CHAPTER 2

PRODUCTION OF SPEECH AND OVERVIEW OF SINGLE CHANNEL ENHANCEMENT TECHNIQUES

2.1 INTRODUCTION

Background information needed for speech signal processing is discussed in this chapter. The main topics covered are production of speech, its source/filter model, summary of existing single channel speech enhancement and the advantages of wavelet based methods. In addition the basic concepts of the classic wavelet transform, Discrete Wavelet Transform and its implementation, selection of wavelet, wavelet denoising procedure and different wavelet de-noising issues as approached in literature are discussed.

2.2 PRODUCTION OF SPEECH AND ITS ACOUSTIC ASPECTS

Speech is an acoustic waveform which is a dynamic, unique and information-bearing signal. These waves are produced due to the sound pressure generated in the mouth of the speaker as a result of some sequence of coordinated movements of a series of structures in the human vocal system. Speech segments could be coarsely divided into voiced and unvoiced sounds. For the production of voiced sounds, the lungs press air through the vocal cords that vibrate and interrupt the air stream to produce a quasi-periodic pressure wave. The frequency of the pressure impulses is
commonly called the pitch frequency. The pitch impulses stimulate the air in the mouth and for certain sounds also the nasal cavity. When the cavities resonate, they radiate a sound wave which is the speech signal. Both cavities act as resonators with characteristic resonance frequencies, called formant frequencies. In the case of unvoiced sounds, the excitation of the vocal tract is more noise-like and no vibration of the vocal cords is involved. Acoustic speech is commonly regarded as resulting from a combination of a source of sound energy (the larynx) modulated by a transfer function (filter) determined by the shape of the vocal tract. The human speech production can be illustrated by a simple model depicted in Figure 2.1. In this model, a switch controls the selection between voiced and unvoiced sound. Excitation of voiced sounds is modeled by an impulse train and excitation of unvoiced sounds is modeled by random noise. In both cases, excitation is fed into a spectrum shaping filter that models the vocal tract. This model is commonly used due to its simplicity. (Ephraim and Cohen 2006)

![Figure 2.1 Speech Production Model](image)

**Figure 2.1 Speech Production Model**

This model is often referred to as the source-filter theory of speech production. For the unvoiced part of speech, the source can be seen as random noise while for the voiced part of speech, pulse train acts as the
energy source. In Figure 2.1, $A_N$ and $A_V$ are the gain for energy sources of the unvoiced and voiced speech respectively. Normally, for short segments or frames of speech (10-30 ms) the shape of the vocal tract remains relatively same, results the valid use of a linear time-invariant filter. Therefore, the speech signal can be seen as a stationary random process, which allows the use of the Short Time Fourier Transform (STFT) or other stationary analysis techniques.

2.3 SUMMARY OF SINGLE CHANNEL SPEECH ENHANCEMENT METHODS

The main purpose of speech enhancement is to reduce additive noise while minimizing the degree of distortion of the desired speech signal. For single channel spectral based techniques, there are generally three approaches to separate speech from noise, although some methods tend to have combined characteristics.

2.3.1 Methods Using Periodicity of the Voiced Speech

Some enhancement methods capitalize on the observation that waveforms of voiced sounds are periodic with a period that corresponds to the fundamental frequency. Two typical methods are presented as follows:

i Adaptive comb filtering

Normally the energy of the speech is concentrated in narrow bands of frequency whereas the noise has energy spread across the whole spectrum. Comb filtering passes the harmonics of speech but rejects the frequency content between those harmonics, which are mainly from the noise. Due to the fact that pitch information (the perceived fundamental frequency) varies from speaker to speaker or even within speech from a
single speaker, the comb filter is required to be adaptive, matching the changing pitch of the input waveform.

