SPEECH ENHANCEMENT UTILIZING PHASE CONTINUITY BETWEEN CONSECUTIVE ANALYSIS WINDOWS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ERDAL MEHMETCİK

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

SEPTEMBER 2011

Approval of the thesis:

SPEECH ENHANCEMENT UTILIZING PHASE CONTINUITY BETWEEN CONSECUTIVE ANALYSIS WINDOWS

submitted by **ERDAL MEHMETCİK** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ________________

Prof. Dr. İsmet Erkmen
Head of Department, **Electrical and Electronics Engineering** ________________

Assoc. Prof. Dr. Tolga Çiloğlu
Supervisor, **Electrical and Electronics Engineering Dept., METU** ________________

Assoc. Prof. Dr. Çağatay Candan
Co-supervisor, **Electrical and Electronics Engineering Dept., METU** ________________

Examining Committee Members:

Prof. Dr. Mübeccel Demirekler
Electrical and Electronics Engineering Dept., METU ________________

Assoc. Prof. Dr. Tolga Çiloğlu
Electrical and Electronics Engineering Dept., METU ________________

Assoc. Prof. Dr. Çağatay Candan
Electrical and Electronics Engineering Dept., METU ________________

Asst. Prof. Dr. Afşar Saranlı
Electrical and Electronics Engineering Dept., METU ________________

Dr. Özgül Salor
TÜBİTAK UZAY ________________

Date: ________________

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: ERDAL MEHMETCİK

Signature :

ABSTRACT

SPEECH ENHANCEMENT UTILIZING PHASE CONTINUITY BETWEEN CONSECUTIVE ANALYSIS WINDOWS

Mehmetcik, Erdal
M.Sc., Department of Electrical and Electronics Engineering
Supervisor : Assoc. Prof. Dr. Tolga Çiloğlu
Co-Supervisor : Assoc. Prof. Dr. Çağatay Candan

September 2011, 75 pages

It is commonly accepted that the induced noise on DFT phase spectrum has a negligible effect on speech intelligibility for short durations of analysis windows, as the early intelligibility studies pointed out. This fact is confirmed by recent intelligibility studies as well. Based on this phenomenon, classical speech enhancement algorithms do not modify DFT phase spectrum and only make changes in the DFT magnitude spectrum. However, in recent studies it is also indicated that these classical speech enhancement algorithms are not capable of improving the intelligibility scores of noise degraded speech signals. In other words, the contained information in a noise degraded signal cannot be increased by classical enhancement methods. Instead the ease of listening, i.e. quality, can be improved. Hence additional effort can be made to increase the amount of quality improvement using both DFT magnitude and DFT phase. Therefore if the performances of the classical methods are to be improved in terms of speech quality, the effect of DFT phase on speech quality needs to be studied.

In this work, the contribution of DFT phase on speech quality is investigated through some simulations using an objective quality assessment criterion. It is concluded from these simulations that, the phase spectrum has a significant effect on speech quality for short durations of analysis windows. Furthermore, phase values of low frequency components are found to have the largest contribution to this quality improvement. Under the motivation of these results, a new enhancement method is proposed which modifies the phase of certain low frequency components as well as the magnitude spectrum. The proposed algorithm is implemented in MATLAB© environment. The results indicate that the proposed system improves the performance of the classical methods in terms of speech quality.

Keywords: speech enhancement, phase estimation, time frequency analysis

ÖZ

ARDIŞIK ANALİZ PENCERELERİ ARASINDAKİ FAZ SÜREKLİLİĞİNİ
SAĞLAYARAK KONUŞMA İYİLEŞTİRME

Mehmetcik, Erdal
Yüksek Lisans, Elektrik ve Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Tolga Çiloğlu
Ortak Tez Yöneticisi : Doç. Dr. Çağatay Candan

Eylül 2011, 75 sayfa

DFT faz spektrumunun konuşma anlaşılabilirliği üzerinde ihmal edilebilir bir katkısının olduğu bilinmektedir. Bu olgu yapılan yakın zamanda yapılan araştırmalarda da doğrulanmıştır. Klasik konuşma iyileştirme algoritmaları, bu bulgulara dayanarak sadece DFT genlik spektrumunu değiştirmekte ve faz spektrumunun gürültülü halini kullanmaktadır. Ancak, yakın zamanda yapılan araştırmalar klasik yöntemlerin anlaşılabilirliği arttıramadığını vurgulamaktadır. Bu yöntemler dinleme rahatlığını, başka bir deyişle konuşma kalitesini, arttırabilmektedir. Bu bağlamda hem DFT genlik hem de DFT faz spektrumu kullanılarak klasik yöntemlerin performansı konuşma kalitesi açısından arttırılabilir. Bu amaç doğrultusunda faz spektrumunun konuşma kalitesine olan katkısı da incelenmelidir.

Bu tez çalışmasında, faz spektrumunun konuşma kalitesine olan katkısı bazı benzetimler aracılığıyla incelenmiştir. Bu benzetimlerde objektif kalite belirleme kriterleri kullanılmıştır. Bu benzetimlerde faz spektrumunun konuşma kalitesine önemli bir katkı sağlayabileceği sonucuna varılmıştır. Özellikle düşük frekans bileşenlerinin fazının bu kalite iyileştirmesindeki etkisinin çok daha fazla olduğu görülmüştür. Bu sonuçlardan yola çıkarak, düşük frekans bileşenlerinin fazını düzeltmeye yönelik yeni bir konuşma iyileştirme algoritması önerilmiştir.

Önerilen yöntem bileşenlerin fazını değiştirdiği gibi, genlik değerlerini de klasik yöntemleri kullanarak değiştirmektedir. Önerilen yöntem MATLAB ortamında gerçeklenmiş ve önerilen yöntemin performansının klasik yöntemlere oranla daha yüksek olduğu görülmüştür.

Anahtar Kelimeler: Konuşma iyilestirme, faz kestirimi, zaman-frekans analizi

To my family

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Assoc. Prof. Dr. Tolga Çiloğlu for his guidance, encouragements and insight throughout the research.

I would also like to thank my co-advisor Assoc. Prof. Dr. Çağatay Candan.

Thanks to the examining committee members Prof. Dr. Mübeccel Demirekler, Asst. Prof. Dr. Afşar Saranlı and Dr. Özgül Salor for evaluating my work.

I would like to express my gratitude to Oktay Sipahigil for his suggestions and support.

I am grateful to The Scientific and Technological Research Council of Turkey (TÜBİTAK) for their financial support during my master's studies.

Deepest thanks to my family for their love, trust, understanding, and every kind of support not only throughout my thesis but also throughout my life.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Speech is a highly non-stationary signal. Because of this characteristic, it is truly difficult to accurately analyze and process speech signals. Moreover, when the signal is corrupted by noise (e.g. in the transmission channel), additional problems emerge. For instance, the transmitted signal may become unintelligible or very disturbing for the listener. It is therefore the aim of speech processing researchers to suppress the induced noise, without degrading the speech signal. The problem here arises in 'not degrading the speech' part, as there is a trade-off between the amount of noise suppression and induced distortion [3]. The noise suppression process is commonly referred to as 'speech enhancement'. The problem is further narrowed down in most of the studies by identifying the input signal on which the noise suppression is to be done. For instance, when the only available signal is the degraded speech signal, the procedure is commonly referred as 'single channel speech enhancement'. When there are multiple input signals (from different microphones etc.) the process is named accordingly, for instance 'dual-microphone speech enhancement' etc. In this study, single channel narrowband speech enhancement is the main concern.

The aim of speech enhancement algorithms is to increase the quality and if possible the intelligibility of speech signals. It is important to make the distinction between the concepts of quality and intelligibility. The quality of speech is related to the ease of listening, whereas the intelligibility is related to the perceived information by the listener. Both quality and intelligibility are subjective quantities; hence it is hard to define a measure to evaluate the quality and intelligibility of a signal. There are many performance measures defined for the evaluation of speech quality (e.g. segmental signal to noise ratio (segmental SNR), perceptual evaluation of speech quality (PESQ), weighted spectral slope (WSS) etc.), though many of these measures contradict with each other in certain cases. On the other hand there is no objective measure defined for the evaluation of speech intelligibility. The intelligibility is generally tested through subjective listening tests.

There are many applications of speech enhancement. For instance the hearing aids (designed for the hearing impaired patients) are known to have a high internal noise which disturbs and tires the patient. Generally speech enhancement algorithms are run in real time on the processors of these hearing aids to ease the listening conditions. Another application area is the communication channels (generally telephone channels). Due to the imperfections in the communication system, the speech signal to be transmitted is degraded by noise (internal and/or external). Hence the enhancement algorithms are used in the receiving end as well.

1.1 Scope of thesis

In this work, single channel narrowband (0-4 kHz) speech enhancement algorithms are studied. Classical methods on the subject, mainly focus on the modification of the DFT magnitude spectra of the degraded speech, assuming that the phase spectra has a negligible effect on the intelligibility of speech, for short analysis frames, [4]. However it is known that none of the existing algorithms increase the intelligibility scores, [5], and only the quality scores can be improved. Hence the effect of phase noise on speech quality needs to be studied. In this thesis study, the effect of phase noise on speech quality is investigated through some simulations in MATLAB© environment, employing the PESQ (Perceptual Evaluation of Speech Quality, [6]) quality measure (explained in Chapter-2). It is concluded from these simulations that the phase spectra can be utilized to increase the quality as well. Then in the following chapters, an enhancement algorithm (using the phase spectra together with the magnitude spectra) is proposed. The implementation and performance evaluation of this algorithm are also explained in detail.

1.2 Outline

In Chapter-2, the basic concepts that are used in speech processing literature are presented. Some commonly used analysis and synthesis procedures are also studied in this chapter.

Chapter-3 focuses on the review of the classical methods on speech enhancement.

