CHAPTER 1

INTRODUCTION TO SPEECH ENHANCEMENT

This chapter gives an introduction to speech enhancement, its applications and common sources of noise that degrade speech. The main challenges and issues related to single channel enhancement motivated this work and an outline of the thesis is also described.

1.1 GENERAL INTRODUCTION

Communication via speech is one of the essential functions of human beings. Humans possess varied ways to retrieve information from the outside world or to communicate with each other and the three most important sources of information are speech, images and written text. For many purposes, speech stands out as the most efficient and convenient one. Speech not only conveys linguistic contents, but also communicates other useful information like the mood of the speaker. When speaker and listener are near to each other in a quiet environment, communication is generally easy and accurate. However, at a distance or in a noisy background, the listener’s ability to understand suffers (Douglas 2005). In many speech communication systems the quality and intelligibility of speech is of greatest importance for ease and accuracy of information exchange. The speech processing systems used to communicate or store speech is usually designed for a noise free environment but in a real-world environment, the presence of background interference in the form of additive background and channel noise drastically degrades the performance of these systems, causing inaccurate information
exchange and listener fatigue. Over the year, researchers have developed a number of methods to enhance speech from the degraded speech. Yet, due to complexities of the speech signal, restoring the desired speech signal from the mixture of speech and background noise still poses a considerable challenge in speech processing and communication system research.

1.2 SPEECH ENHANCEMENT AND IT’S APPLICATIONS

Speech enhancement deals with processing of noisy speech signals, aiming at improving their perception by human or their correct decoding by machines (Berouti et al 1979). Speech enhancement algorithms attempt to improve the performance of communication systems when their input or output signals are corrupted by noise. The presence of background noise causes the quality and intelligibility of speech to degrade. Here, the quality of speech refers how a speaker conveys an utterance and includes such attributes like naturalness and speaker recognizability. Intelligibility is concerned with what the speaker had said, that is, the meaning or information content behind the words (Hu and Loizou 2007). Therefore, a noisy environment reduces the speaker and listeners ability to communicate. To reduce the impact of this problem speech enhancement can be performed. It is usually difficult to reduce noise without distorting speech and thus, the performance of speech enhancement systems is limited by the tradeoff between speech distortion and noise reduction (Boll 1979).

Efforts to achieve higher quality and/or intelligibility of noisy speech may effectively end up improving performance of other speech applications such as speech coding/compression and speech recognition, hearing aids, voice communication systems and so on. The goal of speech enhancement varies according to specific applications, such as to reduce listener fatigue, to boost the overall speech quality, to increase intelligibility and to improve the performance of the voice communication device. Hence
speech enhancement is necessary to avoid the degradation of speech quality and to overcome the limitations of human auditory systems.

1.3 COMMON SOURCES OF NOISE THAT DEGRADE SPEECH

For communication systems, two general objectives depend on the nature of the noise and often on the signal to noise ratio (SNR) of the distorted speech. With medium to high input SNR, reducing the noise level can produce a subjectively natural speech signals at a receiver or can obtain reliable transmission. For low SNR, the objective could be to decrease the noise level, while retaining or increasing the intelligibility and reducing the fatigue caused by heavy noise for example motor and street noise. Figure 1.1 shows the factors that affect the speech signal during transmission at various stages by different noise sources. Sources that degrade speech quality are noisy environment during acquisition, background noise, multi-speaker effect, noisy transmission channel and imperfect speech reproduction. In the transmission side the effect of background noise are added with the desired signal and the signal from other speakers are treated as noise for the desired speaker. The signal with background noise is transmitted through the channel where the transmission channel noise is also added with the desired signal.

![Diagram of common sources of noise](image-url)

**Figure 1.1 Common sources of noise**
The nature of the noise is an important factor in deciding on a speech enhancement method. Therefore, a good model of noise is important for the performance of speech enhancement system and it is important to analyze how well a speech enhancement algorithm/model works with different types of noise (Kamath and Loizou 2002). Noise can be different based on various statistical, spectral or spatial properties. Based on the nature and properties of the noise sources, noise can be classified as additive background noise, interfering speakers (speech like noise), impulse noise, convolutive noise, and multiplicative noise. In general, it is more difficult to deal with non-stationary noise, where there is no prior knowledge available about the characteristics of noise. Since non-stationary noise is time varying, the conventional method of estimating the noise from initial intervals by assuming no speech signal is not suitable for estimation. Noise types, which are similar in temporal, frequency or spatial characteristics to speech, are also difficult to remove or attenuate. For instance, Multi talker babble retains some characteristics of speech and poses a particularly difficult problem for an algorithm intended to isolate speech signal from the background noise.