**Adaptive Noise Cancellation (ANC)**

It is useful for processing speech that results from what is traditionally referred to as the “two microphone” problem. It assumes that speech is corrupted by uncorrelated noise and that a reference noise source is available, which is correlated with the noise to be reduced. Adaptive filtering techniques are employed to reduce noise including least mean square (LMS) algorithm and steepest descent algorithm that requires no a-priori knowledge of the statistics of the noise. In those cases where the reference noise channel is not available. The drawback of these methods is that the intelligibility of the processed speech is reduced over a wide range of SNR ratios when white noise or a competing talker is involved (Sayed Hadei and Lotfizad 2010).

### Methods Based on Subspace

The idea behind subspace methods is to project the noisy signal onto two subspaces: the signal plus noise subspace, or simply signal subspace (since the signal dominates this subspace), and the noise subspace. The noise subspace contains signals from the noise process only; hence an estimate of the clean signal can be made by removing or nulling the components of the signal in the noise subspace and retaining only the components of the signal in the signal subspace will lead to speech enhancement. The decomposition of the space into two subspaces can be done using singular value decomposition (Dendrinos et al 1991), quotient singular value decomposition (Jensen et al 1995) and eigen value decomposition (Ephraim and Van Trees 1995). Mittal and Phamdo (2000) focused on providing proper noise shaping for colored noise without pre-
whitening. To do that, they first classified the noisy speech frames into speech-dominated and noise-dominated frames and used a different Karhunen–Loeve transform (KLT) matrix for these frames to construct the estimator.

Rezayee and Gazor (2001) proposed an adaptive KLT, was a tracking based algorithm for enhancement of speech degraded by colored additive interference. This algorithm decomposes noisy speech into its components along the axes of a KLT based vector space of clean speech. Enhancement is performed by modifying each KLT component due to its noise and clean speech energies. Hu and Loizou (2003) proposed joint diagonalization of the covariance matrices of the clean signal and noise process.

The major difference between the spectral subtraction approach and the signal subspace approach is in the transform used to decompose the vector space of the noisy signal. The signal subspace method didn't work well for low energy speech units in continuous speech such as consonants and is not suitable for speech corrupted by non-stationary noises like real environmental noises (Jensen and Heusdens 2007).

2.3.3 Methods using the Short Time Spectral Amplitude (STSA) of Speech

Spectral subtraction and Wiener filtering are two typical methods that fall into this category of speech enhancement and a detailed discussion on each in the following sections. Both of them require noise estimation from non-speech segments. These methods are motivated from the fact that in terms of speech intelligibility and quality, short time spectral amplitude (STSA) information is more important than the phase information. The spectra of the background noise and of the noisy speech are estimated in
order to reconstruct the enhanced speech. Ephraim Malah filtering (EMF) capitalizes on the minimum mean square (MMSE) STSA estimator to provide enhanced speech, with a significant reduction of the noise (Ephraim and Malah 1984).

2.3.3.1 Spectral Subtraction (SS) Methods

Spectral subtractive algorithm is historically one of the first algorithms proposed for noise reduction (Boll 1979). It is non-parametric method which requires only an estimate of the noise spectrum. In case of single channel, the estimation is done in the periods where the speaker is silent, which is referred as the pause periods. The basic principle of the spectral subtraction method is to subtract the magnitude spectrum of noise from that of the noisy speech. The subtraction process needs to be done carefully to avoid any speech distortion. If too much is subtracted, then some speech information might be removed, whereas if too little is subtracted, then much of the interfering noise remains.

\[ |\hat{S}(\omega)|^2 = |Y(\omega)|^2 - |\hat{N}(\omega)|^2 \]  \hspace{1cm} (2.1)

where \( \hat{S}(\omega) \) is the spectrum of the enhanced signal, \( \hat{N}(\omega) \) is the estimated noise spectrum and \( Y(\omega) \) is the noise corrupted speech spectrum. The noise is assumed to be uncorrelated and additive to the speech signal. The conventional power spectral subtraction method substantially reduces the noise levels in the noisy speech. However, it also introduces an annoying distortion in the speech signal called musical or residual noise. In SS phase information of the corrupted speech is used for the reconstruction of the enhanced speech. This is one of the short comings of the spectral subtraction which produces a roughness in the quality of the synthesized speech. Estimating the phase of the clean speech is a difficult task and greatly
increases the complexity of the enhancement algorithm. SS is suitable for stationary noises or very slowly varying noises.