Chapter-4 illustrates the conducted simulations to investigate the contribution of phase spectra to speech quality. The previous studies on the contribution of the phase spectra to speech intelligibility are briefly explained and then the obtained results and comments on these results are given in this chapter.

In Chapter-5, the proposed speech enhancement system is explained. The block diagram of the system is given and the sub-processes in this system are explained in detail in Chapter-5.3 (Phase estimation) and Chapter-5.4 (Pitch estimation).

The validation of the proposed system is elaborated in Chapter-6.1. The implementation of the proposed system is described in detail in Chapter-6.2. The performance of the proposed system is also tested in this chapter.

Lastly, some concluding remarks are made in Chapter-7 and the suggested future work on the subject is stated.

CHAPTER 2

REVIEW OF FUNDAMENTAL CONCEPTS

In this chapter, some basic concepts that are used in speech enhancement literature will be explained.

2.1 Speech signal

Speech signal has a highly non-stationary nature; i.e. the spectral characteristics change rapidly over time. However over short periods of time (20-40 msec), the signal can be considered as stationary. Generally the analysis is done by using short durations of data frames.

The hearing range of humans typically covers the 20Hz - 20kHz frequency band. This is one of the reasons for the sampling rate selection of audio CDs (44.1 kHz) or digital audio tapes (48 kHz), so as to satisfy the Nyquist sampling criterion [7], (f_s > 2x20 kHz). On the other hand, a much smaller sampling frequency is enough for speech signals to be intelligible. The 0.3-3.4 kHz band allows 97% of all sounds to be understood, as stated in [8]. This frequency band is called the ‘telephone band’ and is used in classical telephony. The sampling rate is fixed to 8 kHz in these applications, hence covering the (0-4 kHz) band. The speech signals sampled at this sampling frequency is called ‘narrowband speech’. Wideband speech on the other hand is defined, in the ITU (International Telecommunication Union) recommendation [9], to cover the 50-7000 Hz range and sampling frequency is set to 16 kHz.

In this study, single channel narrowband speech enhancement methods is studied, hence the signal of interest is band-limited to 0-4 kHz band and sampling frequency is selected as 8 kHz.

2.2 Voiced-unvoiced speech

Sounds within a speech signal can be separated into two different classes, namely ‘voiced’ and ‘unvoiced’ sounds. Voiced parts of speech are generated by the vibrations of the vocal cords and exhibit harmonic characteristics. The frequency of the first harmonic (or the fundamental component) is called the fundamental frequency or the pitch frequency of the sound. Sounds like ‘a’, ‘e’, ‘r’ etc. are voiced sounds.

Unvoiced sounds on the other hand are not driven by the vibrations of the vocal cords and exhibit a noise-like and wideband structure. Sounds like ‘s’, ‘f’ etc. are unvoiced sounds.

The production of speech can be modeled as in Figure-2.1, indicating that the generated speech is either driven by a periodic pulse or by noise. This structure is proposed by Rabiner and Schafer [1].

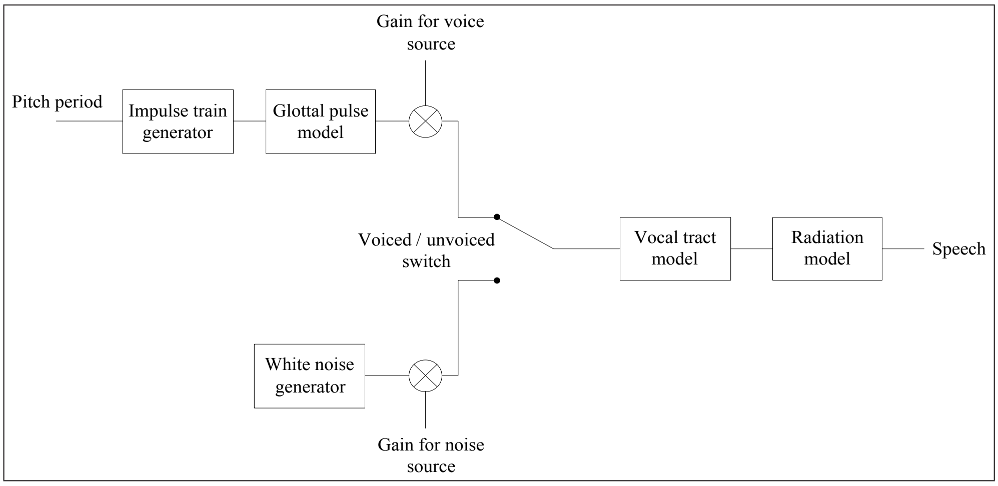


Figure 2.1: Speech production model of Rabiner and Schafer [1]

To elaborate the spectral characteristics of voiced and unvoiced segments of speech, the spectrogram of the word ‘a-s-a’ is given in Figure-2.2.

The oscillatory nature of voiced signals makes it possible to partially model these segments of the speech as a sum of sinusoids with the following form; $\sum_{k=1}^{M} A_k cos\left(2\pi k \frac{f_0}{f_s} n + \phi_k\right)$. This property of the voiced segments of speech signals will be utilized in the context of speech enhancement in the following chapters.

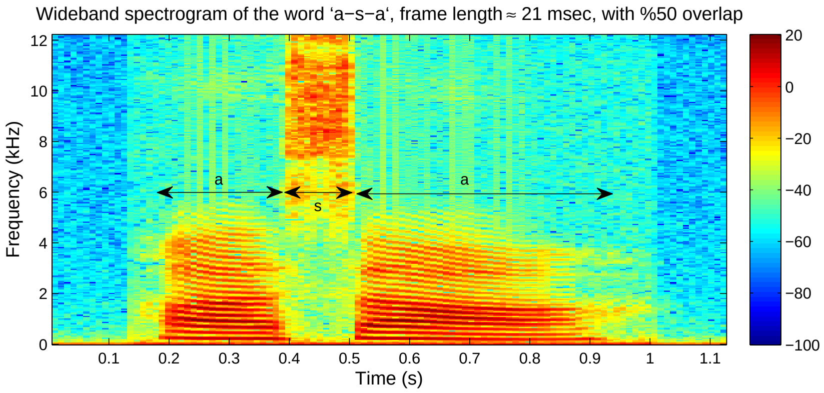


Figure 2.2: Spectrogram of the word ‘asa’, spoken by a female speaker.

The classification of sounds is actually more detailed than voiced-unvoiced discrimination. A detailed classification can be found in [10]. However in the context of this work, it is sufficient to make voiced-unvoiced distinction in a given speech signal.

2.3 Short Time Fourier Transform (STFT)

In this section, a commonly used time-frequency analysis method; namely ‘short time Fourier transform’ (STFT) is briefly explained. A detailed analysis of STFT can be found in many signal processing books, e.g. [7], [11], [12].

2.3.1 STFT analysis

Discrete Fourier Transform is a powerful tool for analyzing LTI (Linear Time Invariant) systems, as the basis functions $\{e^{j\frac{2\pi}{N}kn}\}_{k=0}^{N-1}$ are the characteristic functions of discrete-time LTI systems. However, the signal under consideration must be stationary over the analysis window, in order DFT coefficients to be able to characterize the signal. If this is not the case, the variations of the spectral contents in time will be averaged over the analysis window and these variations in time will not be observed. To overcome this problem, short durations of analysis windows can be used. Assuming that the signal is stationary in a short analysis window, one can have a better characterization of the variations of the spectral content in time. This procedure is called ‘Short Time Fourier Transform’ (STFT). The definition of continuous STFT of a signal $x(t)$, is as follows;

$$X(t,\omega) = \int_{-\infty}^{\infty} x(\tau)w(t-\tau)e^{-j\omega\tau}d\tau \tag{2.1}$$

where $w(t)$ is the window function and is non-zero for $0 < t < T$, T being the window length. The continuous STFT, $X(t,\omega)$, is a complex signal and can be written in polar form as follows;

$$X(t,\omega) = |X(t,\omega)|\, e^{\angle X(t,\omega)} \tag{2.2}$$

In this form, $|X(t,\omega)|$ is called the short time magnitude spectrum and $\angle X(t,\omega)$ is called the short time phase spectrum.

The previous definitions are made for continuous time signal $x(t)$. The discrete-time STFT for the signal $x[n]$, is defined as follows;

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\omega m} \tag{2.3}$$

Notice that the frequency variable ω is a continuous variable in discrete-time STFT definition. In order to be able to compute the transform numerically for an arbitrary signal, the transform must also be sampled in frequency. Hence the discrete STFT (discrete both in time and frequency) is defined as in equation-2.4, which is just the sampled version (in frequency) of equation-2.3 at $\omega = \frac{2\pi k}{N}$.

$$X[n,k] = \sum_{m=-\infty}^{\infty} x[m]w[n-m]e^{-j\frac{2\pi}{N}km} \tag{2.4}$$

where N is the window length in samples and $w[n]$ is the window function.

Although STFT is an efficient way for time-frequency analysis, it has also some (fundamental) drawbacks. For instance, the time resolution and frequency resolution can not be increased at the same time, by changing the analysis window length or sampling frequency. To increase the time resolution, i.e. to observe the changes in time with increased accuracy, one needs to use shorter analysis windows . This causes the frequency resolution of the DFT ($\frac{f_s}{N}$) to decrease and vice-versa. As a result the window length becomes a critical parameter in STFT analysis. Although it might be possible to increase the time resolution by increasing the overlap ratio, doing so is undesirable because of the increased computational load.

2.3.2 STFT synthesis

There are two commonly used synthesis methods for STFT, namely filter bank summation (FBS) and overlap-add (OLA). The methods were first studied in the work of Allen and Rabiner [13]. These methods will be briefly explained in the following sections.

