block is presented separately.

5.1 Data Preparation

The noise database consists of conveyor band sounds recorded from internet websites of which sampling frequency is 44.1 kHz and resolution is 16 bit per sample (Conveyor band sounds (n.d.), Conveyor belt sounds (n.d.), Conveyor Belt Sound Effects (n.d.), Conveyor Belt2 sound effect (n.d.)). The air hammer, ambiance and traffic sounds which have the same sampling frequency and resolution are also downloaded from the websites (Air hammer sounds (n.d.), Ambiance sounds (n.d.), Traffic sounds (n.d.)).

The PTDB-TUG speech database provided by Graz University of Technology is used (Pirker, Wohlmayr, Petrik, & Pernkopf, 2011). The database has phonetically rich sentences, which are taken TIMIT corpus with different male-female speakers (Lamel, Kassel & Seneff, 1986). Since the sampling frequency is 48 kHz and resolution is 16 bit per sample, the samples taken from this database are resampled.

In order to show the performance under various noise levels, Signal-to-Noise Ratio (SNR) is used which is a common measurement in speech and signal analysis. It means that what is the ratio of the noise signal in the speech signal or any desired signal. It is computed from power of speech signal respect to power of noise signal. The ratio is measured in decibels (Equation 5.1). It is clearly seen that a clean speech signal has high SNR. Besides, if noise is dominant in the environment, then it means that SNR is low.

$$SNR_{dB} = 10\, log_{10}(\frac{P_{Signal}}{P_{noise}}) \quad (5.1)$$

where P_{Signal} is energy of speech signal and P_{noise} energy of noise signal.

The dataset used in this thesis contains 11600 sounds of which 12 seconds. While the first and the last four seconds contain both the speech and noise, the remaining four seconds contains only the noise. There are 870 conveyor band, 580 traffic, 522 ambiance and 348 air hammer sounds which of length is twelve seconds. Different ratios of the speech and noise signals are mixed to obtain 5 different Signal-to-Noise Ratio (SNR) levels which are 10db, 5db, 0db,-5db and -10db.

5.2 VAD Performance Results

The performance of G.729 VAD block was measured in this section. The results of VAD block in various noise types and SNR conditions are plotted. In order to provide various noise types and SNR conditions, data preparation and noise addition which are crucial for the VAD block are mentioned.

5.2.1 G.729 VAD with Different Sounds in a Constant SNR Ratio

G.729 system makes a decision in each frame which has length of 10ms. To see the performances of VAD block in various sound types (conveyor band, air hammer, ambiance and traffic noise), at fixed SNR level which is chosen -5db where the noise is dominant. In each 10ms (4410 samples), VAD returns one or zero according to the speech or noise decision. The decision results of the G.729 VAD block are shown from Figure 5.1 to 5.4.

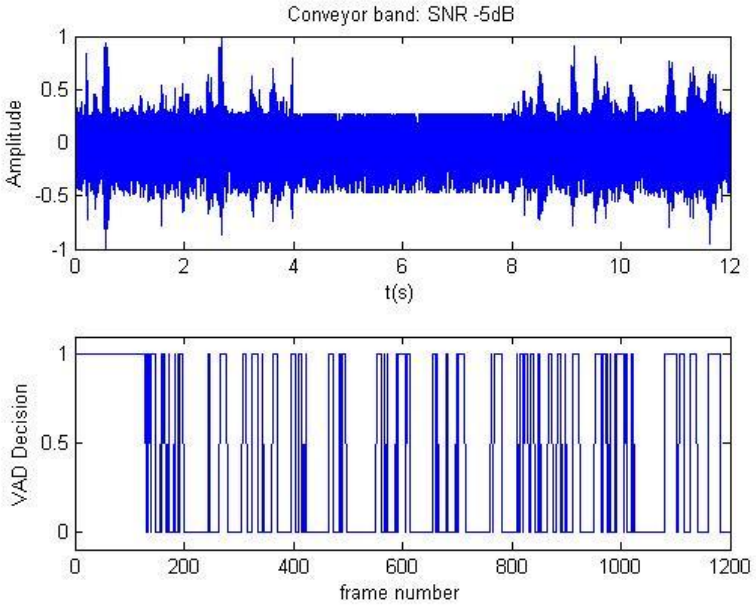

Figure 5.1 G.729 VAD with conveyor band examples in -5db SNR

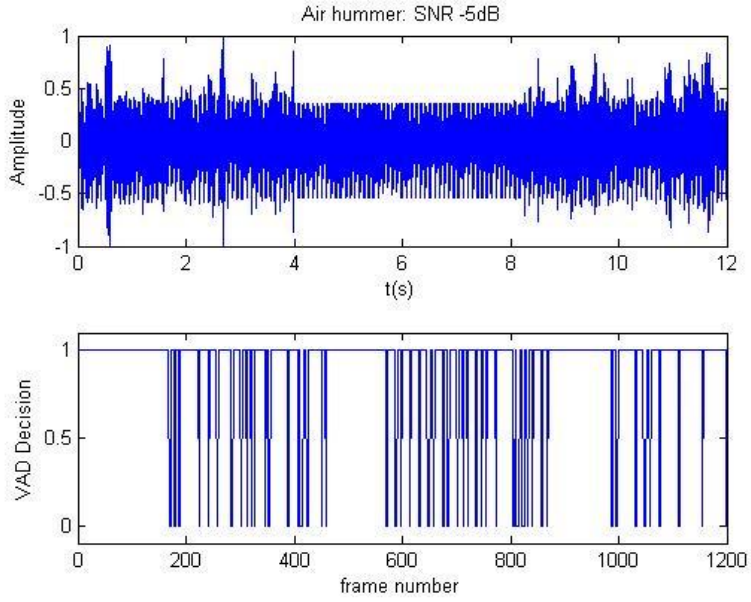

Figure 5.2 G.729 VAD with air hammer noise examples in -5db SNR

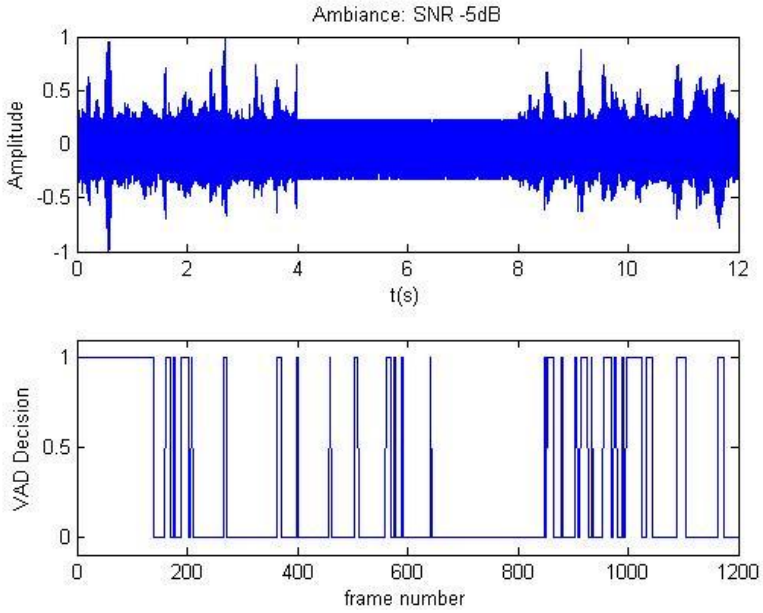

Figure 5.3 G.729 VAD with ambiance noise examples in -5db SNR

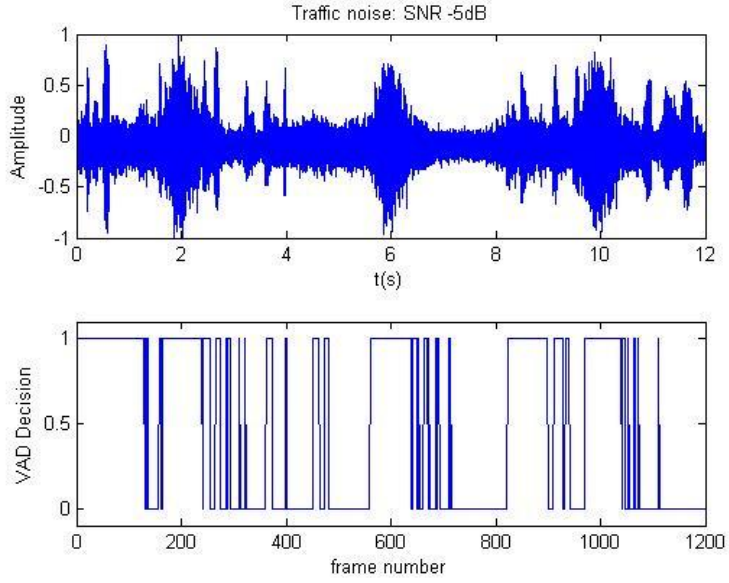

Figure 5.4 G.729 VAD with traffic noise examples in -5db SNR

5.2.2 G.729 VAD with Conveyor Band Sounds in Different SNRs

To measure the performance of VAD block in a specific noise type, the conveyor band sounds are used. The wav sounds are generated as the section 5.1 at five different SNR levels such as 10db, 5db, 0db, -5db and -10db. The decision results of

five different SNR levels with the conveyor band noises are shown from Figure 5.6 to 5.10. Also, clean speech is given to the VAD block to see its behavior in Figure 5.5.