In most spectral subtraction algorithms it is assumed that an estimate of the noise spectrum is available. Such an estimate can be obtained using a VAD or a noise estimation algorithm which will increase the system complexity. Sovka et al (1996) proposed an alternative and computationally simple approach that continually estimates the noise spectrum without requiring a VAD algorithm termed extended spectral subtraction (ESS). ESS method is based on the combination of adaptive wiener filtering and spectral subtraction principles. Sim et al (1998) developed a MMSE spectral subtraction method for optimally selecting the subtractive parameters in the mean squared error sense and it was based on parametric formulation of the generalized spectral subtraction algorithm. Virag (1999) proposed a technique based on the masking properties of the human auditory system, the masking thresholds are calculated by first performing power spectral subtraction. Therefore, when the threshold is high, the subtraction parameters are kept minimal, thereby reducing speech distortion. When the masking threshold is low, the residual noise is not masked and the subtraction parameters are maximized. Teddy Surya Gunawan et al (2010) suggested a method based on a short term temporal masking threshold to noise ratio in which a novel functional model for forward masking is incorporated into a speech enhancement framework based on speech boosting.

Kamath and Loizou (2002) suggested multiband spectral subtraction (MBSS) to reduce the distortions to a large extent while maintaining a high level of speech quality. In multi band spectral subtraction, frequency dependent subtraction factor is calculated for each frequency component of the spectra. It provides a definite improvement over
the conventional power spectral subtraction method. The added computational complexity of the algorithm is minimal. The main limitation to SS method lies in the degradation of the intelligibility of the enhanced speech, especially at low SNR levels.

2.3.3.2 Iterative Wiener Filtering

The Wiener filter is a popular adaptive technique that has been used in many enhancement methods. The basic principle of the Wiener filter is to estimate an optimal filter from the noisy input speech by minimizing the Mean Square Error (MSE) between the desired signal and the estimated signal (Haykin 2000). The Wiener filter can be given in the frequency domain by:

$$H(\omega) = \left( \frac{S_s(\omega)}{S_s(\omega) + S_n(\omega)} \right)$$

(2.2)

where $S_s(\omega)$ is the power spectral density (PSD) of the speech and $S_n(\omega)$ is the PSD of the noise spectrum calculated during periods of non-speech activity. From equation (2.2), it is obvious that a-priori knowledge of the speech and noise power spectra is necessary. The speech power spectrum is estimated using the estimated speech model parameters. In the single channel case, noise statistics normally have to be estimated during silence frames and a-priori knowledge of the speech statistics have to be obtained via an iterative estimation process. Lim and Oppenheim (1979) proposed the estimation of speech parameters in an all pole model corrupted by white Gaussian noise and later it is generalized for the colored noise case. The enhanced speech signal spectrum is estimated in $(i+1)^{th}$ iteration can be represented as,
\[ \hat{S}_{i+1}(\omega) = H_i(\omega) Y(\omega) \]  \hspace{1cm} (2.3)

where \( \hat{S}_{i+1}(\omega) \) is the estimated speech spectrum in \((i+1)\)th iteration, \(Y(\omega)\) represents the noisy speech spectrum and \(H_i(\omega)\) denotes the Wiener filter obtained in the \(i\)th iteration,

\[ H_i(\omega) = \left( \frac{\hat{S}_s(\omega)}{\hat{S}_s(\omega) + \hat{S}_n(\omega)} \right) \]  \hspace{1cm} (2.4)

where \( \hat{S}_s(\omega) \) represents the estimated power spectrum of the speech for each iteration and \( \hat{S}_n(\omega) \) is the estimate of the noise power spectrum (Philipos 2007).