2.3.2.1 Filter-bank summation (FBS) method

The STFT can be viewed as a set of filters. For instance, consider equation-2.4 in the following form;

$$\begin{aligned} X[n,k] &= \sum_{m=\infty}^{\infty}\left(x[m]e^{-j\frac{2\pi}{N}km}\right)w[n-m] \\ &= \left(x[n]e^{-j\omega_k n}\right)*w[n], \qquad \omega_k=\frac{2\pi}{N}k && (2.5)\\ &= \left(x[n]*w[n]e^{j\omega_k n}\right)e^{-j\omega_k n} && (2.6)\end{aligned}$$

The last two equations can be viewed as modulating the signal then low-pass filtering with the window function or simply bandpass filtering the signal with the modulated window functions. With this perspective, the signal can be reconstructed by modulating back each filter output $X[n,k]$ with $e^{j\omega_k n}$ then summing up the results;

$$y_{rec}[n] = \frac{1}{Nw[0]}\sum_{k=0}^{N-1} X[n,k]e^{j\frac{2\pi kn}{N}} \tag{2.7}$$

2.3.2.2 Overlap-add (OLA) method

Another commonly used method for STFT synthesis is the overlap-add procedure. The flowchart of the method can be seen in Figure-2.3. This flowchart is taken from [12] (page 274).

If the STFT $X(n,\omega)$ is sampled in time every R samples and a window length of N samples is

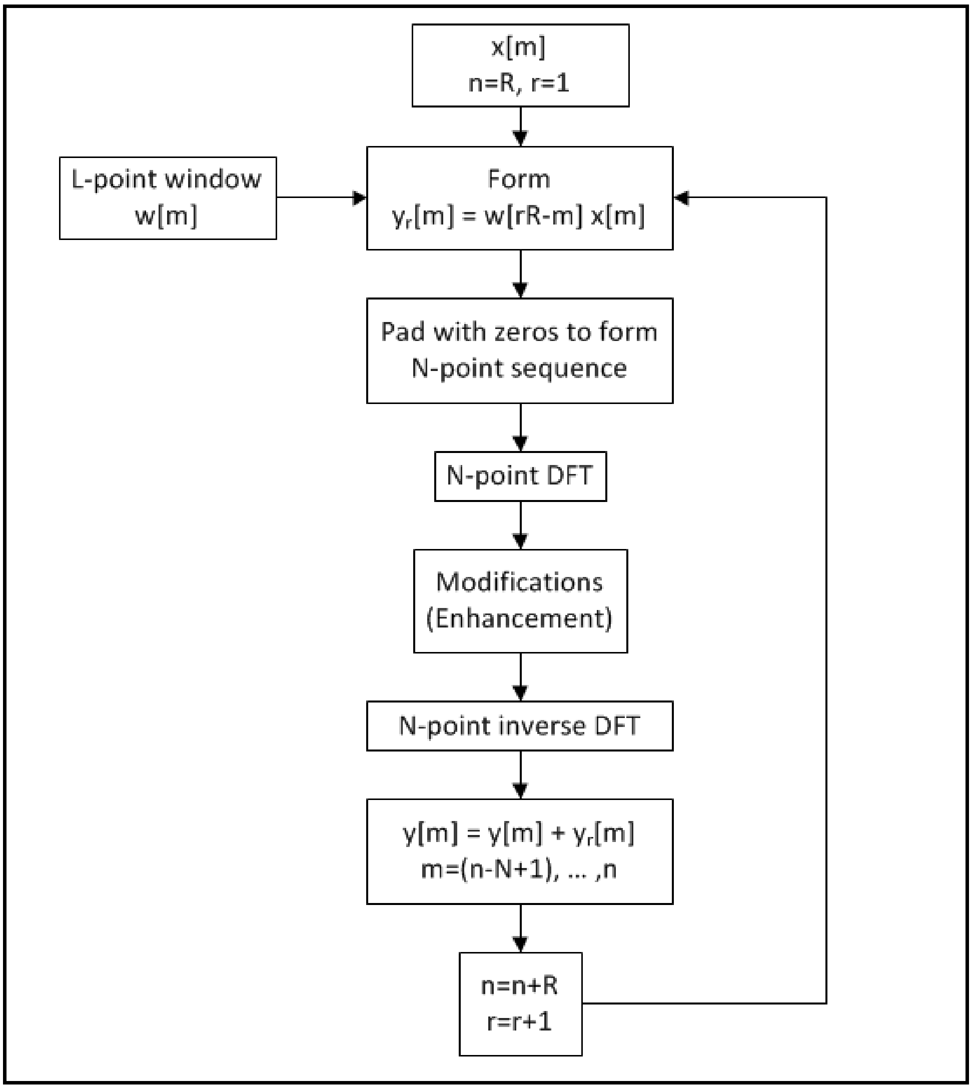


Figure 2.3: Overlap-add method

used, then using OLA method the reconstructed signal ($y_{rec}[n]$) will have the following form;

$$y_{rec} = \sum_{r=-\infty}^{\infty} \left[\frac{1}{N} \sum_{k=0}^{N-1} X[rR, \omega_k] e^{j\omega_k n} \right] \tag{2.8}$$

$$= \sum_{r=-\infty}^{\infty} x[n] w[rR - n] \tag{2.9}$$

$$= \sum_{r=-\infty}^{\infty} y_r[n] = x[n] \sum_{r=-\infty}^{\infty} w[rR - n] \tag{2.10}$$

$$\tag{2.11}$$

As seen in the last equation, $x[n]$ can be perfectly reconstructed when OLA method is used, if the term $\sum_{r=-\infty}^{\infty} w[rR-n]$ is equal to a constant for all n. This can be achieved if the STFT, $X(n,\omega)$, is sampled properly in time. This constraint can be stated as follows;

$$\sum_{r=-\infty}^{\infty} w[rR-n] = \frac{W(0)}{R} \tag{2.12}$$

For Hamming window ($w[n] = 0.54 - 0.46cos(\frac{2\pi n}{N-1})$), which is a common choice in speech enhancement applications, the above condition is satisfied for $R = L/4$. This means that %75 overlap is needed between analysis frames for perfect reconstruction, (when the window function is Hamming). In spite of this perfect reconstruction constraint, a common practice in speech enhancement algorithms is to use Hamming window with %50 overlap. In this case a small distortion is introduced to the reconstructed signal which is quite negligible.

2.4 Performance evaluation

As mentioned earlier, the aim of the speech enhancement algorithms is to increase the ease of listening and if possible, increase the amount of perceived information. These two concepts are known as ‘quality’ and ‘intelligibility’ respectively. Since the aim of enhancement algorithms is to improve these two attributes, one needs a performance measure in order to evaluate how good the proposed algorithm is.

Many distance measures are defined for the evaluation of speech quality over the past three decades. Some of these methods are as follows; segmental SNR, weighted spectral slope (WSS), Bark distortion measures, perceptual evaluation of speech quality (PESQ, [14]) etc. In some cases most of these methods contradict with each other. In this work, PESQ measure will be used to evaluate the performance of the proposed algorithm, as it is the ITU standard for automatic assessment of speech quality and used by phone manufacturers and telecom operators. The details of PESQ measure will be given in the next section.

Although many performance measures are defined (mathematically) for speech quality, there isn’t an objective method for intelligibility assessment. Intelligibility performance of an enhancement algorithm is generally evaluated by listening tests. There are different types of listening tests applied for this purpose; such as nonsense syllable, word or sentence tests. In these tests speech intelligibility is quantified in terms of percentage of words identified correctly by the listener. Such listening tests are somewhat unconventional, as it requires a long process of listening and subjective evaluation of many listeners. Nevertheless the ground truth is taken as the result of such listening tests.

2.4.1 Perceptual Evaluation of Speech Quality (PESQ)

The aforementioned performance measures other than PESQ (segmental SNR, weighted spectral slope (WSS), Bark distortion measures) are suitable for assessing the quality for a limited range of distortions which do not include the commonly encountered distortions in communication channels; for instance packet loss, delay, codec distortions etc. Such kind of deformities in the signal causes the aforementioned methods to produce unreasonably low quality scores. To cope with this problem, in 2000 the International Telecommunication Union (ITU) organized a competition to select a new objective measure which is capable of handling the stated problems. The perceptual evaluation of speech quality (PESQ) measure was selected as the new ITU recommendation P.862 [6]. The method has a high correlation ($\rho > 0.92$) with subjective listening tests as stated in [3].

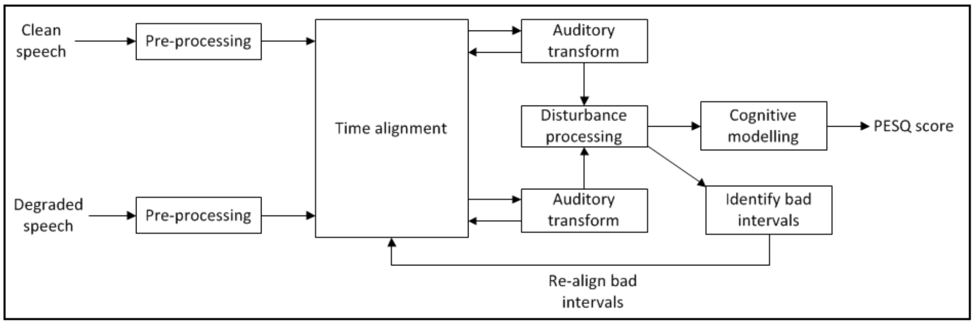


Figure 2.4: Block diagram of PESQ method

The method has the structure given in Figure-2.4. In this block diagram, the ‘time alignment’ and ‘identify bad intervals’ blocks are the novelties of PESQ method with respect to the previous ITU recommendation. The system takes the clean and degraded speech, then it computes the quality score which is between 4.5 and -0.5, with 4.5 corresponding to distortionless and -0.5 corresponding to noise. Although the lower limit is -0.5, the scores below 1 indicate an unacceptable level of distortion. The perceptual meaning of PESQ scores are described in Table-2.1.