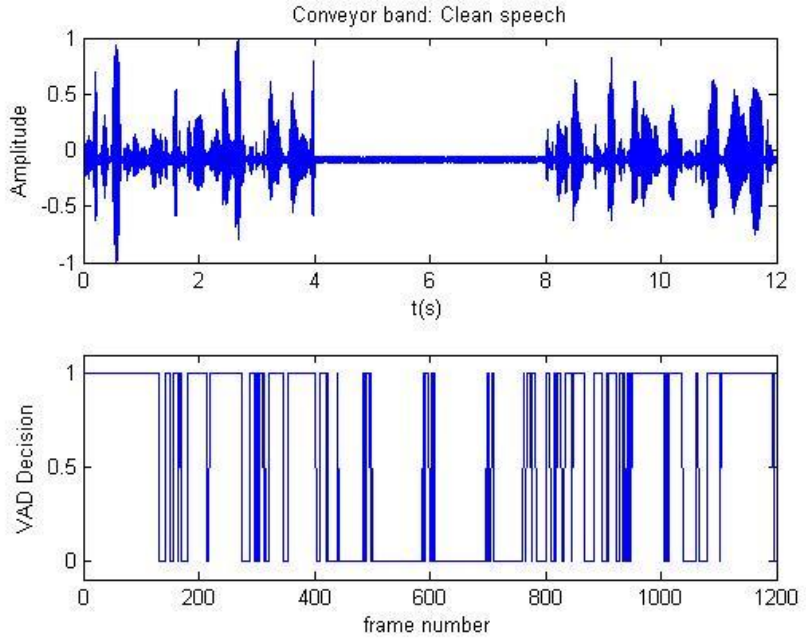

Figure 5.5 G.729 VAD with conveyor band sound examples in clean speech

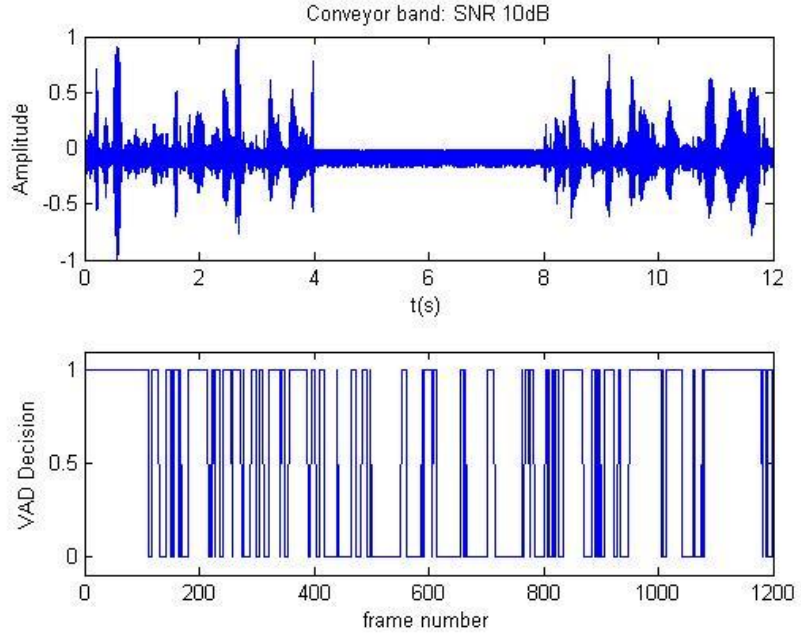

Figure 5.6 G.729 VAD with conveyor band sound examples in 10db SNR

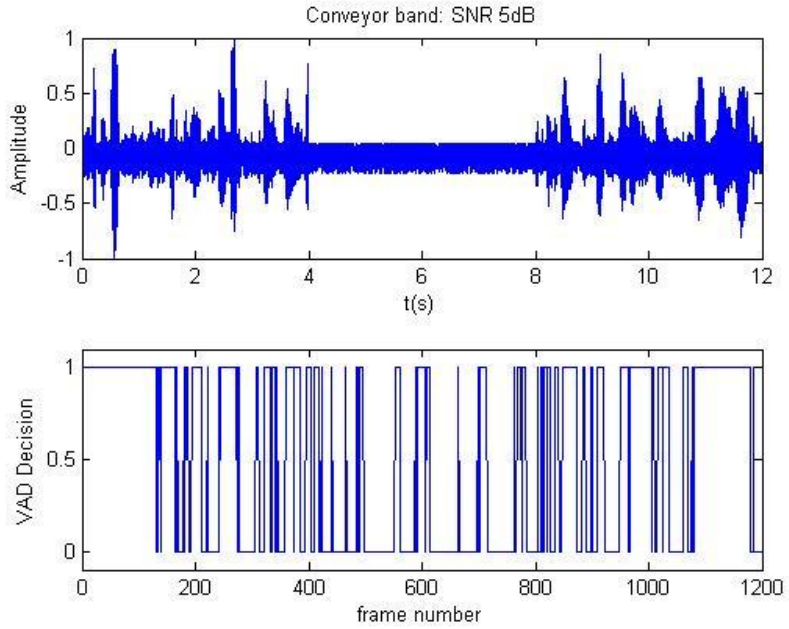

Figure 5.7 G.729 VAD with conveyor band sound examples in 5db SNR

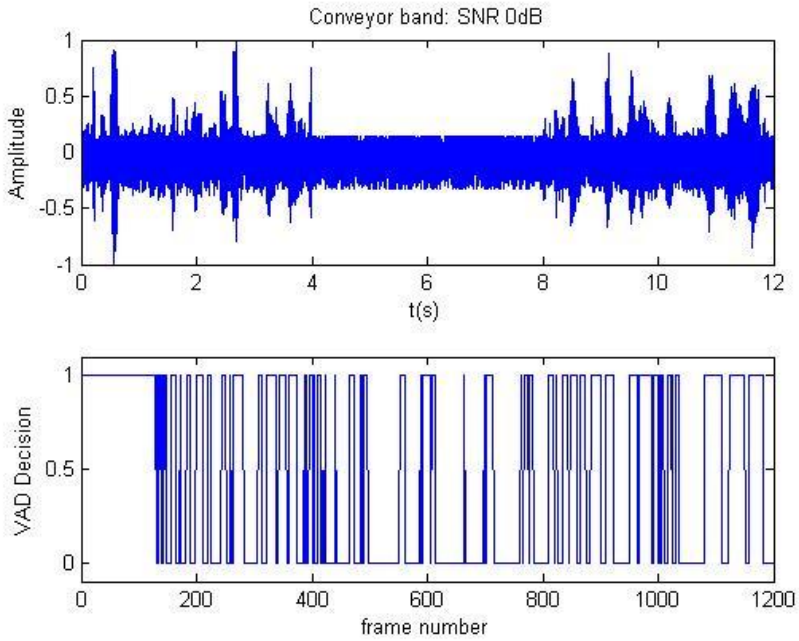

Figure 5.8 G.729 VAD with conveyor band sound examples in 0db SNR

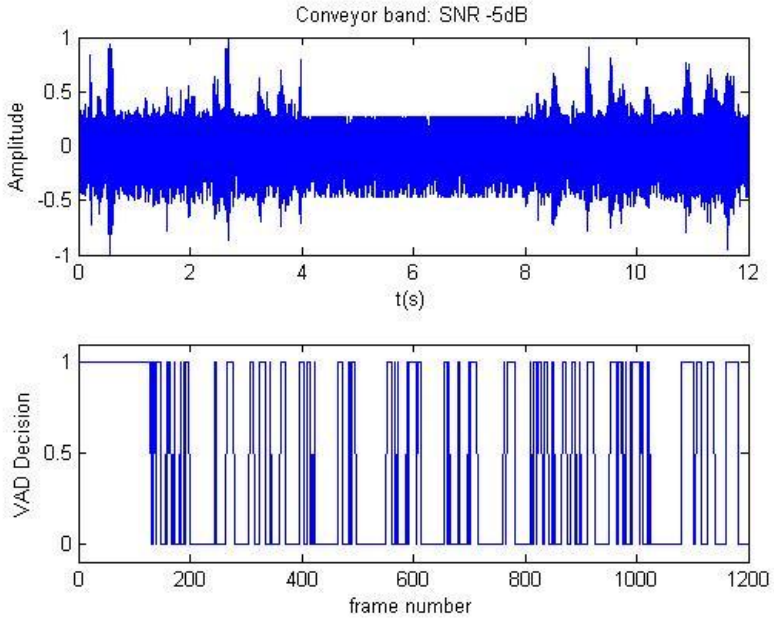

Figure 5.9 G.729 VAD with conveyor band sound examples in -5db SNR

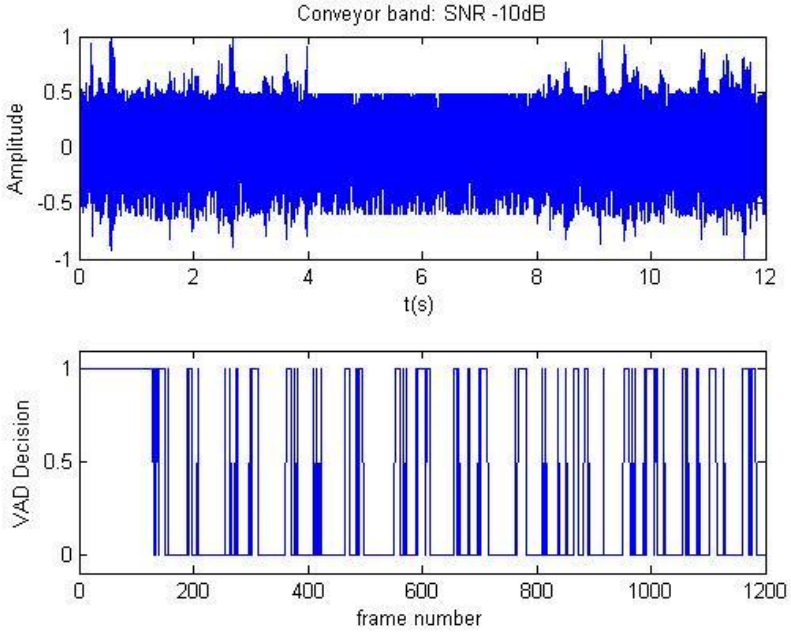

Figure 5.10 G.729 VAD with conveyor band sound examples in -10db SNR

5.2.3 Comparison of Features with Modified Mel-spectrum Feature

Speech hit rate and non-speech hit rate are the measures that determine the performance of VAD classifier.

$$HR0 = \frac{N_{0,0}}{N_0^{ref}} \tag{5.2}$$

$$HR1 = \frac{N_{1,1}}{N_1^{ref}} \tag{5.3}$$

where N_0^{ref} are the number of non-speech frames and N_1^{ref} speech frames in the whole database and $N_{0,0}$ and $N_{1,1}$ are the number of correctly classified non-speech and speech frames (Ramirez, Segura, Benitez, Torre & Rubio, 2004).

For better understanding, performance of the modified Mel-spectrum feature is compared with different features proposed in the literature. Note that the frame length is chosen 300 ms for this feature. Since the other features are generally short-time features, their mean or variance values are taken in order to have same frame length. If the frame length is chosen short-time windows (10-25ms), ICA cannot estimate the source signals. So, this is the other reason to choose 300ms for the frame length. To conclude, one feature symbolizes 300ms in this section. To see VAD block performance in a specific noise type, the conveyor band sounds are used. The wav sounds are generated as the section 5.1 at five different SNR levels such as 10db, 5db, 0db, -5db and -10db. Overall performances are tested with 870 conveyor band sounds.

In order to make a classification, neural networks are used. In this part, a multi layer perceptron is used with one input neuron which represents feature, one hidden layer with 10 hidden neurons and 1 output neuron that determines the input feature vector is speech or non-speech. Modified Mel-spectrum feature and other features are tested separately by using dataset. Therefore, one input neuron is used. Matlab and its function ‘trainlm’ which symbolizes Levenberg-Marquardt backpropagation

algorithm are used for training algorithm. Although it requires more memory than other training algorithms, it is highly recommended as a first-choice supervised algorithm due to the fastest algorithm in toolbox (Ho, 1998). This algorithm updates weight values according to Levenberg-Marquardt method. The algorithm is mentioned in Section (2.2) before. The update rule for weight values is computed by using Equation (2.21). Transfer function for the network is chosen default option ('tansig function'). Also, the training epoch parameter is 1000. The database is divided into three parts as it is well known train set, test set and validation set with the ratio of 0.33, 0.33 and 0.33. The average speech, non-speech and total hit rates of five different SNR levels with the conveyor band noises are shown Table 5.1, Table 5.2 and Table 5.3.

Table 5.1 Average speech hit rates of all features with conveyor band sounds

HR1 (%)	10db	5db	0db	-5db	-10db
ZCR(var)	79.88	88.37	100	100	100
SF(var)	96.01	90.86	82.82	92.78	99.80
SR(var)	93.30	91.91	83.68	88.42	99.99
SC(var)	95.99	93.60	91.74	86.60	91.53
SE(mean)	82.40	88.29	90.01	90.32	92.63
V2W(var)	77.95	86.51	92.84	100	100
Modified Mel-spectrum feature	97.74	96.25	91.50	87.90	83.84

Table 5.2 Average non-speech hit rates of all features with conveyor band sounds