1.4 CLASSIFICATION OF SPEECH ENHANCEMENT METHODS

There are many ways to classify speech enhancement methods. It is usually difficult for a typical algorithm to be able to perform homogenously across all noise types. Therefore, usually a speech enhancement system is based on certain assumptions and constraints that are typically dependent on the application and the environment. In general the performance of a speech enhancement algorithm is limited by the following factors: limitations on the number of noise sources available, making different uses of a priori information about the signal of interest and/or the corrupting signal, limitations in the time variations (non-stationary) allowed for the corrupting
signal, model based limitations like the restriction of the algorithm to uncorrelated noise. From the proposed works of Lim and Oppenheim (1979), Berouti et al (1979), Gong (1995), Ephraim (1992), Virag (1999), Kamath and Loizou (2002) and Athanasios Mouchtaris et al (2007) the speech enhancement systems can be classified based on number of input channels (one/two/multiple), domain of processing (time/frequency) and type of algorithm (Non-adaptive/Adaptive). Typically, the speech enhancement literature broadly divides the various speech processing strategies under single and multichannel enhancement techniques (Philipos 2007).

1.4.1 Single Channel Speech Enhancement Systems

In most of the real-time applications like speaker and speech recognition, mobile communication and hearing aids, usually a second channel is not available. These systems are easy to build and comparatively less expensive than the multiple input systems. They constitute one of the most difficult situations of speech enhancement, since no reference signal to the noise is available and the clean speech cannot be preprocessed prior to being affected by the noise. Usually single channel systems make use of different statistics of speech and unwanted noise. The performance of these methods are usually limited in the presence of non-stationary noise as most of the methods make an assumption that noise is stationary during speech intervals and also, the performance drastically degrades at lower signal to noise ratios. Many of these methods are presented in more detail and strategies employed to overcome the above mentioned limitations are discussed in chapter 2.

Various research groups are working with single channel speech enhancement around the globe, a few of which are listed here. Speech enhancement using wavelet filtering method is one of the active researches in Polandelphia University, USA. Signal Processing Laboratory in Griffith
University is working on fast converging iterative Kalman filtering for speech enhancement using long and overlapped tapered windows with large side lobe attenuation. A corpus based approach to speech enhancement from non-stationary noise is also one of the research works in Queen's University at Belfast. In India, Bangalore Institute of Speech and Hearing (BISH), Bangalore is currently working on the enhancement of speech in cochlear implants and Indian Institute of Science, Bangalore and Scientific Analysis Group, New Delhi is working on Speech enhancement using Gaussian Mixture Models. Speech enhancement using bionic wavelet de-noising is one of the research projects in Marquette Speech and Signal Processing Lab. Centre for Robust Speech Systems, University of Texas at Dallas, Center for Spoken Language Understanding and Oregon Health and Science University are working on noise robust automatic speech and speaker recognition system with a speech enhancement system as a preprocessor.

1.4.2 Multi Channel Speech Enhancement Systems

These systems take advantage of the availability of multiple signal inputs to the system and make use of the noise reference in an adaptive noise cancellation device, the use of phase alignment to reject undesired noise components, or even the use of phase alignment and noise cancellation stages into a combined scheme (Kokkinakis and Loizou 2010). By taking into account the spatial properties of the signal and the noise source, the limitations inherent to one channel systems, particularly non-stationarity of noises can be better addressed. These systems tend to be more complex.

From this point onwards, the discussion in this thesis would be restricted to single channel enhancement techniques, since this forms the most common scenario on which enhancement algorithms is applied.
1.5 ISSUES IN SINGLE CHANNEL SPEECH ENHANCEMENT

Several automatic speech processing systems have also found their way in everyday life through their use in mobile communication, speech and speaker recognition, aid for the hearing impaired and numerous other applications (Krishnamurthy and Prasanna 2007). In all these speech communication systems the quality and intelligibility of speech is of utmost importance for ease and accuracy of information exchange. Both human and automatic speech communications are effective in controlled environments. This is due to the high quality and intelligibility of speech. However, in many situations of practical interest, speech signals are affected by various types of degradations like background noise, reverberation and speech from other speakers. The degraded speech needs to be processed to enhance the speech components present in the signal. The methods employed in practice take the nature of degradation into consideration for enhancing the speech components. This is because the signal characteristics will be different for each type of degradation.