Table 2.1: Meaning of PESQ scores

Distortion level	**PESQ score**
Distortionless	4.5
Perceptible but not annoying	3.5
Slightly annoying	3
Annoying	2
Noise-like	<1

The technical details of the method will not be discussed, as it is too comprehensive to be included in this thesis. The PESQ method is explained in detail in [3] and [14].

CHAPTER 3

CLASSICAL SPEECH ENHANCEMENT METHODS

Speech enhancement has been a major research topic for more than four decades. As a result there is a vast literature of different enhancement algorithms. However almost all of these algorithms only modify the DFT magnitude spectra and use the noisy DFT phase in the reconstruction. In this chapter, the main motivation between two of the classical speech enhancement methods is explained and an overview of other enhancement algorithms is presented.

3.1 Spectral subtraction based algorithms

Spectral subtraction based algorithms utilize a simple idea. The algorithms assume additive noise and estimate the noise spectrum; then the clean spectrum estimate is simply obtained by subtracting the noise spectrum from the noisy speech spectrum. There are many versions of spectral subtraction based algorithms, for instance [15], [16]. The details of classical spectral subtraction algorithm can be found in almost any speech enhancement book, e.g. [3]. The method can be summarized as follows;

Let the input speech (noise corrupted) be denoted by $y[n]$, clean speech by $x[n]$ and noise signal (or ‘disturbance’) by $d[n]$. Assuming additive noise;

$$y[n] = x[n] + d[n] \tag{3.1}$$

For practical reasons, consider the DFT coefficients (instead of DTFT) of both sides;

$$Y[k] = X[k] + D[k] \tag{3.2}$$

For single channel speech enhancement algorithms, only $Y[k]$ (DFT of noisy signal) is known.

Hence to obtain $X[k]$, $D[k]$ must somehow be estimated. If equation-(3.2) is written in polar form, we get;

$$|Y[k]|\,e^{j\phi_y[k]} = |X[k]|\,e^{j\phi_x[k]} + |D[k]|\,e^{j\phi_d[k]} \tag{3.3}$$

The phase spectra of noise ($\phi_d[k]$) can be replaced by the phase spectra of noisy speech ($\phi_y[k]$) assuming it has little effect on intelligibility, [4]. As a result, the estimated clean speech will have the following form;

$$\left|\hat{X}[k]\right| = |Y[k]| - \left|\hat{D}[k]\right| \tag{3.4}$$

$$\hat{X}[k] = \left(|Y[k]| - |\hat{D}[k]|\right) e^{j\phi_y[k]} \tag{3.5}$$

$\hat{X}$ and $\hat{D}$ are used to indicate that the signals are 'estimated' versions of the clean speech and the disturbance signal respectively. It is clear that, the noise power estimation will have a crucial effect on the performance of the algorithm.

Substituting the noise phase spectra with the noisy speech phase is known to have negligible effect on speech intelligibility as studied in [4]. Actually this is one of the common assumptions in spectral subtraction based algorithms. However, the effect of this assumption on speech quality may not be negligible, as discussed in Chapter-4.

The problem of estimating $\hat{D}[k]$ can be solved by simply averaging the noise in 'silence' regions, which require a 'Voice Activity Detector' (VAD). The noise power can be estimated by using some other methods as well. An interesting approach is the so called 'minimum statistics' method, described in [17]. In this method, noise power estimate is done for each frequency bin by tracking the minimum power in long observation windows. In the article [17], it is (empirically) stated that these minimum values are proportional with the actual noise levels in the corresponding frequency bins.

Notice that, in equation-(3.5) $|\hat{X}[k]|$ can be negative if the estimated noise power exceeds the noisy signal power at a specific frequency bin, which is of course meaningless and is one of the main problems of spectral subtraction. To prevent such an occasion, usually some extra constraints are imposed. For instance, if the magnitude of the estimated coefficient is negative, it can simply be half wave rectified (negative values are set to zero or set to the minimum value in the spectrum). However, this procedure creates isolated peaks in different frequency bins of the spectrum. The locations of these peaks also change at each frame. As a result a somewhat 'tonal' noise is generated by the process. This type of noise (introduced by the enhancement process) is referred as the ‘musical noise’ in speech enhancement literature and it is the main problem in spectral subtraction based algorithms.

To improve the performance of the algorithms in terms of generation of the ‘musical noise’, a method called ’oversubtraction’ can be employed [18]. In this method, the ‘overestimate’ of the noise power spectrum is subtracted while preventing the the resulting spectral components to fall below a minimum value, called the ‘spectral floor’. The magnitude estimator in this method has the following form;

$$\left|\hat{X}[k]\right|^2 = \begin{cases} |Y[k]|^2 - \alpha \left|\hat{D}[k]\right|^2 & , \quad if \quad |Y[k]|^2 > (\alpha + \beta) \left|\hat{D}[k]\right|^2 \\ \beta \left|\hat{D}[k]\right|^2 & , \quad else \end{cases} \tag{3.6}$$

In this equation α is called the oversubtraction factor ($\alpha \geq 1$), and β is called the spectral floor parameter ($0 < \beta << 1$). The main idea behind the oversubtraction method is to decrease the amplitudes of the peaks in the spectrum that are artificially generated by the spectral subtraction algorithm itself. It is known that, speech processed by this oversubtraction algorithm possesses less amount of musical noise than the original spectral subtraction method as in equation-3.5. However the musical noise is still present. Although the oversubtraction factor suppresses the artificial peaks, it introduces additional distortion to the speech signal. The additional parameters α and β gives the control of making an adjustment between the musical noise suppression and introduced distortion. This adjustment can be optimized in the mean square sense and the values of α and β can be determined accordingly. This optimization is proposed by Sim et. al. [19]. The details of their method will not be discussed in this study. However it is worth mentioning that the method also introduces musical noise. Hence the spectral subtraction based algorithms has certain limitations. Depending on the induced phase noise, the performance of the spectral subtraction based algorithms inevitably decreases.

3.2 MMSE estimator

Several researchers have proposed methods that minimize the mean squared error between the estimated magnitude spectra and the true magnitude spectra. More specifically;

$$e_k = E\left\{\left(\hat{X}_k - X_k\right)^2\right\} \tag{3.7}$$

where $\hat{X}_k$ is the estimated magnitude for the k^{th} frequency bin and X_k is the magnitude of the k^{th} bin for the clean signal. The optimal (in MSE sense) coefficients can be obtained by minimizing the Bayesian MSE error given by;

$$I(\hat{X}_k) = \int\int (X_k - \hat{X}_k)^2 p(\boldsymbol{Y}, X_k) d\boldsymbol{Y} dX_k \tag{3.8}$$

The minimization of the Bayesian MSE with respect to $\hat{X}_k$ yields the optimal MMSE estimator as follows;

$$\hat{X}_k = \int X_k p(X_k|\boldsymbol{Y}) dX_k \tag{3.9}$$

$$= E\{X_k|\boldsymbol{Y}\} \tag{3.10}$$

$$= E\{X_k|Y(\omega_0), Y(\omega_1), ..., Y(\omega_{N-1})\} \tag{3.11}$$

where $\boldsymbol{Y} = [Y(\omega_0), Y(\omega_1), ..., Y(\omega_{N-1})]$ is the vector containing the DFT coefficients of the observed noisy speech.

To calculate equation-3.11, one needs the distribution functions of the DFT coefficients. However it is not possible to measure the density functions of the DFT coefficients by evaluating the histograms from a large amount of data, simply because of the fact that speech is not a stationary and ergodic process. That is to say the statistics of the coefficients will change over time and the time averages will not correspond to the actual density function. At this point Eprahim and Malah proposed a method [16] by assuming that the DFT coefficients have a Gaussian distribution and the coefficients are uncorrelated (since it is Gaussian and uncorrelated then independent as well). The assumptions are justified by utilizing the central limit theorem [20] and by pointing out the fact that as the analysis frame gets longer the correlation between the coefficients decays to zero.

The MMSE magnitude estimator will not be derived here. The details of the derivation can be found in the original paper of Eprahim and Malah [16] and almost in any speech enhancement book (e.g. [3], [8], [21]).

The mean-squared error criterion is actually questioned if it was the right cost function to increase intelligibility, in a recent study by Loizou and Kim [5]. In this study the distortions caused by the additive noise and enhancement process are divided into two groups as amplification and attenuation. It is stated that the effects of these two types of distortions can not be the same aiming the mean squared error criterion, as the MSE metric can not make a distinction between a +5 or -5 difference between the actual and estimated values.

3.3 Other methods

There are many derivatives of the spectral subtraction based methods and MMSE estimators that are briefly explained in the previous sections. There are some other main classes of enhancement algorithms such as Wiener filtering and subspace methods [3]. To apply Wiener filtering, it is assumed that the clean speech can be obtained by a linear filtering operation on noisy speech. Then the ‘optimal’ filter coefficients are determined in the MSE (Mean Squared Error) sense. Wiener filtering is a classical signal processing subject, the details of which can be found in many statistical signal processing books (e.g. [22], [23]). Wiener filtering can be applied to speech enhancement by imposing different constraints on the signal as well, as explained in [3].

In literature there are some methods that modify the phase spectra as well. One of those methods is presented in [24]. In this method the phase of the noisy signal is intentionally distorted by adding a real number to first half of the DFT coefficients and subtracting the same real number from the second half. Then using the phase of the resulting coefficient set and the magnitude of the original coefficient set, the inverse STFT is computed. By adding and subtracting a real number, the phases of weak components are shifted almost 180 degrees out of phase to its conjugate counterparts. These components cancel each other in the reconstruction process.With this property, the method seems like an efficient way of linear filtering, where the linear filter is designed so that the suppression of the components is inversely proportional to the strength of the corresponding component.

CHAPTER 4

CONTRIBUTION OF PHASE INFORMATION TO SPEECH QUALITY

4.1 Introduction

In this chapter the importance of phase information on speech quality is studied. Previous work on the subject is briefly explained and the results of the conducted experiments are given.