HR0 (%)	10db	5db	0db	-5db	-10db
ZCR(var)	63.58	74.50	0	0	0
SF(var)	88.91	82.62	73.55	16.85	0
SR(var)	98.73	94.38	89.29	38.29	0
SC(var)	97.47	93.85	74.44	67.72	30.95
SE(mean)	82.25	78.30	68.49	48.47	30.24
V2W(var)	80.60	35.35	10.44	0	0
Modified Mel-spectrum feature	99.39	99.36	96.40	81.55	52.52

Table 5.3 Total hit rates of all features with conveyor band sounds

THR (%)	10db	5db	0db	-5db	-10db
ZCR(var)	74.94	69.59	70.01	69.81	70.00
SF(var)	93.88	88.37	80.04	70.02	70.00
SR(var)	94.98	92.64	85.36	73.44	69.74
SC(var)	96.43	93.67	86.55	80.97	69.25
SE(mean)	82.35	85.31	83.59	77.90	73.92
V2W(var)	78.74	71.26	68.00	70.27	70.13
Modified Mel-spectrum feature	98.23	97.19	92.97	86.00	74.44

According to Table 5.2, the average HR0 rates of the other features are too low. That means the noise signals (in some cases all of them) are wrongly labeled as speech. Therefore, it seems that in some cases the performances of the other features are better than the proposed one at low SNR levels in Table 5.1. But total hit rates are taken into consideration. As it is seen in Table 5.3, the overall performance of the proposed feature is better than the other features which are frequently used in speech detection.

5.2.4 Performance of G.729 VAD

As it mentioned before, standard G.729 VAD system makes a decision in each frame (10ms). Since the sampling frequency is equal to 44.1 kHz, the frame contains 441 samples. It is sufficient to run with the ICA and ANC. However, the output of ICA is not known whether speech or not. To overcome this problem, the frame length is chosen 300ms. Standard G.729 VAD produces 30 decisions. If the 70 % percentage of these decisions returns zero, the signal contains only noise signal. The G.729 VAD algorithm is evaluated from five different SNR levels such as 10db, 5db, 0db, -5db and -10db. G.729 VAD algorithm is applied to 11600 sounds and average speech/non-speech hit rates and the total hit rates are given Table 5.4, 5.5 and 5.6. The speech and non-speech decision of G.729 VAD are plotted in Figure 5.11 and 5.12.

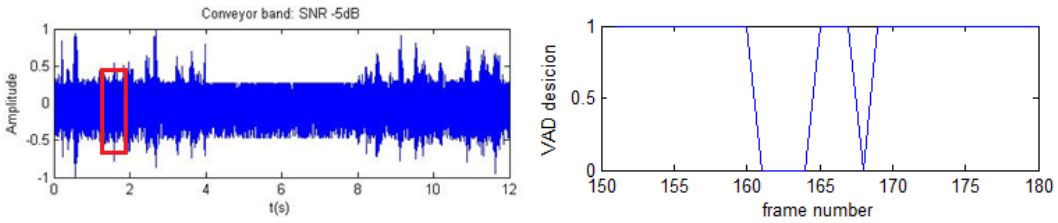

Figure 5.11 Speech decision of G.729 VAD with in -5db SNR

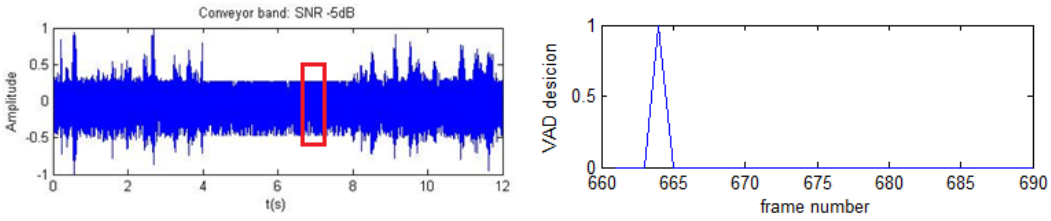

Figure 5.12 Non-speech decision of G.729 VAD with -5db SNR

Table 5.4 Average speech hit rates of G.729 VAD with different sounds

HR1 (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	57.49	84.48	94.76	71.24
5db	51.50	84.75	92.63	62.57
0db	37.51	80.63	89.18	55.06
-5db	26.14	81.28	87.70	48.63
-10db	19.64	80.04	82.56	41.80

Table 5.5 Average non-speech hit rates of G.729 VAD with different sounds

HR0 (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	100	48.92	42.86	83.29
5db	99.97	42.52	43.34	82.86
0db	99.91	41.95	47.55	81.35
-5db	100	34.05	43.53	79.55
-10db	100	31.75	41.52	78.86

Table 5.6 Total hit rate of G.729 VAD with different sounds

THR (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	70.77	73.81	79.19	74.72
5db	63.92	72.09	77.84	68.85
0db	56.25	69.03	75.83	62.95
-5db	48.30	67.11	74.02	57.91
-10db	43.75	65.56	71.46	53.56

The effect of SNR levels is analyzed. It is seen that the VAD system is insufficient when the SNR level is -5db or other low SNR levels. Due to the effects of the noise, the performance of the VAD decreases significantly. Also, the average performance speech hit-rates and non-speech hit rates are calculated for the conveyor band sounds and other sounds. The air hammer sounds reaches up to 84.48% speech hit-rates at the 10db SNR level, the speech hit-rate at the -10db SNR level falls to 80.04%. The performance of ambiance and air hammer is slightly higher than the

performance of conveyor band and traffic noise for the VAD block. When non-speech hit rates are investigated, air hammer has the lowest percentage.