Typically, more the processing employed for reducing the degrading component, more will be the distortion introduced. Hence speech enhancement is a trade-off between the actual reduction of degrading component and its own distortion. Therefore the performance of the speech enhancement methods is measured in terms of quality and intelligibility of the processed signal. The two performance measures are not correlated. It is also well known fact that improving the quality of the noisy signal does not necessarily elevate its intelligibility. On the contrary, quality improvement is usually associated with loss of intelligibility relative to that of the degraded signal (Epharim and Cohen 2006).
1.6 MOTIVATION

In the past decades, research in the field of speech enhancement has focused on the suppression of additive background noise (Lim and Oppenheim (1979), Ephraim (1992), He and Zweig (1999)). The presence of background noise in speech significantly reduces the intelligibility of speech. Degradation of speech severely affects the ability of a person, whether impaired or normal hearing, to understand what the speaker is saying. The ultimate goal of speech enhancement is to eliminate the additive noise present in speech signal and restore the speech signal to its original form. Several methods have been developed as a result of these research efforts. Most of these methods have been developed with some or the other auditory, perceptual or statistical constraints placed on the speech and noise signals. However, in real world situations, it is very difficult to reliably predict the characteristics of the interfering noise signal or the exact characteristics of the speech waveform. Hence, in effect the speech enhancement methods are sub-optimal and can only reduce the amount of noise in the signal to some extent. Due to the sub-optimal nature of these methods, some of the speech signal can be distorted during the process. In effect, there is a trade-off between distortions in the processed speech and the amount of noise suppressed.

The effectiveness of the speech enhancement system can therefore be measured based on how well it performs in light of this trade-off. Exploration of the same is the motivation for this research work. Complexity and ease of implementation of the noise reduction algorithms is also of concern in applications especially those related to portable devices such as mobile communications and digital hearing aids. In view of the above constraints and challenges there is a need to search for methods towards the improved performance of the speech enhancement algorithms. In this research the main focus is on the enhancement of speech for human listeners and for
speech processing systems in noisy environment conditions, as this is the most commonly encountered problem. Hence the main objectives of the research are:

i. To investigate the use of bionic wavelet and wavelet packet transform for speech enhancement and to propose new enhancement methods to attenuate the additive background noise without introducing speech distortion for an extended range of noise types and noise levels, thereby making the proposed methods more suitable than the existing algorithms for speech processing applications like speech recognition, speaker recognition and aid for the hearing impaired.

ii. To compare the performance of the proposed schemes with existing methods by conducting both objective and subjective test.

iii. To find whether the proposed speech enhancement methods are suitable to increase the recognition accuracy of speaker recognition system under real life noise conditions.

iv. To propose a new noise robust automatic speaker recognition system by incorporating speech enhancement algorithm as a front end processor to achieve improved overall recognition accuracy and to carry out a comprehensive evaluation and performance comparison of various speech enhancement algorithms including the proposed one on speaker recognition task under noisy environments.

1.7 CONTRIBUTIONS TO THE PROPOSED WORK

Current speech enhancement methods belongs to two categories namely, time domain methods such as the subspace approach and frequency
domain methods such as the spectral subtraction, Minimum Mean Square Error (MMSE) estimator (Ephraim and Malah 1984) and Wiener filtering (Lim and Oppenheim 1978). Both methods have their own advantages and drawbacks. The subspace methods provide a mechanism to control the tradeoff between speech distortion and residual noise, but with the cost of a heavy computational load. Frequency domain methods, on the other hand, usually consume less computational resources, but do not have a theoretically established mechanism to control tradeoff between speech distortion and residual noise. Among them, spectral subtraction is computationally efficient and has a simple mechanism to control tradeoff between speech distortion and residual noise, but suffers from a notorious artifact known as “musical noise”. The MMSE estimators and Wiener estimator have a moderate computation load, but have no mechanism to control tradeoff between speech distortion and residual noise.