4.2 Previous work

As mentioned in the previous chapters, classical speech enhancement methods (e.g. [15], [16], [25] and many others) rely on the assumption that the human perception is less sensitive to phase distortions, citing the study of Wang and Lim [4]. In their article [4], Wang and Lim presented the results of their intelligibility tests. These tests are conducted to measure the contribution of DFT phase and magnitude to speech intelligibility. They used two analysis blocks in parallel to estimate the phase and magnitude spectra by using both clean and noisy speech. By altering the amount of induced noise for phase and magnitude estimation seperately, the structure is capable of controlling the amount of phase and magnitude distortions independently. Using this structure and carrying out listening tests, they conclude that [4];

”…It is unwarranted to make an effort to more accurately estimate the phase from the noisy speech in the context of speech enhancement if the estimate is used to reconstruct a signal by combining it with an independently estimated magnitude or to reconstruct the signal using the phase-only signal reconstruction algorithm”, [4].

A recent study, by Paliwal and Alsteris [26], confirms the results of Wang and Lim [4] for short durations of analysis windows and also points out that phase information can be important when the analysis window is long (in the order of 500 msec), in terms of intelligibility.

These studies were done in the context of speech intelligibility. However, as stated in [5] current enhancement algorithms are not capable of improving the intelligibility scores of degraded speech signals. What they are capable of is to improve the ease of listening, i.e. quality; hence the contribution of phase information to speech quality needs to be investigated. For that purpose, some simulations are carried out in MATLAB©. The details of these simulations are given in the next section.

4.2.1 Simulations on the effect of phase information on speech quality

The simulations are carried out using the structure given in Figure-4.1.

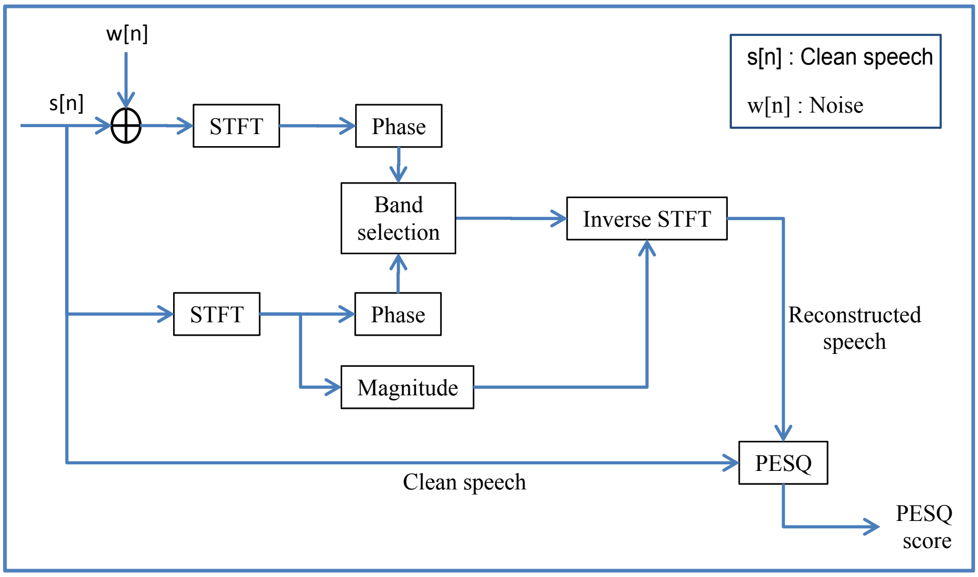


Figure 4.1: Implemented test structure which is designed to investigate the contribution of the phase of a specific frequency band

As seen in Figure-4.1 a hybrid signal is generated using the magnitude spectra of the clean speech and phase spectra of both noisy and clean speech. The clean phase is used for a particular frequency band using the band selection block. The reason for doing this, is to observe the quality improvement when a specific part of the phase spectra is estimated perfectly. The band selection block in Figure-4.1 is elaborated in Figure-4.2. As the figure indicates, the task of this block is to arrange the frequency band, in which the clean phase will be used and these bands are shown in Figure-4.3.

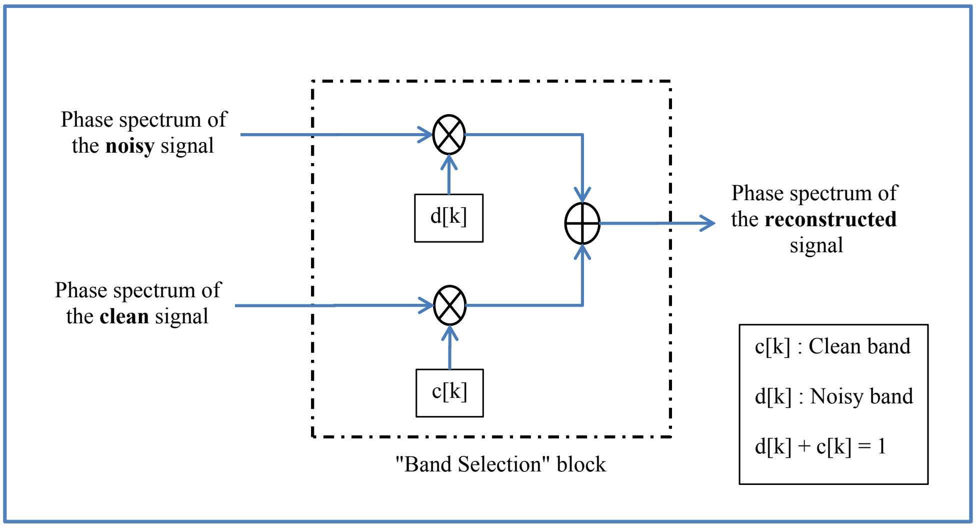

Phase spectrum of the **noisy** signal
Phase spectrum of the **clean** signal
d[k]
c[k]
Phase spectrum of the **reconstructed** signal
c[k] : Clean band
d[k] : Noisy band
d[k] + c[k] = 1
"Band Selection" block

Figure 4.2: Details of the band selection block

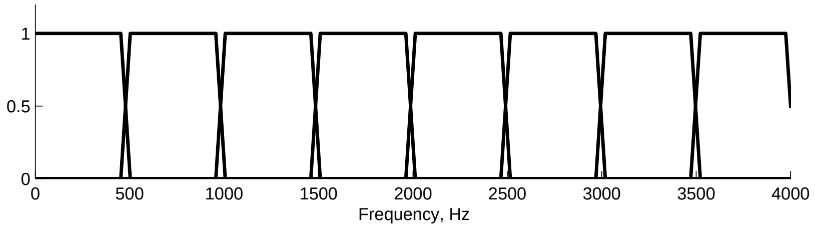


Figure 4.3: Frequency bands over which the clean phase is used

The sampling frequency for the narrowband speech is 8 kHz, hence the frequency spectrum is limited to 4 kHz for these tests. As seen in Figure-4.3, the 4kHz band is divided into 8 non-overlapping subbands over which the clean phase is used.

While generating these hybrid signals ‘NOIZEUS’ speech database is used. The details of this database are given in [27] and [28]. The database consists of 30 IEEE sentences ([29]) recorded under 8 different colored noises at 4 different SNR levels. The sentences are pro-

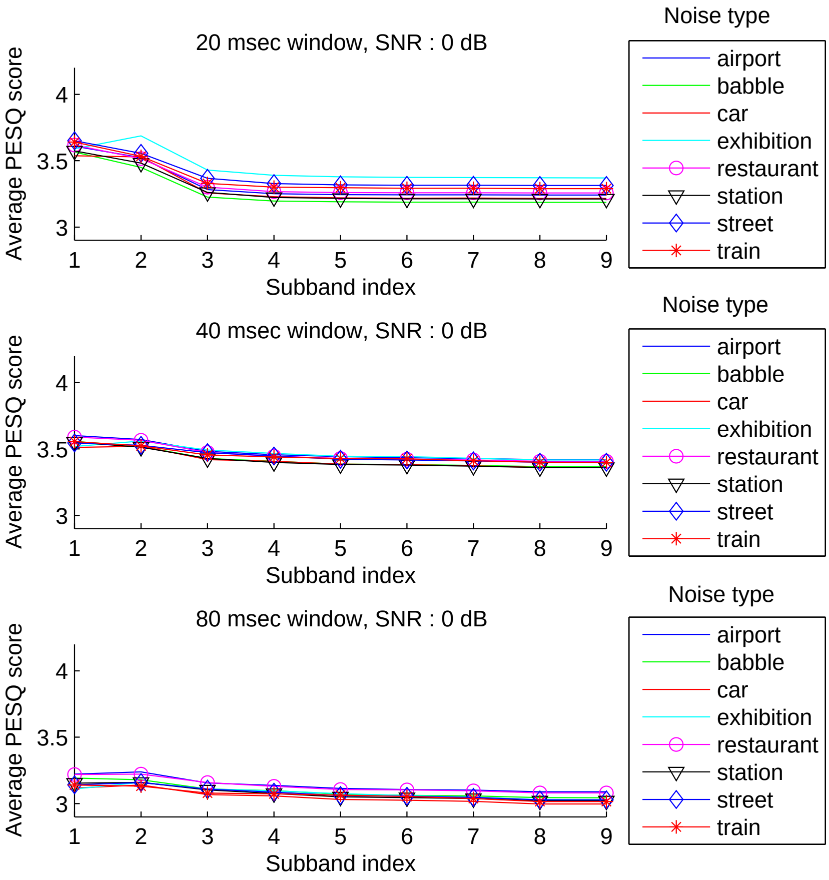


Figure 4.4: Average PESQ scores scores of 30 sentences for 0 dB signal to noise ratio with window lengths of 20, 40 and 80 msec. The first 8 subbands in the x-axis correspond to the subbands shown in Figure-4.3, over which clean phase is used. The 9th subband corresponds to the all noisy phase case (base score).

nounced by 3 male and 3 female speakers. Each audio file in this database is processed using the system in Figure-4.1 and the resulting PESQ scores of the reconstructed signals are recorded.