5.2.5 Performance of the Modified Mel-spectrum Feature for Different Noise Types

Modified Mel-spectrum feature is proposed for VAD system and neural networks have been used for making a speech or non-speech decision. As stated before, neural networks are widely used in speech recognition and speech detection areas to find the best separation plane. A multilayer perception is a neural network model which has one or more layers between the input (features) and output (speech/non-speech decision). The parameters for training and testing the features are same as the previous section (section 5.2.3).

The proposed feature is evaluated from five different SNR levels such as 10db, 5db, 0db, -5db and -10db. The proposed VAD algorithm is also applied to 11600 sounds. Because test set has 33% rate, 3866 of them are test data and test set contains 154640 samples in order to perform the proposed VAD. Average speech/non-speech hit rates and the total hit rates are given Table 5.7, 5.8 and 5.9.

Table 5.7 Average speech hit rates of the proposed feature with different sounds

HR1 (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	97.74	95.19	98.15	96.28
5db	96.25	93.10	97.16	95.36
0db	91.50	88.30	95.90	94.42
-5db	87.90	88.94	91.56	91.53
-10db	83.84	99.2	90.96	89.91

Table 5.8 Average non-speech hit rates of the proposed feature with different sounds

HR0 (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	99.39	96.98	99.13	98.49
5db	99.36	90.20	98.99	93.55
0db	96.40	67.45	98.10	81.27
-5db	81.55	43.89	93.48	71.30
-10db	52.52	5.60	77.44	58.67

Table 5.9 Total hit rates of the proposed feature with different sounds

THR (%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	98.23	95.73	98.44	96.94
5db	97.19	91.12	97.71	94.81
0db	92.97	82.04	96.56	90.47
-5db	86.00	75.43	92.14	85.46
-10db	74.44	70.58	86.91	80.54

HR1 ratio of G.729 is lower than the proposed one. This error is critic because in noise control system, the signal which is labeled as noise is suppressed directly with the ANC block. The noise signal which is wrongly labeled as speech can be detected by the ICA block and this error can be reduced by the system. Overall performance of the proposed feature is better than the G.729 algorithm in all SNR levels with different types of noise.

5.3 ICA Results

The selection of mixture matrix is very important. The nonsingular symmetric random matrix Equation (3.10) is used for the ICA block. ICA analysis is applied to various types of noise (conveyor band, ambiance, air hammer and traffic sounds) at different SNR levels by using Fast-ICA toolbox in Matlab (Hyvärinen & Oja, 2000).

5.3.1 Performance of ICA

Several different performance algorithms are obtained in the literature. To perform this system, the global mixing matrix is preferred.

$$G = WA \tag{5.4}$$

where A is 2×2 the unknown mixing matrix, W is 2×2 the de-mixing matrix. As stated before, only the observation signals are known. After running ICA algorithm, there is W matrix which separates the mixing signals.

$$x(t) = As(t) \tag{5.5}$$

$$\hat{s}(t) = Wx(t) \tag{5.6}$$

where $s_i(t)$ are 2 source signals. $x_i(t)$ are 2 observation signals where represents $x_1(t),\ x_2(t)$ and $\hat{s}(t)$ is the estimation of 2 source signals.

$$\hat{s}(t) = WAs(t) \tag{5.7}$$

$$\begin{bmatrix} \widehat{s_1} \\ \widehat{s_2} \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} \tag{5.8}$$

$$\hat{s}_1(t) = as_1(t) + bs_2(t) \tag{5.9}$$

where *a*, *b*, *c*, *d* are the elements of the global mixing matrix (WA) and it is assumed that for the notation $\hat{s}_1(t)$ term represents estimated speech signal. The interference of the noise signal to speech signal can be calculated with the Equation (5.8) and (5.9).

For better understanding, it is explained with an example,

$$G_1 = \begin{pmatrix} 1 & 10 \\ 0 & 1 \end{pmatrix} \tag{5.10}$$

$$G_2 = \begin{pmatrix} 5 & 0.1 \\ 0.1 & 2 \end{pmatrix} \tag{5.11}$$

In case of G_1, the second source is estimated but the first one has affected by the other sources. Hence, this is not true estimation. It is clear that G_2 means that the source signals have been estimated correctly and the effect of non-diagonal elements should have take into account when computing SNR.

The ICA algorithm performance is performed for all 11600 sounds. The global matrix is used when computing SNR calculation after noise reduction in section 5.5. The residual of ICA $P_{residual}(ICA)$ is the interference of the noise to the speech signal can be used to calculation of the overall performance. To conclude, the aim of performing the global matrix is to determine whether the estimation has an acceptable solution or not.

5.3.2 Performance of ICA block

In some cases, ICA block cannot estimate the unmixed matrix elements. In other words, ICA block doesn't find a suitable plane for the separation of observation signals. This is given as percentage ratio of the non-separated case to separated one. The results are shown in Table 5.10.

Table 5.10 Ratio of non-separation case to separation case for ICA block

(%)	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	0.086	0.129	0.387	0.309
5db	0.129	0.064	0.515	0.332
0db	0.094	0.064	0.632	0.244
-5db	0.094	0.086	0.502	0.270
-10db	0.112	0	0.517	0.296

In some cases, ICA block decided that the mixed signals contain only one source. This situation occurs when the VAD block decision is wrong. This is related with the non-speech hit rate (HR0). VAD block can decide the signal contains speech even it does not. In this case ICA output gives the noise signal only and this signal suppressed with the ANC block. By this way, HR0 error of VAD is reduced by ICA.

Table 5.11 Percentage of ICA block finds only one signal in mixing signal

%	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	0.18	0.90	0.28	0.45
5db	0.18	2.04	0.33	1.94
0db	1.08	9.82	0.64	5.64
-5db	5.59	16.91	2.11	8.64
-10db	14.37	29.07	7.11	12.56

5.3.3 The Output of ICA

The outputs of ICA block are estimated sources and it is not known whether speech is in the output channel 1 or 2. To determine this case, a simple feature is used, namely variance of autocorrelation signals. The variance of autocorrelation speech signals is low comparing to noise signals.

.

5.4 ANC Results

ANC block performs after the ICA block. If the other algorithms perform perfectly, ANC block will be applied only for the noise signal. The power spectrum density is computed for the various types of noise (conveyor band, ambiance, air hammer and traffic sounds).

5.4.1 Feedforward ANC with Different Sounds

The time domain and frequency domain of the input and error signals are represented for various sounds. It is important to determine the filter parameter in the Fx-LMS ANC system. Step size is 0.008 and the filter length is chosen 256 as a result of the extensive simulations. Power spectrum densities of the signals are negative because the signal magnitudes are smaller than 1. The noise eliminated by ANC is desired signal in Figure 5.14, 5.16, 5.18 and 5.20.