Wavelet based methods using thresholding techniques are promising for coping with real life noise of various kinds. However many improvements have yet to be made to render this approach more flexible and robust. The main purpose of this Thesis is to propose new enhancement scheme that can handle any type of real world noise which does not affect the speech signal uniformly over the entire spectrum. In this work, two wavelet based approaches are proposed to reduce the additive non stationary noises effectively which occur in real life like street noise. The schematic overview and the outcome of the thesis are shown in Figure 1.2. First method is a new time adaptive discrete wavelet thresholding (TADWT) based on bionic wavelet transform (BWT) to achieve better enhancement of speech with improved perceptual quality. The BWT used in this work is discrete in nature and is based on the Daubechies mother wavelet named as Time Adaptive Discrete Wavelet Transform (TADWT). In this method DWT is taken for pre-filtered noisy speech initially and later to derive time adaptive
Figure 1.2 Schematic overview of the Thesis

In this method DWT is taken for pre-filtered noisy speech initially and later to derive time adaptive discrete wavelet transform coefficients derived. The
adaptive nature is captured by introducing a time varying linear factor at each scale over time to update the coefficients. Then threshold value is determined by using Stein’s Unbiased Risk Estimator (SURE) and de-noising of coefficients is made by using the new modified soft thresholding function. Due to the time adaptation nature of wavelet coefficients and the advantages of the new modified soft thresholding strategy, significant improvement is obtained in speech quality with reduced speech distortion.

Second approach is the combined spectral subtraction and wavelet packet based thresholding (SSWPT) method to suppress the noise without introducing any perceptible distortion in the signal for the speech that has been corrupted extremely by high levels of noise like cockpit noise in air-ground communication scenario. In this approach dual band spectral subtraction (DSSS) method with adaptive noise estimator is used as the pre-processor for reducing the noise initially. Further, WPT denoising is done to handle noise dominated speeches in real world environment (with SNR less than zero) and to get better quality and intelligibility of speech. This proposed technique provides a mechanism to control the tradeoff between speech distortion and residual noise. The performance of the proposed TADWT and SSWPT are evaluated by conducting experiments to find various objective quality measures like output SNR, segmental SNR, MMSE, IS distance and subjective measure by collecting opinion score depending on the nature of degradation.

Finally, an experimental evaluation of the reported TADWT and SSWPT methods is made for speaker recognition task under uncontrolled environments. Usually, ASR system gives good performance in controlled environments than in uncontrolled environments. New ASR system is suggested and its performance is studied
in both controlled and uncontrolled environmental conditions. The reported TADWT and SSWPT SE algorithms are incorporated in to the proposed noise robust Automatic Speaker Recognition (ASR) system as a front end processing stage to improve the speaker specific features and hence speaker recognition performance in noisy environment. The recognition results showed that the suggested methods gave relatively higher and competitive performance than other front end processing methods.

1.8 ORGANISATION OF THE THESIS

The Thesis is organized in the following manner:

Chapter 2 is a presentation of background information including the production of speech and its acoustic aspects, summary of single channel speech enhancement approaches. Further, the overview of literature on various existing single channel algorithms and various research groups working with single channel SE around the globe and it brings out the need for the present work is also discussed. In addition the fundamental concept of wavelet de-noising procedure is detailed.

Chapter 3 deals with the details of the first contribution made in this thesis. The proposed TADWT for enhancing speech with modified soft rule is explained in this chapter. In addition, different experimental studies based on objective and subjective quality measure are performed and performance evaluation with existing methods is also described.

Chapter 4 presents another hybrid scheme by combining spectral subtraction and wavelet packet based thresholding named as SSWPT for enhancing speech capable of handling noise dominated speeches in real life conditions. In addition this chapter discusses the several performance
measures to assure the efficiency improvement of the proposed method than the existing baseline methods.

In chapter 5, the effectiveness of the two proposed speech enhancement algorithms for increasing the noise robustness of the new Automatic Speaker Recognition system (ASR) is discussed in detail. In addition the performance of the ASR is analyzed based on various feature extraction and speaker modeling techniques. The results of ASR experiments under clean speech condition and under noisy environment are described. Improvement over the performance of the proposed ASR, in presence of different types of background noises by including front end processing stage (SE) is also discussed in this chapter.

Chapter 6 summarizes the work presented in this thesis, highlights the main contributions of the work, draws the conclusion and provides suggestions for future work. Finally the references used in this research work and publications made out of this research work are listed.