### 2.3.3.3 Ephraim-Malah Filtering (EMF)

Unlike spectral subtraction, where the spectral magnitude is averaged regardless of whether the frame contains speech or noise or both, the Ephraim Malah minimum mean-square error (MMSE) short-time spectral amplitude estimator (STSA) applies nonlinear smoothing mostly when the analysis frame is predominantly noise. Interestingly, Ephraim and Malah found that this optimal MMSE STSA estimator provides colorless residual noise rather than the musical tone artifacts often incurred by spectral subtraction methods while drastically reducing the noise without significant distortion of speech contents. Ephraim and Malah (1984) proposed optimal MMSE STSA for speech enhancement. This spectral amplitude estimator is obtained optimally, in the maximum likelihood sense, via modeling speech and noise spectral components as statistically independent Gaussian random variables. The background noise in the real world is most often non-stationary and non-ergodic as well.
Ephraim and Malah (1985) developed a log spectral amplitudes estimator which gives somewhat better enhancement results. The overall result is an algorithm which adaptively tracks and adjusts estimates of both noise and signal amplitudes and uses these estimates to adjust the degree of enhancement, which has significant impact on reducing artifacts often present in spectral subtraction and Wiener filtering. Furthermore in dealing with non-stationary noise, the Ephraim and Malah filter (EMF) possesses significant advantages over other classical spectral based methods. It has the theoretical advantage of finding the optimal solution according to the short time spectral amplitude changes and the implementation is quite straightforward despite the complex mathematical basis. As with SS and Wiener filtering, the basic problem encountered in applying EMF is again how to estimate the noise statistics to the most precise extent possibly. In order to circumvent all the limitations of existing single channel enhancement methods, wavelet based denoising is used for speech signals which will be discussed in the subsequent chapters.

2.4 WAVELET BASED METHODS

So far all the speech enhancement techniques discussed are based on the spectral information obtained through the short time Fourier transform analysis of the target signal. These are all frequency based methods intending to preserve the slow varying short time spectral characteristics of the speech such as the low frequency harmonics of vowels, which is still not enough to maintain speech quality after the processing. Over the last decade much work has been done in applying time frequency transforms to the problem of signal representation and classification. Signals possessing non-stationary character are not well suited for detection and classification by traditional Fourier methods. It has been shown that
wavelets can approximate time varying non-stationary signals in a better way than the Fourier transform representing the signal on both time and frequency domains. Hence they can easily detect local features in a signal. Furthermore, wavelet decomposition allows analyzing a signal at different resolution levels.

As a powerful time-frequency tool, the wavelet transform has established a reputation as a tool for signal analysis: having high frequency resolution (and low time resolution) for the low frequency content of the signal while having low frequency resolution (and high time resolution) for the high frequency content of the signal. The effective noise suppression may be achieved by transforming the noisy signal into the wavelet domain and preserving only the local maxima of the transform. Alternatively, a reconstruction that uses only the large magnitude coefficients has been shown to approximate well the uncorrupted signal. 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 for the extended range of noise types and SNR levels.

2.4.1 Introduction to Wavelet Transform

Wavelet transform has been intensively used in various fields of signal processing areas, especially in signal and image processing. It has the advantage of using variable size time-windows for different frequency bands. This results in a high frequency-resolution (and low time-resolution) in low bands and low frequency-resolution in high bands (Donoho and Johnstone 1995). Consequently, wavelet transform is a powerful tool for modeling non-stationary signals like speech that exhibit slow temporal variations in low frequency and abrupt temporal changes in high frequency. Generally the use of the sub band processing can result in a better
performance when one is restricted to use only one (noisy) signal (as in single-microphone speech enhancement). Therefore, wavelet transform can provide an appropriate model for speech signal denoising applications.