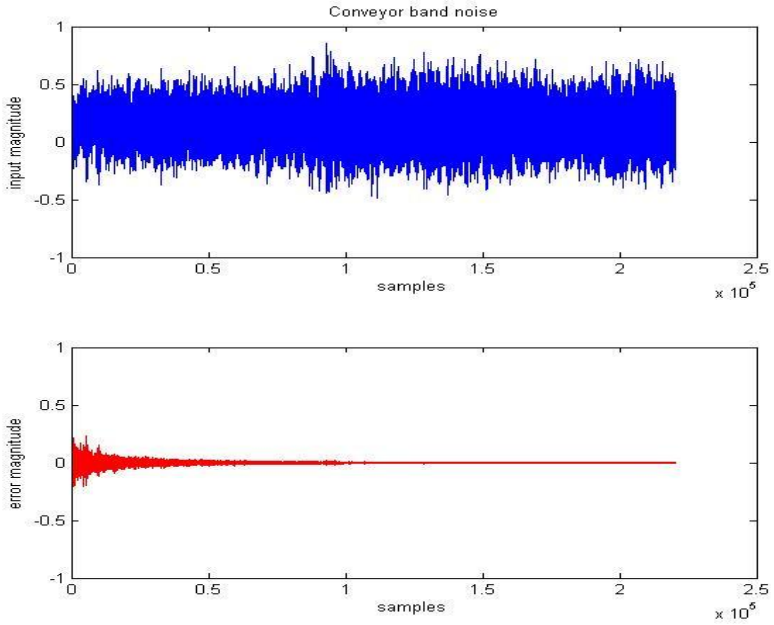

Figure 5.13 Time domain representation of input and error signal of conveyor band sound with Fx-LMS algorithm

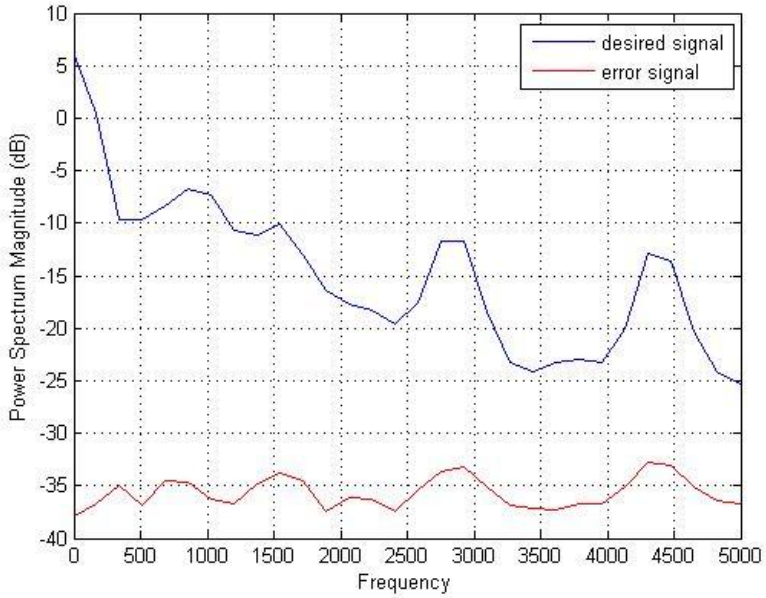

Figure 5.14 Frequency domain representation of input and error signal of conveyor band sound with Fx-LMS algorithm

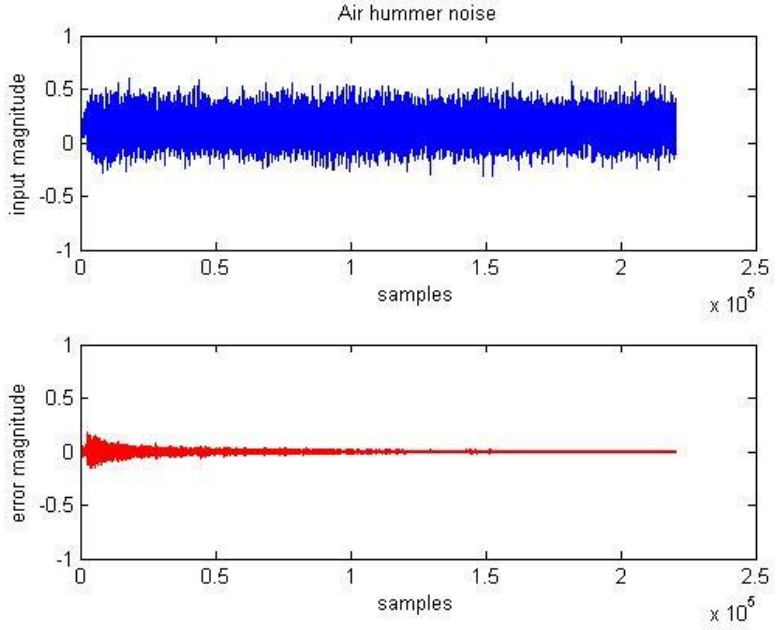

Figure 5.15 Time domain representation of input and error signal of air hammer sound with Fx-LMS algorithm

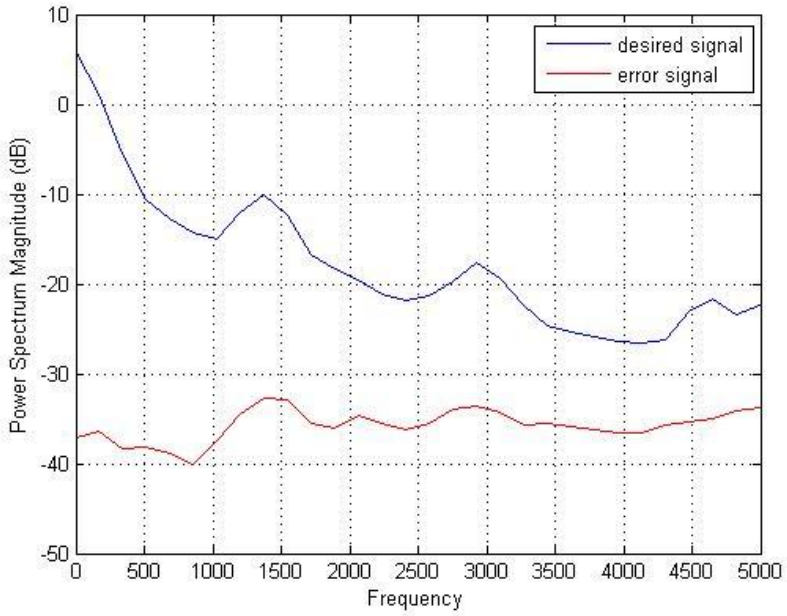

Figure 5.16 Frequency domain representation of input and error signal of air hammer sound with Fx-LMS algorithm

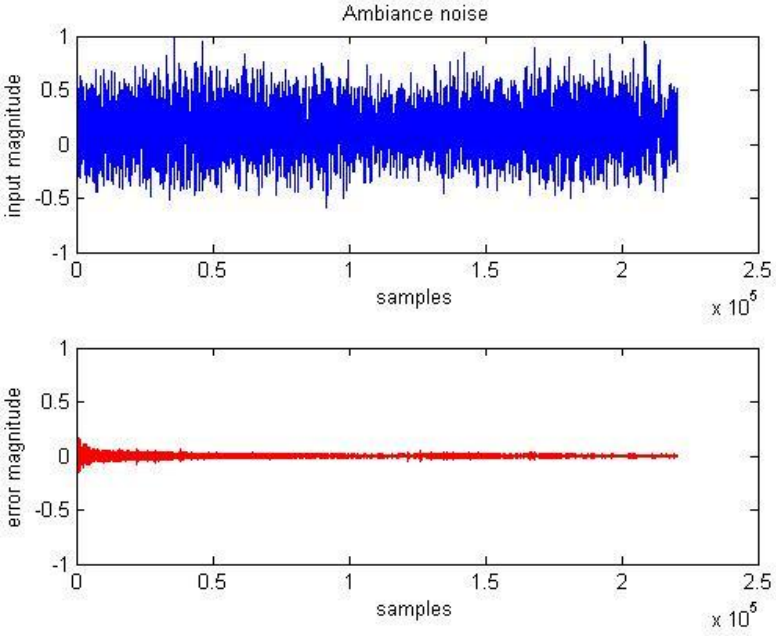

Figure 5.17 Time domain representation of input and error signal of ambiance sound with Fx-LMS algorithm

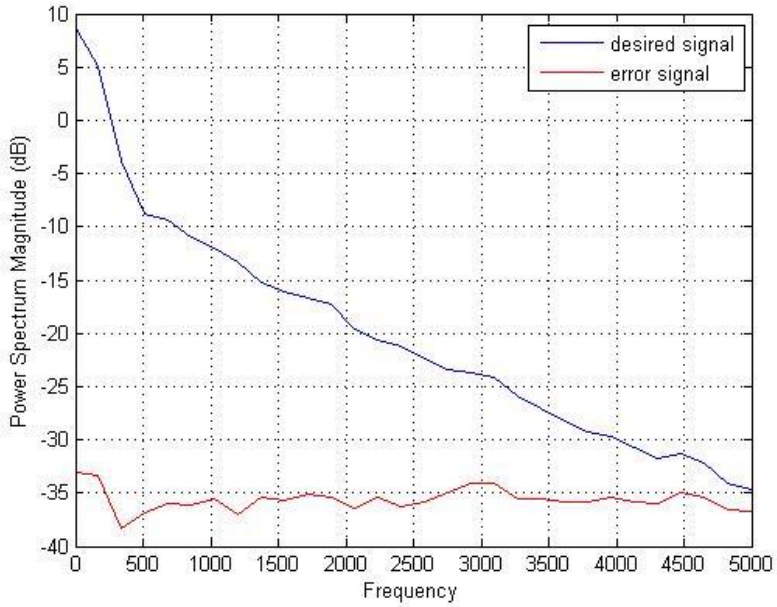

Figure 5.18 Frequency domain representation of input and error signal of ambiance sound with Fx-LMS algorithm