The tests are conducted for 20, 40 and 80 msec long analysis frames with %50 overlap, using Hamming window. The STFT reconstruction procedure is explained in Chapter-2.

The PESQ scores are evaluated (for the structure in Figure-4.1) for 8 different noise types which are given in Figure-4.4, Figure-4.5, Figure-4.6 for 0dB, 5dB and 10dB noise levels, respectively. The lines in each figure correspond to a specific type of exposed noise. The PESQ scores are obtained for 30 different audio files and the average score of these 30 sentences

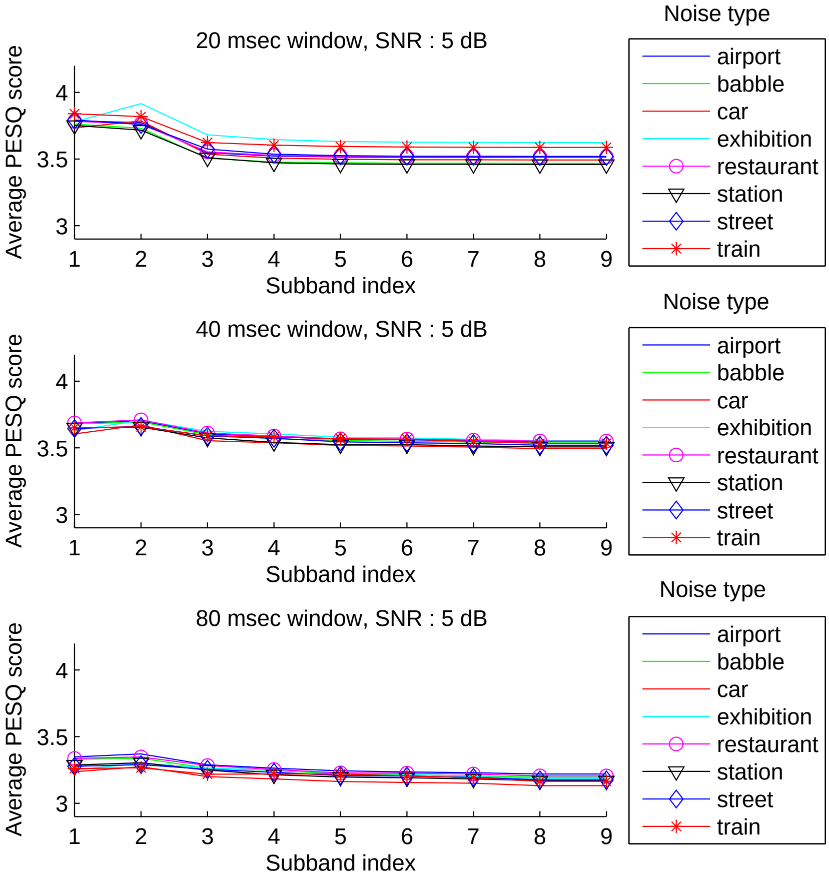


Figure 4.5: Average PESQ scores scores of 30 sentences for 5 dB signal to noise ratio with window lengths of 20, 40 and 80 msec. The first 8 subbands in the x-axis correspond to the subbands shown in Figure-4.3, over which clean phase is used. The 9^{th} subband corresponds to the all noisy phase case (base score).

for each band is plotted. X-axis in the figures corresponds to the subband index over which the phase of the clean speech is used (rest of the phase spectrum uses noisy phase). The first subband corresponds to 0-500Hz, the second corresponds to 500-1000Hz and so on, as shown in Figure-4.3.

The results of the conducted simulations are quite interesting. As seen in Figure-4.4, for 0dB noise level, most of the quality gain is obtained from the 0-500 and 500-1000 Hz bands. These bands actually contain the first few harmonics, when the speech is voiced. When the noise level decreases, Figure- 4.5 and Figure-4.6, we observe that the effect of the first subband (0-500Hz) decreases and the second subband (500-1000Hz) becomes dominant.

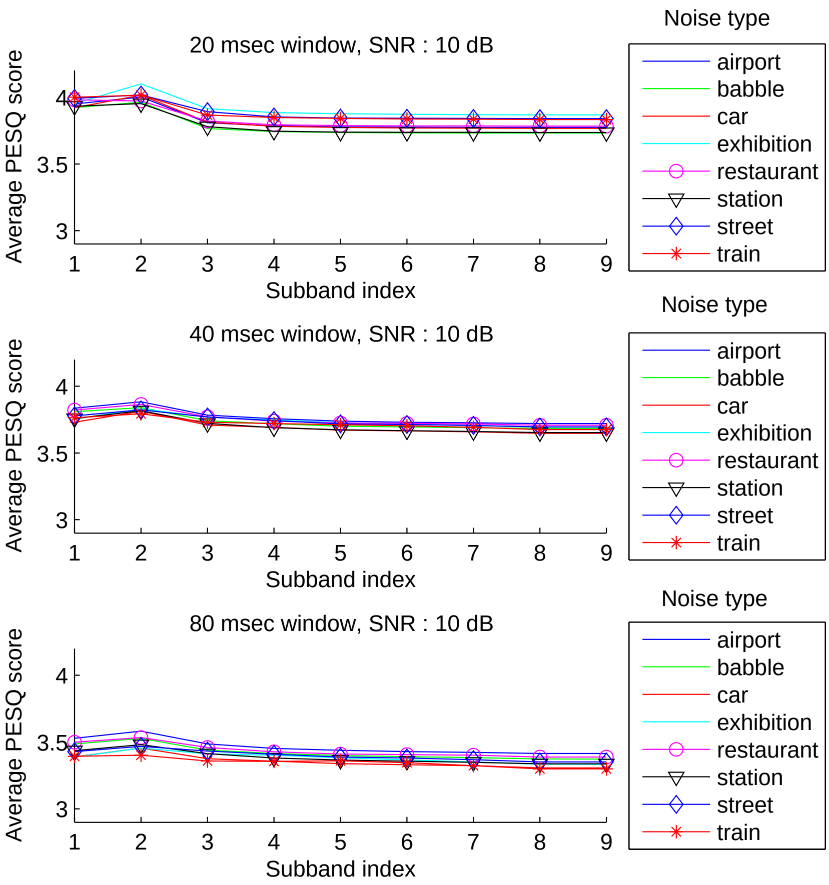


Figure 4.6: Average PESQ scores scores of 30 sentences for 10 dB signal to noise ratio with window lengths of 20, 40 and 80 msec. The first 8 subbands in the x-axis correspond to the subbands shown in Figure-4.3, over which clean phase is used. The 9^{th} subband corresponds to the all noisy phase case (base score).

Another observation is that; for short durations of analysis windows (20msec) the contribution of higher frequency bands are somewhat negligible compared to the low frequency bands. For longer analysis windows, the contribution of higher frequency bands are still less than low frequency bands however the quality gain is somewhat spread to the whole spectrum in a more balanced manner. This fact is better observed for lower SNR values. As seen in Figure-4.4, for 20 msec window; the average PESQ score for the reconstructed signals using the clean phase in the first subband is about 3.6, however the score drops to 3.2 when the clean phase is used in either one of the 4^{th}, 5^{th}, 6^{th}, 7^{th} or 8^{th} subband. When the analysis window length is 80 msec, the gap between the average PESQ scores reduces to 0.1 points (3.2 to 3.1) which indicates a more balanced contribution from all subbands compared to 20 msec window.

It can be concluded that, most of the quality gain can be attained by correcting the phase spectra of the low frequency components, especially for short analysis windows when the SNR is low. Hence a better phase estimate of the low frequency components should provide a considerable improvement. Considering the fact that such low frequency band (0-500Hz) is dominated by a few sinusoids (fundamental component and its harmonics) in voiced speech; the phase estimation problem can be narrowed down to a few frequency bins. Therefore, the problem is converted to the estimation of phase values of these frequency bins. The proposed solution for the phase estimation of low frequency components is presented in Chapter-5.

CHAPTER 5

THE PROPOSED METHOD

In this chapter, the structure of the proposed method for maintaining the phase continuity in the classical speech enhancement algorithms is explained. The advantages as well as the drawbacks of the system are also stated. The methods used to solve the phase estimation and pitch estimation problems are also explained in detail. The validation of the system and the implementation results will be given in the following chapters.

5.1 Voiced segments of speech signals

Voiced segments of speech signals exhibit highly tonal characteristics. These parts of speech signals can be modeled as a sum of sinusoids in the following form; $\sum_{k=1}^{M} A_k cos\left(2\pi k \frac{f_0}{f_s} n + \phi_k\right)$. This sinusoidal structure imposes a certain constraint on the signal, as it will be explained in the next section. If the imposed constraint is taken into consideration in the signal reconstruction procedure, a better phase estimate can be obtained for the frequency bins that encompass the fundamental frequency and even for the frequency bins that cover the higher harmonics.

5.2 General formulation

As mentioned in the previous chapters, classical speech enhancement algorithms do not modify the phase spectra of the corrupted signal and simply use the phase spectra of the noisy signal in the reconstruction process. The effects of this phase distortion were explained in Chapter-4.

The main idea in the proposed system is to maintain the phase continuity in the reconstructed signal. To achieve this, the following simple fact will be utilized;

- Let $x(t) = cos\,(2\pi f_0 t + \phi_0)$ be continuous time signal and $x[n] = cos\left(2\pi \frac{f_0}{fs} n + \phi_0\right)$ be the sampled version of $x(t)$.
- The instantaneous phase of a sinusoid, in continuous time, is defined as follows;

  $$\theta(t) = \int_{-\infty}^{t} \omega(\tau) d\tau \tag{5.1}$$

  where $\omega(t)$ is the instantaneous radial frequency and is equal to $2\pi f_0(t)$.