The word “wavelet” literally means “a small wave”. A wavelet is a function that has finite energy and zero mean. It is a powerful tool for the analysis of transient, non-stationary characteristics such as drift, trends, abrupt changes, beginning and ends of events, breakdown points, and discontinuities in higher derivatives and self-similarity. Many kinds of wavelets are available like Haar, Morlet and Daubechies. They look different and have different properties like orthogonal, bi-orthogonal and normalized. The wavelet representation $W(a,b)$ of a time domain energy signal $f(t)$ can be obtained using the continuous wavelet transform (CWT) in terms of the wavelet function as.

$$W(a,b) = \int_{-\infty}^{\infty} f(t) \psi_{a,b}^*(t) dt$$

(3.1)

where $\psi_{a,b}(t)$ are dilated and shifted versions of the basis function $\psi(t)$ obtained by,

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}} \psi \left( \frac{t - b}{a} \right)$$

(3.2)

and * denotes the complex conjugate. In this representation, $a$ is known as the dilation parameter and $b$ is the time shift.

2.4.2 Implementation of Discrete Wavelet Transform

Similar to the CWT, the DWT represents an energy signal in terms of dilated and shifted versions of the wavelet function. The difference
is that with the DWT, the signal is represented in terms of wavelets for discrete values of the dilation parameter \( a = 2^m \) for \( m = 0,1,2,3\ldots N \) and the time shift \( \tau \). DWT allows the implementation of a fast dyadic wavelet transform and its inverse with filter banks. High-pass filter removes the low frequency components of the signal and the corresponding filter parameters become the detailing part of the wavelet coefficients. Low pass filter removes the high frequency components of the signal and the corresponding filter parameters become the smoothing part of the wavelet coefficients. Partly due to the efficient implementation and auditory and visual cortex like properties of dyadic wavelets, a large part of wavelet theory has involved finding dyadic wavelet bases that are orthogonal and useful in a variety of applications (Quatieri 2001).

The main advantage of DWT over CWT is its ease of implementation; the DWT and IDWT can be implemented using discrete time finite impulse response (FIR) filters, which makes them convenient for implementation in microprocessors, digital signal processors, field programmable gate arrays and application specific integrated circuits. One more drawback of the CWT is that the representation of the signal is often redundant. Unlike the continuous wavelet transform, which can operate on every scale, the discrete wavelet transform (DWT) chooses a subset of scales and positions to calculate.

2.4.3 Selection of wavelet

Choosing a wavelet that has compact support in both time and frequency in addition to significant number of vanishing moments is essential for an algorithm. Several criteria can be used in selecting an optimal wavelet function. The objective is to minimize reconstructed error variance and maximize signal to noise ratio (SNR). Optimum wavelets can
be selected based on the energy conservation properties in the approximation part of the coefficients. Wavelets with more vanishing moments should be selected as it provides better reconstruction quality and introduce less distortion into processed speech and concentrate more signal energy in few coefficients. Computational complexity of DWT increases with the number of vanishing moments and hence for real time applications it cannot be suggested with high number of vanishing moments (Mahesh S. Chavan et al 2010).

Unlike Fourier and discrete cosine transforms, which have only one basis function, the discrete wavelet transform has many possible choices of basis function. These basis functions have significantly different properties. For example, a wavelet basis function can be orthogonal, bi-orthogonal, symmetric or non symmetric. While this variety allows the system designer more flexibility, it is important to decide which wavelet is the best for a specific application. More than a dozen different wavelets are currently known in the literature and recognized in the Matlab libraries. The different wavelets have various advantages and disadvantages depending on the targeted application. For example the complex wavelets family, the Gaussian wavelet, the Morlet wavelet and the Mexican hat wavelet cannot be used because a reconstruction filter does not exist for them, while the Meyer wavelet is disregarded because it cannot be implemented with a discrete FIR filter. This leaves only the orthogonal, compactly supported wavelets, namely Haar, Daubechies, Coiflets and Symlets, as candidates for real time denoising systems. Ingrid Daubechies (1998) invented compactly supported orthonormal wavelets named Daubechies (Db) wavelet. It could be implemented using short digital filters, thus making discrete wavelet analysis practicable. The names of the Daubechies family wavelets are written DbN, where N is the order and db the "surname" of the wavelet.
(Soman and Ramachandran 2006). Hence Db4 wavelet is selected for this work.