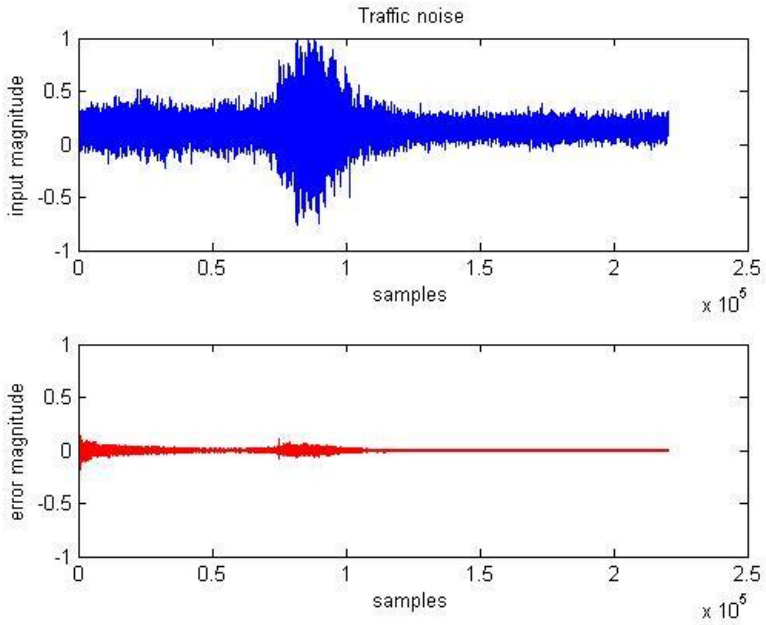

Figure 5.19 Time domain representation of input and error signal of traffic sound with Fx-LMS algorithm

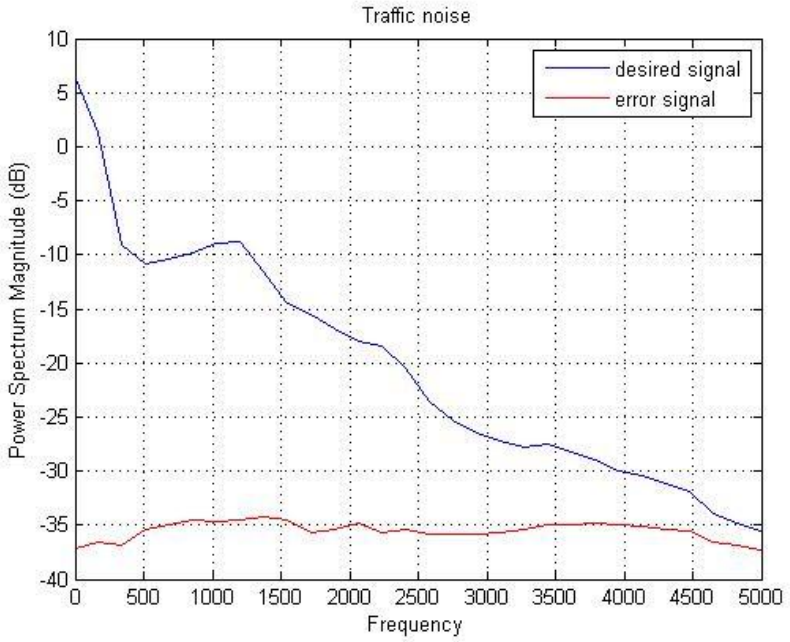

Figure 5.20 Frequency domain representation of input and error signal of traffic sound with Fx-LMS algorithm

5.4.2 The Output of ANC

The output of ICA and ANC block representations are given in Figure 5.21 and 5.22. If the VAD system returns '1', the ICA system works on and separates the mixture signals. Then ANC system eliminates the noise only. If the VAD system returns zero, ICA system doesn't work, ANC system directly eliminates the noise signal. ANC block must be run for all cases so ANC block is applied to 11600 sounds and its performance is given in section 5.5. The psd of source signals is given Figure 5.23. The psd of one of mixture signals is plotted (ANC is not used) and the psd of summation speech and residual error is plotted (ANC is used) in Figure 5.24.

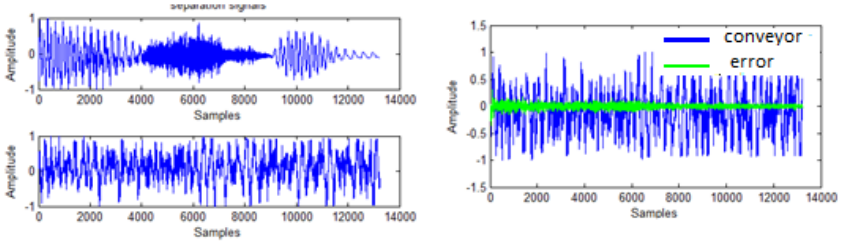

Figure 5.21 The output of blocks if the VAD returns '1'

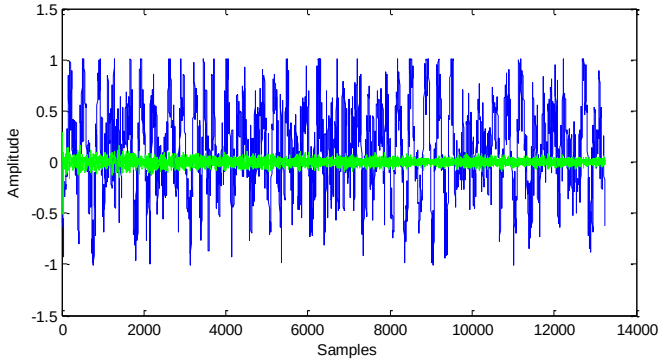

Figure 5.22 The outputs of blocks if the VAD returns '0'

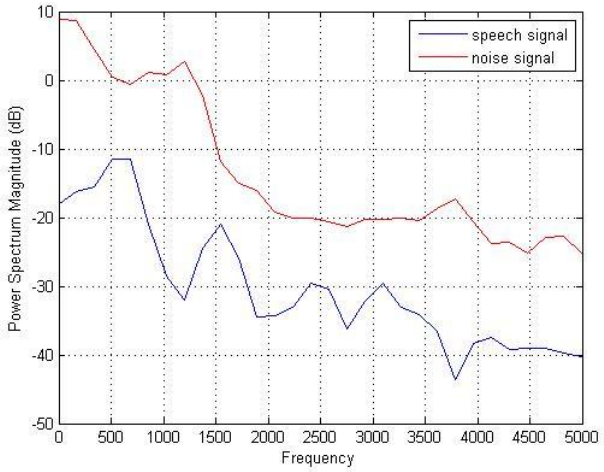

Figure 5.23 The PSD of two source signals

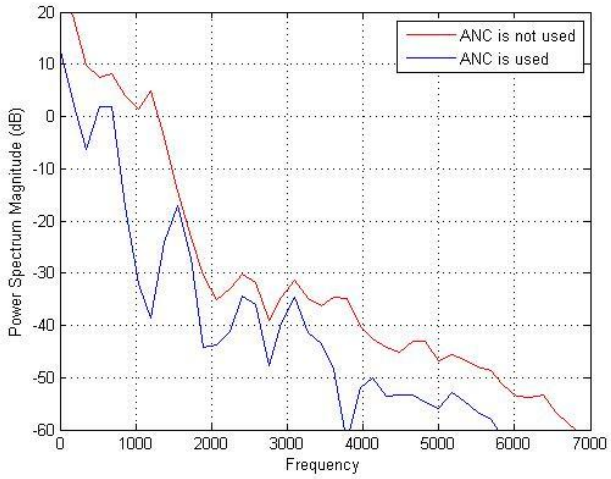

Figure 5.24 The PSD of signals at the system output when ANC is used or not.

5.5 The General Performance of the Proposed System

As shown in Figure 5.25, the proposed system contains different algorithm blocks. Performance of the system is calculated considering the performances of each block. At first, the power of input speech and noise are computed. The ratio of speech power and noise power is SNR_{before}. Then each block runs, the speech and residual error (it comes from ANC output) are obtained at the output. Note that the effect of ICA takes into consideration as mentioned in section 5.3.1. As it is indicated before, the residual of ICA $P_{residual}(ICA)$ which is the interference of the noise to the speech signal can be used to calculation of the overall performance. The general performance of this system is performed for all 11600 sounds by using Equation (5.12) and (5.13). The average results are shown in Table 5.12.