  The instantaneous phase can be calculated by numerically evaluating the above integral, using the discrete time data. However, it is not practically efficient to calculate pitch frequency estimates densely in time. Instead the fundamental frequency (or pitch) can be estimated over a rather longer time frame. In this case the estimate will represent the average value of the $f_0(t)$ over that time frame. This estimate can be utilized to numerically evaluate the above integral by using the rectangle rule. This fact can be stated as follows;

  Let the instantaneous phase at time $n = n_0$, $(t = \frac{n_0}{f_s})$ be equal to $\theta[n_0] = \phi_0$ and the fundamental frequency estimate over the time interval $[t_0, t_0 + \frac{N}{f_s}]$ be f_1. In this case the instantaneous phase estimate for $n = n_0 + N$, $(t = t_0 + \frac{N}{f_s})$ will be;

  $$\theta[n_0 + N] = \phi_0 + 2\pi \int_{t_0}^{t_0 + \frac{N}{f_s}} f_0(\tau) d\tau \tag{5.2}$$

  $$\begin{aligned} \theta[n_0 + N] &\approx \phi_0 + 2\pi f_1 \Delta t \\ &\approx \phi_0 + 2\pi f_1 \frac{N}{f_s} \end{aligned} \tag{5.3}$$

- The difference between the instantaneous phase values is seen in equation 5.4. Notice that the phase difference is independent of the initial phase ϕ_0.

  $$\Delta\theta = \theta[n_0 + N] - \theta[n_0] = 2\pi \frac{f_1}{f_s} N \tag{5.4}$$

This simple fact (5.4) can be applied to speech enhancement algorithms. If the fundamental frequency is somehow estimated, equation-5.4 can be used to correct the phase difference between two consecutive frames, since the voiced segments of the input signals have a tonal harmonic structure. The phase values $\theta[n_0 + N]$ and $\theta[n_0]$ correspond to the instantaneous phase values at the beginning of these two consecutive analysis frames.

Figure-5.1 summarizes the proposed method. The system is built upon classical speech enhancement methods which only modify the magnitude spectra of the noisy signal. As seen in the figure, there are 3 analysis processes acting in parallel, namely; ‘spectral analysis’, ‘phase estimation’ and ‘pitch extraction’. Spectral analysis can be done by using STFT (Short Time Fourier Transform, see Chapter-2). Due to the inadequacy of STFT in phase and pitch estimation, other methods are employed as explained in detail in the following chapters. Using the outputs of the three analysis bocks, at each frame, the phase of the current frame will be corrected by using the phase of the previous frame and the extracted pitch. In order to achieve this, the input speech must be voiced within these two consecutive analysis frames. And the mentioned ‘phase’ refers to the phase of the corresponding frequency bins of the extracted pitch and its harmonics.

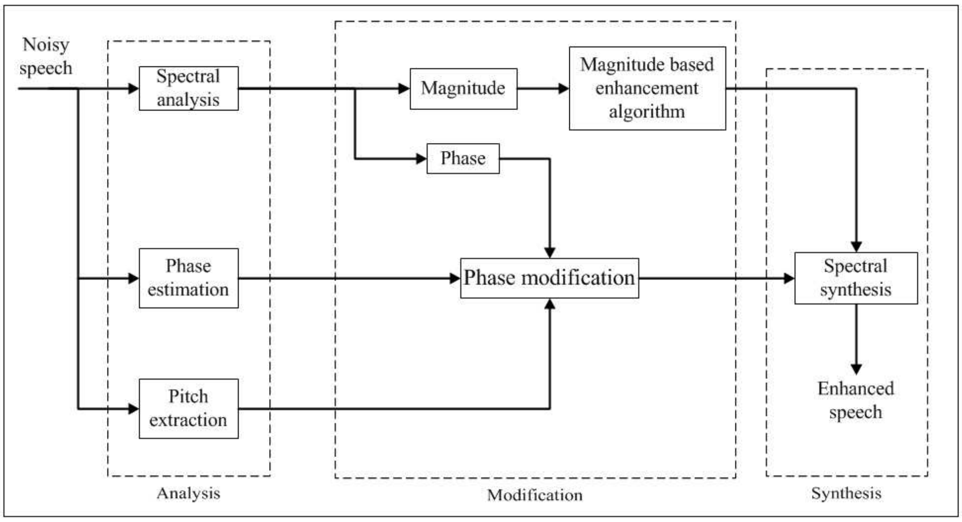


Figure 5.1: Proposed system

The proposed method modifies the phase spectra of the input speech. In that respect, the method differs from the classical enhancement algorithms. Moreover, the algorithm keeps the improvements obtained by the classical algorithms, as it uses the output of the classical methods and combines it with its own.

In the following sections; phase estimation and pitch estimation algorithms will be presented.

Then, in the following chapters; validation of the algorithm is demonstrated using the simulation results and showing that equation-(5.4) holds. Lastly the implementation of the proposed structure and the results will be presented.

Note : The idea presented in this chapter has been thought to be original, however a similar approach is used in [30] in the context of speech synthesis. Nevertheless the concept is not applied to speech enhancement algorithms in any previous work.

5.3 Phase estimation of tonal signals

In this section, an efficient method for phase estimation and spectral analysis, namely the all-phase DFT (although we believe that the method is not named appropriately, the same naming as the original paper [2] will be used throughout this work), is presented. A rigorous derivation of the method, which the original paper [2] lacked, is stated. Furthermore, a reconstruction method for the all-phase DFT analysis is proposed.

5.3.1 Introduction

Discrete Fourier Transform is one of the most commonly used signal analysis method, due to the fact that its basis functions, $\{e^{j\frac{2\pi}{N}kn}\}_{k=0}^{N-1}$, are the eigen-functions of discrete-time linear time invariant (LTI) systems. Hence it is a good way to characterize LTI systems. Also the fast implementation algorithms (FFT: [7], FFTW: [31], [32]) made this transform even more popular.

Although the aforementioned properties of DFT are very attractive, there are some drawbacks of this transform for certain applications; for instance, the phase estimation of a sinusoid is problematic when the period of the signal does not fit to the observation window. In other words, when the observation window length is not an integer multiple of the period of the (tonal) signal to be observed, the phase spectrum gives a function of the desired phase value, instead of directly giving the phase value of the sinusoid. The mathematical derivation of this fact is as follows;

Consider the simplest case where the input signal is a single complex exponential as in equation-5.5.

$$x[n] = A_0 e^{j\left(2\pi \frac{f_0}{f_s} n + \phi_0\right)}, A_0 \in \mathbb{R} \tag{5.5}$$

The DFT of $x[n]$ (using rectangular window of length N) will be equal to;

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn} \tag{5.6}$$

$$= \sum_{n=0}^{N-1} A_0 e^{j\left(2\pi \frac{f_0}{f_s} n + \phi_0\right)} e^{-j\frac{2\pi}{N}kn} \tag{5.7}$$

$$X[k] = A_0 e^{j\phi} \sum_{n=0}^{N-1} e^{j2\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)n} \tag{5.8}$$

$$= A_0 e^{j\phi} \frac{1 - e^{j2\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N}}{1 - e^{j2\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)}} \tag{5.9}$$

$$= A_0 e^{j\phi} \frac{e^{j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N}}{e^{j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)}} \frac{e^{-j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N} - e^{j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N}}{e^{-j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)} - e^{j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)}} \tag{5.10}$$

$$= A_0 e^{j\phi} e^{j\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)(N-1)} \frac{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)} \tag{5.11}$$

The magnitude spectrum of $X[k]$ is in the following form;

$$|X[k]| = |A_0| \frac{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)} \tag{5.12}$$

It is important to note that the input signal is actually multiplied by the window function in time domain. This operation corresponds to convolution operation in the frequency domain and in order not to corrupt the signal properties, the frequency response of the window function should be an impulse function. Since the window function is bounded in time domain, its frequency response will be unbounded and can never be equal to an impulse function. Hence in the transform domain, the energy of the signal components will spread along the entire spectrum. This problem is called the ”spectral leakage” and it can be observed in equation 5.12, as the energy of a single complex exponential spreads over the entire spectrum because of the $\frac{sin(.)}{sin(..)}$ term.

There is also a problem with the phase value obtained through $X[k]$. The phase of the $X[k]$ is;

$$\angle X[k] = arctan\left(\frac{Im\{X[k]\}}{Re\{X[k]\}}\right) \tag{5.13}$$

$$\angle X[k] = \phi_0 + \pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)(N-1) \tag{5.14}$$

As seen in equation (5.14), unless the condition in (5.15) is satisfied, the phase of DFT result will be a function of both the $\frac{f_0}{f_s}$ ratio as well as the window length N, and it will not be possible to directly observe ϕ_0 via the phase of DFT.

$$\begin{aligned} \pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right) &\equiv 0, \quad (\text{mod } 2\pi) \\ \frac{\pi}{f_s}\left(f_0 - \frac{kf_s}{N}\right) &\equiv 0, \quad (\text{mod } 2\pi) \qquad (5.15) \end{aligned}$$

The condition in (5.15) actually states that, in order to observe the phase of the sinusoid directly from the phase of DFT, the frequency of the sinusoid (f_0) must be equal to the center frequency of the k^{th} frequency bin ($k\frac{f_s}{N}$).

In other words, the duration of the observation window (N/f_s) must be an integer multiple of the signal period ($1/f_0$). Note that, it is also possible to find the exact frequency of the signal in this case, since the $\frac{sin(.)}{sin(..)}$ function in equation (5.12) yields one when (5.15) is satisfied. However, this is hardly the case in practice, i.e. the condition in (5.15) is rarely satisfied. Also the input signal in general does not consist of a single tone and the aforementioned ‘spectral leakage’ effect creates other problems as well.