2.4.4 De-noising using discrete wavelet thresholding

The wavelet de-noising technique is a non-linear algorithm, called as thresholding. This technique is simple and efficient. However it relies heavily on the choice of the threshold, which in its turn depends on the noise distribution. In the wavelet thresholding first a threshold is selected and the components of wavelet transform of the noisy signal are processed in order to improve signal to noise ratio (SNR). Signals can be de-noised in the wavelet domain by implementing the following three steps:

i. Obtain the discrete wavelet domain representation of the signal. The particular wavelet and number of levels to be used are parameters that need to be specifically determined according to the circumstances and the conditions of the de-noising problem.

ii. Apply thresholding to the detail components of the signal obtained in step 1. The thresholding strategy and the threshold values are also parameters that need to be specifically determined according to the circumstances and the conditions of the de-noising problem.

iii. Reconstruct the de-noised signal by applying the IDWT to the approximation and the thresholded detail components.

Figure 2.2 shows a block diagram of a two level DWT denoising system. As can be seen in the figure, only the detail components in the signal are thresholded; the approximation components pass through without
thresholding. In above figure Similarly gA and gs represents analysis and synthesis HPF respectively.

![Block diagram of a two level DWT de-noising system](image)

**Figure 2.2 Block diagram of a two level DWT de-noising system**

2.5 Wavelet denoising systems and their variations

Wavelet denoising is commonly used for speech enhancement because of the simplicity of its implementation. Mallat (1989) suggested a method of wavelet decomposition according to the principle of multi resolution. The method of wavelet threshold denoising is based on the principle of the multi resolution analysis. The discrete detail coefficients and the discrete approximation coefficients can be obtained by a multi-level wavelet decompose (Taswell 2000).

Mallat and Hwang (1992) have shown that effective noise suppression may be achieved by transforming the noisy signal into the wavelet domain and preserving only the local maxima of the transform. Alternatively, a reconstruction that uses only the large-magnitude
coefficients has been shown to approximate well the uncorrupted signal. In other words, noise suppression is achieved by thresholding the wavelet transform of the contaminated signal.

Donoho and Johnstone (1992) firstly proposed a universal threshold for removing the additive white Gaussian noise. Later Donoho (1995) introduced wavelet shrinkage as a powerful tool in denoising signals corrupted by additive white noise. Wavelet shrinkage employs non-linear thresholding in the wavelet domain and has proved broad asymptotically near-optimal properties for a wide class of signals corrupted by additive noise.

Johnston and Silverman (1997) proposed a level-dependent threshold to remove colored noise. The wavelet transform performs a correlation analysis; therefore, the output is expected to be maximal when the input signal most resembles the wavelet. Other variations of wavelet de-noising include multi-wavelets, which have been suggested as an improvement over conventional (scalar) wavelets (Downie and Silverman 1998). It delivers a better denoising performance at the expense of a more complex system and more latency. Yao and Zhang (2001) stated the Bionic Wavelet Transform (BWT) method for wavelet decomposition of speech signals. “Bionic” means that it is rooted in an active biological mechanism. The BWT was originally designed for applications in speech coding, with particular emphasis on the possibility of using it for encoding of cochlear implant signals.