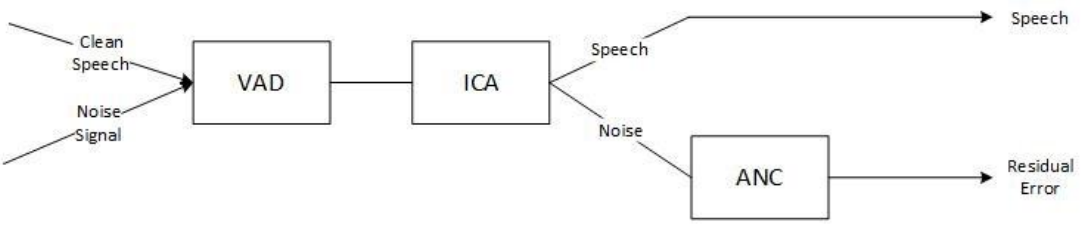

Figure 5.25 Block diagram of the proposed system

$$\mathrm{SNR}_{\mathrm{before}} = 10\log_{10}\left(\frac{\mathrm{P}_{\mathrm{speech(in)}}}{\mathrm{P}_{\mathrm{noise(in)}}}\right) \tag{5.12}$$

$$\mathrm{SNR}_{\mathrm{after}} = 10\log_{10}\left(\frac{\mathrm{P}_{\mathrm{speech(out)}}}{\mathrm{P}_{\mathrm{residual\ noise(out)}}+\mathrm{P}_{\mathrm{residual\ ICA}}}\right) \tag{5.13}$$

Table 5.12 The performance comparison of the proposed system

SNR before:	Conveyor band	Air Hammer	Ambiance	Traffic Noise
10db	15.97db	13.24db	7.65db	15.16db
5db	12.52db	10.65db	6.07db	13.55db
0db	9.41db	7.64db	3.93db	10.26db
-5db	6.44db	3.32db	1.29db	8.53db
-10db	2.22db	-2.86db	2.30db	5.32db

The result shows that the proposed system performance is great, even for the low SNR levels. The important part of system is VAD. If VAD makes a right decision, the other parts run correctly. When the proposed feature is used, HR1 and HR0 rates have the best performances. Moreover, the proposed system well performs for all noise types with different sound levels.

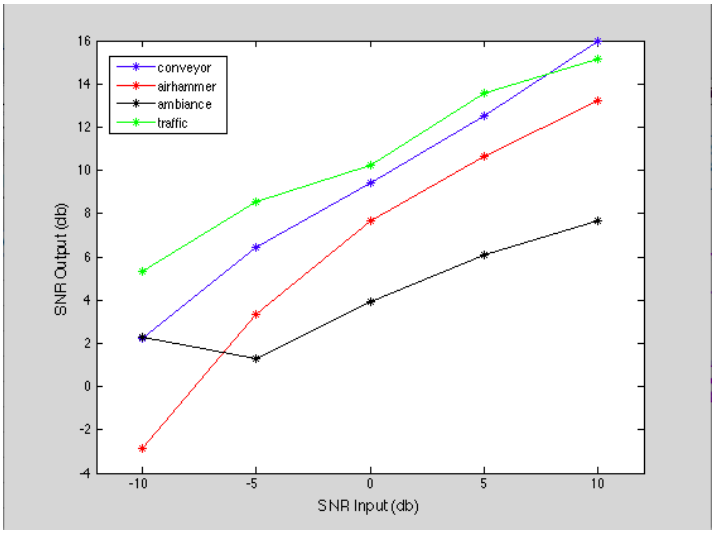

Figure 5.26 Comparison of SNR performance between input and output

CHAPTER SIX
CONCLUSION

In this thesis, a system which is a combination of VAD, ICA and ANC methods is proposed for the smart noise elimination. The usual ANC systems eliminate both unwanted and desired sounds. This proposed system provides that while the noise is eliminated, the speech is preserved in a noisy environment. (Up to our knowledge, the system is suggested for the first time). With the help of this system a worker will be able to hear speech in a noisy environment.

The noise dataset consists of several different samples for each different noise type such as conveyor band, ambiance, air hammer and traffic noise sounds. These sound samples are mixed with the speech sound taken from PTDB-TUG speech database (Pirker, Wohlmayr, Petrik, & Pernkopf, 2011) under five different SNR condition. So the dataset used in this thesis contains 11600 sounds of which 12 seconds. These sounds contain both speech and noise in the first and last four seconds; and the remaining four seconds contains only the noise.

As a first block, the properties of standard VAD block are discussed. VAD block has significant importance in the proposed system. The performance of VAD block will affect the whole system. For example, if VAD block decides as noise for a sound that contains speech, ICA block will be skipped and ANC block will eliminate both noise and speech. In the contrary if it decides as speech for a noise sound, noise will be estimated by the ICA block, this block recovers the fault of VAD. Then, estimated noise will be eliminated by ANC. ICA block can be run in the first step. In that case, the noise and speech are separated by the ICA block. However, the evaluation time for ICA block is greater than VAD block. Therefore, it is decided to run VAD block in the first step. Furthermore, ICA block is only active when the speech exists. This is more suitable case for real time applications.

The VAD block is tested with the different noise types such as conveyor band, ambiance, air hammer and traffic noises with different SNR levels (10db, 5db, 0db,

-5db, -10db). Since the proposed system will be used in real environments, especially in industrial workplaces, its performance must be high in all SNR levels. The first G.729 VAD system is used for VAD in the proposed system but it is seen that the performance decreases significantly in the low SNR level.

Since the performance of the whole system highly depends on the VAD block, more professional algorithms that can perform well even in the lowest SNR level must be used. Modified Mel-spectrum feature is proposed. Its performance is compared with the well known features. This feature has better performance compared to the other features such as spectral flux, spectral roll-off, and spectral centroid. Although the performances of frequency features are good at high SNR level, they are ineffective at the low SNR level and the non-speech hit rates are generally worse. The modified Mel-spectrum feature is used in VAD system design. Speech hit rate and non-speech hit rates are higher than the G.729 VAD even if the G.729 VAD uses four features (zero crossing rate, low-band energy, full-band energy and line spectral frequencies). The proposed VAD has a better performance than the standard G.729 VAD. Also speech hit rate (HR1) has critic importance in the proposed system. Because, when VAD detects the speech frame as a non-speech frame, the ICA block does not run. Thus, this case cannot be tolerated. By using the proposed feature, the best HR1 rates are obtained.

In the second step, the simulations of ICA system are performed. The observation signals are obtained by non-singular symmetric matrix. The performance of ICA is investigated by using the global mixing matrix. ICA block doesn't find a suitable weight matrix (it is not converged during the estimation) for the observation signals but this case occurs rarely. The case when ICA block finds only one signal in the mixture, can be mostly caused by the wrong VAD decision. ICA block recovers the wrong VAD decision (VAD makes a decision that there is a speech signal but actually it is not.) and this increases the system performance. At the end of the ICA block, it gives two signals one of them noise and other one is speech. It is important to detect the speech signal is in which output channel before the noise elimination.

Otherwise, the input of ANC block can be speech signal. Thus, a simple feature is used to detect the speech signal and the other signal, namely noise.

As a third step, ANC block is designed with the Fx-LMS algorithm. The PSD of the error and input signals are calculated. The parameters of ANC block (step size and filter length) are determined as a result of the extensive simulations. The time domain and frequency domain graphs are plotted. It is clearly seen that ANC block eliminates the conveyor band and air hammer sounds very efficiently. Human audible range is sensitive from 0Hz to 5kHz. ANC block reduces these sounds approximately 25db on the average in this range. When the ANC block is applied to the traffic and ambiance sounds, their average noise reduction performance is lower than conveyor and air hammer noises performance which is 20db on the average.

The whole system performance is evaluated by comparing the input SNR level with the output SNR level. The results show that the proposed system performs well on these noises with different noise levels and improves the SNR level in almost all cases.

REFERENCES

Air hammer sounds, (n.d.). Retrieved August, 17, 2014 from https://www.freesound.org/

Alpaydın, E. (2010). *Introduction to machine learning (second edition).* The MIT Press Cambridge, Massachusetts London, England.

Ambiance sounds, (n.d.). Retrieved August, 17, 2014 from https://www.freesound.org/

Alexendre, E., Rosa, M., Cuadra, L., & Gil-Pita, R. (2006). Application of Fisher Linear Discriminant Analysis to Speech/Music Classification. *Audio Engineering Society Convention Paper, 6678,* 1-6.

Baig, M., Masud, S. & Awais, M. (2006). Support vector machine based voice activity detection. *IEEE, Intelligent Signal processing and Communications ISPACS'06 International Symposium,* 319-322.

Benyassine, A., Shlomot, E., Huan-yu, S. & Massaloux, D. (1997). A robust low complexity voice activity detection algorithm for speech communication systems. *Speech Coding For Telecommunications Proceeeding IEEE, Workshop,* 97-98.

Benyassine, A., Shlomot, E., Huan-yu, S. & Massaloux, D. (1997). ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *Communications Magazine, IEEE, 35,* (9), 64-73.

Chang, J., Kim, N. S. & Mitra, S. K. (2006). Voice activity detection based on multiple statistical models. *IEEE Transactions on Signal Processing, 54,* (6), 1965-1976.

Christensen, K. B. (1992). The application of digital signal processing to large-scale simulation of room acoustics: frequency response modeling and optimization software for a multichannel DSP engine. *Journal of Audio Engineering Society Vol.40 (4),* 260-276.

Conveyor band sounds, (n.d.). Retrieved August, 17, 2014 from https://www.freesound.org/

Conveyor belt sounds, (n.d.). Retrieved August, 17, 2014 from http://www.audiomicro.com/

Conveyor belt sound effects, (n.d.). Retrieved August, 17, 2014 from http://www.twistedtracks.com/sound-effects/machine-sound-effects/conveyor-belt-sound-effects/

Conveyor belt2 sound effect, (n.d.). Retrieved August, 17, 2014 from http://www.audioblocks.com/stock-audio/conveyor-belt-2-sound-effect.html#

Ekštein, K., & Pavelka, T. (2004) *Entropy And Entropy-based Features In Signal Processing.* Retrieved June 24, 2014, from http://daedalus.scl.sztaki.hu/phdws2004/abstract/phdws2004_abstract_4.pdf

Elliott, S. J. & Nelson, P. A. (1993). Active noise control. *IEEE Signal Processing Magazine, 10 (4),* 12-35.

Enqing, D., Heming, Z. & Yongli, L. (2002). Low bit and variable rate speech coding using local cosine transform. *Tencon'02 Proceedings. IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 1,* 423-426.

Erkan, F. (2009). *Design and implementation of a fixed point digital active noise controller headphone.* Retrieved September, 4, 2014 from http://etd.lib.metu.edu.tr/upload/12610758/index.pdf

Fredrik, F. (2011). *Active noise control in forest machines*. Retrieved September, 2, 2014 from http://umu.diva-portal.org/smash/get/diva2:451753/FULLTEXT01.pdf

Ghosh, P. K., Tsiartas, A., & Narayanan, S., (2011). Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio,Speech,and Language Processing, 19,* (3), 600-613.

Haigh, J., & Mason, J. (1993). Robust voice activity detection using cepstral features. *TENCON'93 Proceedings. Computer, Communication, Control and Power Engineering IEEE Region 10 Conference, 3,* 321-324.

Hao, Y., & Wilamowski, B. M., (2010). *Levenberg-Marquardt Training*. Retrieved December, 26, 2014 from http://www.eng.auburn.edu/~wilambm/pap/2011/K10149_C012.pdf

Ho, T. K. (1998). Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8),* 832-843.

Hong-yan, L. & Guang-long, R. (2010). Blind separation of noisy mixed speech signals based Independent Component Analysis. *Persive Computing Signal Processing and Applications (PCSPA), 2010 First International Conference on,* 586-589.

Hong-yan, L., Qing-hua, Z., Guang-long, R. & Bao-jin, X. (2009). Speech enhancement algorithm based on independent component analysis. *Natural Computation, ICNC'09. Fifth International Conference, 2,* 598-602.

Hyvärinen, A. & Oja, E. (2000). Independent component analysis: algorithms and applications. *ScienceDirect Publication: Neural Networks 13,* (4-5), 411-430.

Itoh, K. & Mizushima, M. (1997). Environmental noise reduction based on speech/non-speech identification for hearing aids. *Acoustics, Speech and Signal Processing, ICASSP-97, IEEE International Conference, 1,* 419-422.

ITU-T Series G: Transmission Systems and Media, Digital Systems and Networks. Digital terminal equipments-Coding of voice and audio signals, (2012). Retrieved August, 10, 2014, from https://www.itu.int/rec/T-REC-G.729-199610-S!AnnB/en

Kajikawa, Y. & Hirayama, R. (2010). Feedback active noise control system combining linear prediction filter. *18th European Signal Processing Conference (EUSIPCO),* 31-35.

Kataja, J. (2012). *Development of a robust and computationally-efficient active sound profiling algorithm in a passenger car.* Retrieved September, 4, 2014 from http://www.vtt.fi/inf/pdf/science/2012/S5.pdf

Khoa, P. C. (2012). *Noise robust voice activity detection, School of Computer Engineering.* Retrieved August, 8, 2014 from http://www.ntu.edu.sg/home/aseschng/SpeechTechWeb/members/Conformation_theses/Ben_main.pdf

Kos, M., Grašič, M. & Kačič, Z. (2009). Online Speech/Music Segmentation Based on the Variance Mean of Filter Bank Energy. *Hindawi Publishing Corporation EURASIP Journal on Advances in Signal Processing*, 1-13.

Kos, M., Kačič, Z. & Vlaj, D. (2013). Acoustic classification and segmentation using modified spectral roll-off and variance-based features. *Digital Signal Processing, 23,* (2), 659-674.

Lamel, L., Kassel, R. & Seneff, S. (1986). *Speech database development: design and analyisis of the acoustic-phonic corpus.* Proceedings of the DARPA speech Recognition Workshop Report No. SAIC-86/1546.

Li, W., Gu, F., Ball, A. D., Leung, Y. T. & Phipps, C. E. (2001). A study of the noise from diesel engines using the independent component analysis. *Mechanical Systems and Signal Processing, 15(6),* 1165-1184.

Lu, H., Pan, W., Lane, N. D., Choudhury, T. & Campbel,l A. T. (2009). SoundSense: scalable sound sensing for people-Centric applications on mobile phones. *MobiSys'09*, 14.

Mitianoudis, N. & Davies, M. E. (2003). *Audio source separation of convolutive mixtures. IEEE Transactions on Speech and Audio Processing 11*, 5, 489-497.

Moattar, M. H. & Homayounpour, M. M.(2009). A simple but efficient real-time voice activity detection algorithm. *17th European Signal Processing Conference (EUSIPCO 2009),* 2549-2553.

Noise Control: A practical approach to controlling noise in the workplace, (n.d.). Retrieved July 15, 2014, from http://www.acc.co.nz/PRD_EXT_CSMP/groups/external_ip/documents/publications_promotion/wpc088755.pdf

Pham, T. V., Tang, C. T. & Statdtschitzer, M. (2009). Using artifical neural network for robust voice activity detection under adverse conditions. *Computing and Communication Technoloiges, RIVF'09 International Conference,* 1-8.

Prasad, R. V., Sangwan, A., Jamadagni, H.S., Chiranth M. C. & Sah R. (2002). comparison of voice activity detection algorithms for VoIP. *IEEE, Proceedings of the Seventh International Symposium on Computers and Communications (ISCC'02)*, 530-535.

Prahallad, K. (2008). Speech Technology: A Practical Introduction Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis. Retrieved July 10, 2014, from http://www.speech.cs.cmu.edu/15-492/slides/03_mfcc.pdf

Pirker, G., Wohlmayr, M., Petrik, S. & Pernkopf, F. (2011). A pitch tracking corpus with evaluation on multipitch tracking scenirio. *Interspeech Conference,* 1509-1512.

Quatieri, T. F. (2002). *Discrete-time speech signal processing: principles and practice.* (1st ed.) USA: Prentice Hall PTR

Ramirez, J., Segura, J. C., Benitez, C & Torre, A., Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication, 42,* (3-4), 271-287.

Sohn, J. & Lee, M. (2000). Selective attention system using new active noise controller. *Neurocomputing 31,* 197-204.

Sohn, J., Kim, N. S. & Sung, W. (1999). A statistical model-based voice activity detection. *Signal Processing Letters, IEEE, 6,* (1), 1-3.

Sonnleitner, R., Niedermayer, B., Widmer, G. & Schlüter, J. (2012). A simple and effective spectral feature for speech detection in mixed audio signals. *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12),* 1-7.

Speech database PTDB-TUG, Signal Processing and Speech Communication Laboratory, (n.d). Retrieved August, 17, 2014 from http://www.spsc.tugraz.at/tools/extra-example-based-automatic-phonetic-transcription

Stergiou C. & Siganos D., (n.d). *Neural Networks*. Retrieved August, 18, 2014 from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html

Thom, J., Peters, C., McIntyre, E. & Winters, M. (2005). *Active noise control communication headsets for the entertainment industry.* Retrieved from August 9, 2014, from https://circle.ubc.ca/bitstream/handle/2429/815/ANCHeadsetsFinalRevised.pdf?sequence=1

Traffic sounds, (n.d.). Retrieved August, 17, 2014 from https://www.freesound.org/

Trajkovic, M., Gutta, S., & Cohen-Solal, E. (2002). *Active noise cancelling headset and devices with selective noise suppression.* Retrieved August, 9, 2014 http://www.google.com/patents/US20020141599

Tüske, Z., Mihajlik, P., Tobler[+], Z., & Fegyó[+], T. (2005). Robust voice activity detection based on the entropy of noise-suppressed spectrum. *INTERSPEECH 2005*, 245-248

Verteletskaya, E., & Sakhnov, K. (2010). Voice activity detection for speech enhancement applications. *Acta Polytechnica, 50,* (4), 100-105.

Voice activity detector (VAD) for adaptive multi-Rate (AMR) speech traffic channels, (1998). Retrieved August, 10, 2014, from http://www.etsi.org/deliver/etsi_en/301700_301799/301708/07.01.00_40/en_301708v070100o.pdf

Wu, B., & Wang[+], K. (2006). Voice activity detection based on auto-correlation function using wavelet transform and teager energy operator. *Computational Linguistics and Chinese Language Processing, 11,* (1), 87-100.

Wu, J., & Chiu, S. (2001). Independent component analysis using potts models. *IEEE Transactions on Neural Networks, 12,* (2), 202-211.

Ying, D., Yan, Y., Dang, J., & Soong, F. K. (2011). Voice activity detection based on an unsupervised learning framework. *Audio, Speech and Language Processing, IEEE Transaction, 19*, (8), 2624-2633.

Zhang, Y., (2003). *Seminar speech recognition.* Retrieved June 25, 2014, from http://www.liacs.nl/~erwin/SR2003/Students/04_Mel-spectrum%20Computation.ppt