Bahoura and Rouat (2001) proposed the method of level dependent wavelet thresholding. They utilized the Teager energy operator to improve the discriminability for determining whether a speech segment is
speech dominated or noise dominated. This method can greatly reduce noise and also threshold adaptation in time domain can prevent speech from quality deterioration during thresholding process. Jansen and Bultheel (2002) suggested the universal threshold that should be multiplied by an adjustment factor to obtain minimum MSE. Motivated by the communicative connection between the speech production system and the auditory system, Michael et al (2007) developed the Bionic Wavelet Transform (BWT) in combination with the existing wavelet denoising techniques to construct a new adaptive wavelet thresholding method for speech enhancement. It leads to better separation of signal and noise components within the coefficients and therefore better enhancement results.

Lu and Wang (2003) proposed a method in which the background noise can be almost removed by adjusting the wavelet coefficient threshold (WCT) according to the value of SNR. After that, the adaptive wavelet-based methods in speech enhancement are widely presented. They utilize adequately WCT to improve the performance of speech enhancement. For noisy speech, energies of unvoiced segments are comparable to those of noise. Most of the wavelet thresholding techniques not only suppress additional noise but also suppress some speech components like unvoiced ones are suppressed. Consequently, the detection of the voiced/unvoiced segments of the speech signals is a main problem in wavelet based methods. Sheikhzadeh and Abutalebi (2001) suggested an improved scheme, which categorizes speech into either a voiced frame or an unvoiced frame. They increased WCT for high bands in a voiced frame and decreased the threshold values for high bands in an unvoiced frame. As a result, both low frequency components of voiced segments and high frequency components of unvoiced segments are reserved by the soft thresholding algorithm.
Hu and Loizou (2003) suggested a new perceptually motivated approach for the enhancement of speech corrupted by colored noise. The approach takes into account the frequency masking properties of the human auditory system and reduces the perceptual effect of the residual noise. Further Hu and Loizou (2004) proposed a method using the low-variance short-time spectral amplitude spectral estimators to reduce the effect of musical residual noise. It is based on wavelet thresholding of the multitaper spectra for enhancing speech. However, the conventional methods generate the musical residual noise while thresholding the background noise. The unvoiced components of speech are often eliminated in this method.

Yasser Ghanbari and Mohammad Reza Karami-Mollaei (2006) suggested a new speech enhancement system using the wavelet thresholding algorithm. The basic wavelet thresholding algorithm has some defects including the assumption of white Gaussian noise (WGN), malfunction in unvoiced segments and bad auditory quality. In this an adaptive node dependent wavelet thresholding and modified thresholding functions are introduced to improve the speech enhancement performance as well as the automatic speech recognition (ASR) accuracy. Khaled Daqrouq et al (2010) has investigated the utilization of wavelet filters via multistage convolution by reverse biorthogonal wavelets in high and low pass band frequency parts of speech signal. Speech signal is decomposed into two pass bands of frequency and then the noise is removed in each band individually in different stages via wavelet filters.

Kun-Ching Wang (2010) proposed a novel algorithm of wavelet coefficient threshold (WCT) based on time-frequency adaptation. In addition, an unvoiced speech enhancement algorithm is also integrated into the system to improve the intelligibility of speech. The wavelet coefficient threshold (WCT) of each sub band is first temporally adjusted according to
the value of a posterior signal to noise ratio (SNR). Kadam et al (2011) proposed the generalized perceptual wavelet denoising method to reduce the residual noise and improve the quality of speech, which advantageously exploits the wavelet multirate signal representation to preserve the critical transient information. The wavelet coefficients are used to calculate the Bark spreading energy and temporal spreading energy, from which a time-frequency masking threshold is deduced to adaptively adjust the subtraction parameters.

2.6 CONCLUSION

In this chapter, the basics related to speech production model and the background work related to single channel speech enhancement systems like spectral subtraction, subspace, Wiener filtering, Ephraim and Malah filtering and its limitations were discussed. In addition the need for wavelet based speech signal analysis, de-noising procedure and various approaches in wavelet domain are presented in this chapter.