As explained above, even in the simplest case (single complex exponential) the actual phase of a tonal signal cannot be observed directly by using the DFT phase, unless a very stringent condition is satisfied. Hence, the instantaneous phase of a tonal signal must be estimated by using some other means.

5.3.2 All-phase DFT analysis

As explained in the previous section it is not possible to observe the actual phase values of the signal components directly from the DFT phase.

A computationally efficient method for instantaneous phase estimation and spectral analysis is proposed in [2], called the all-phase DFT. In this method N-point DFT of the signal is estimated using 2N-1 observation points. The procedure is as follows;

- Get a (2N-1) point frame, centered at time n = 0.
  $x[n] = [x_{(-N+1)} \quad x_{(-N+2)} \quad \dots \quad x_{-1} \quad x_0 \quad x_1 \quad \dots \quad x_{(N-2)} \quad x_{(N-1)}],$
  $where \quad x[0] = x_0.$

- Obtain all shifted windows of length N from this frame.

$$\begin{array}{lllllllll}
x_0[n] & = [& x_0 & x_1 & \dots & x_{(N-2)} & x_{(N-1)} &], & x_0[0] = x_0 \\
x_1[n] & = [& x_{-1} & x_0 & \dots & x_{(N-3)} & x_{(N-2)} &], & x_1[0] = x_{-1} \\
x_2[n] & = [& x_{-2} & x_{-1} & \dots & x_{(N-4)} & x_{(N-3)} &], & x_2[0] = x_{-2} \\
\vdots & & & & & & & & \\
x_{(N-1)}[n] & = [& x_{-N+1} & x_{-N+2} & \dots & x_{-1} & x_0 &], & x_{N-1}[0] = x_{-N+1}
\end{array}$$

  The general form of these windows is as follows;

$$x_i[m] = x[m-i], \qquad i = 0, \dots, N-1 \quad m = 0, \dots, N-1 \tag{5.16}$$

  Notice that all windows include the sample x_0.

- Circularly shift each window, such that the sample x_0 is the first element in all windows.

$$\begin{array}{lllllllll}
x'_0[n] & = & [& x_0 & x_{(-N+1)} & x_{(-N+2)} & x_{(-N+3)} & x_{-2} & x_{(-1)} &] \\
x'_1[n] & = & [& x_0 & x_1 & x_{(-N+2)} & x_{(-N+3)} & x_{-2} & x_{(-1)} &] \\
x'_2[n] & = & [& x_0 & x_1 & x_2 & x_{(-N+3)} & x_{-2} & x_{(-1)} &] \\
\vdots & & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
x'_{(N-2)}[n] & = & [& x_0 & x_1 & x_2 & x_3 & x_{(N-2)} & x_{(-1)} &] \\
x'_{(N-1)}[n] & = & [& x_0 & x_1 & x_2 & x_3 & x_{(N-2)} & x_{(N-1)} &]
\end{array}$$

  Notice that the DFT coefficients of the circularly shifted sequences will be in the following form;

$$X'_i[k] = X_i[k]e^{j\frac{2\pi}{N}ki} \tag{5.17}$$

- Lastly, take the average of the signals $\{x'_0[n] \dots x'_{(N-1)}[n]\}$ and take the DFT of the result. The subscript 'ap' in the latter equations stands for 'all-phase'.

$$x_{ap}[n] = \frac{1}{N}\sum_{i=1}^{N-1} x'_i[n] \tag{5.18}$$

$$X_{ap}[k] = \frac{1}{N}\sum_{i=1}^{N-1} X_i[k] e^{j\frac{2\pi}{N}ki} \tag{5.19}$$

We can write the all-phase DFT of $x[n]$, in terms of $x[n]$ as follows;

$$\begin{aligned} X_{ap}[k] &= \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{n=0}^{N-1} x_i[n]e^{-j\frac{2\pi}{N}kn}\right)e^{j\frac{2\pi}{N}ki} \\ &= \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{n=0}^{N-1} x[n-i]e^{-j\frac{2\pi}{N}kn}\right)e^{j\frac{2\pi}{N}ki} \end{aligned} \tag{5.20}$$

Equation 5.20 summarizes the implementation of all-phase DFT analysis. To observe the properties of the all-phase DFT, consider a single complex exponential in the following form; $y[n] = A_0 e^{j\left(2\pi\frac{f_0}{f_s}n+\phi_0\right)}, A_0 \in \mathbb{R}$. The all-phase DFT of this signal is as follows;

$$\begin{aligned} Y_{ap}[k] &= \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{n=0}^{N-1} y[n-i]e^{-j\frac{2\pi}{N}kn}\right)e^{j\frac{2\pi}{N}ki} \\ &= \frac{1}{N}\sum_{i=0}^{N-1}\left(\sum_{n=0}^{N-1} A_0 e^{j\left(2\pi\frac{f_0}{f_s}(n-i)+\phi_0\right)}e^{-j\frac{2\pi}{N}kn}\right)e^{j\frac{2\pi}{N}ki} \\ &= \frac{A_0e^{j\phi_0}}{N}\sum_{i=0}^{N-1}\left(\sum_{n=0}^{N-1} e^{j2\pi\frac{f_0}{f_s}n}e^{-j2\pi\frac{f_0}{f_s}i}e^{-j\frac{2\pi}{N}kn}\right)e^{j\frac{2\pi}{N}ki} \\ &= \frac{A_0e^{j\phi_0}}{N}\sum_{i=0}^{N-1}\sum_{n=0}^{N-1} e^{j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)n}e^{-j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)i} \\ &= \frac{A_0e^{j\phi_0}}{N}\frac{1-e^{j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)N}}{1-e^{j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)}}\frac{1-e^{-j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)N}}{1-e^{-j2\pi\left(\frac{f_0}{f_s}-\frac{k}{N}\right)}} \end{aligned} \tag{5.21}$$

Let $\theta = 2\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)$. Then;

$$Y_{ap}[k] = \frac{A_0 e^{j\phi_0}}{N} \frac{1 - e^{j\theta N}}{1 - e^{j\theta}} \frac{1 - e^{-j\theta N}}{1 - e^{-j\theta}}$$

$$= \frac{A_0 e^{j\phi_0}}{N} \left| \frac{1 - e^{j\theta N}}{1 - e^{j\theta}} \right|^2$$

$$= \frac{A_0 e^{j\phi_0}}{N} \frac{[1 - cos(\theta N)]^2 + sin^2(\theta N)}{[1 - cos(\theta)]^2 + sin^2(\theta)}$$

$$= \frac{A_0 e^{j\phi_0}}{N} \frac{1 - 2cos(\theta N) + cos^2(\theta N) + sin^2(\theta N)}{1 - 2cos(\theta) + cos^2(\theta) + sin^2(\theta)}$$

$$= \frac{A_0 e^{j\phi_0}}{N} \frac{2 - 2cos(\theta N)}{2 - 2cos(\theta)} \tag{5.22}$$

Using the trigonometric identity $cos(\theta) = 1 - 2sin^2\left(\frac{\theta}{2}\right)$;

$$Y_{ap}[k] = \frac{A_0 e^{j\phi_0}}{N} \frac{sin^2\left(\frac{\theta N}{2}\right)}{sin^2\left(\frac{\theta}{2}\right)} \tag{5.23}$$

And putting back $\theta = 2\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)$. Then;

$$Y_{ap}[k] = e^{j\phi_0} \frac{A_0}{N} \frac{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)} \tag{5.24}$$

$$\left|Y_{ap}[k]\right| = \frac{|A_0|}{N} \frac{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)} \tag{5.25}$$

$$\angle Y_{ap}[k] = \phi_0 \tag{5.26}$$

Equations (5.25) and (5.26) depict the most important properties of all-phase DFT analysis. The phase of the k-th coefficient of all-phase DFT simply gives the phase value of the complex exponential whose frequency falls into the k^{th} frequency bin. The key point here is that the phase value is not a function of the signal frequency, unlike the traditional DFT. Also notice that, the magnitudes of the coefficients have better side-lobe suppression than that of the traditional DFTs'.

5.3.2.1 Efficient structure for calculating all-phase DFT

Up to this point, the mathematical derivation of all-phase DFT and some of its important properties are presented. However it is highly inefficient to use equation (5.20) to calculate the all-phase DFT of a signal, as it requires the calculation of N-point DFTs of N different sequences and their average. Also the circular shift operation is added to this computational load. Fortunately there is an efficient way to calculate the transform. Instead of calculating N different N-point DFTs, consider equation (5.18) again;

$$x_{ap}[n] = \frac{1}{N}\sum_{i=1}^{N-1} x_i[n] = \frac{1}{N}\sum_{i=0}^{N-1} x[n-i] \tag{5.27}$$

$$= \frac{1}{N}\begin{bmatrix} Nx[0] \\ (N-1)x[1] + x[-N+1] \\ (N-2)x[2] + 2x[-N+2] \\ \cdots \\ x[N-1] + (N-1)x[-1] \end{bmatrix}^T \tag{5.28}$$

Equation (5.28) can be written in the following form;

$$x_{ap}[n] = \frac{1}{N}\begin{bmatrix} 0 \\ x[-N+1] \\ 2x[-N+2] \\ 3x[-N+3] \\ \vdots \\ (N-2)x[-2] \\ (N-1)x[-1] \end{bmatrix}^T + \frac{1}{N}\begin{bmatrix} Nx[0] \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}^T + \frac{1}{N}\begin{bmatrix} 0 \\ (N-1)x[1] \\ (N-2)x[2] \\ (N-3)x[3] \\ \vdots \\ 2x[N-2] \\ x[N-1] \end{bmatrix}^T \tag{5.29}$$

The important point in equation (5.29) is that, the elements of each vector can be obtained from the windowed version of the original $2N-1$ point data. Equation (5.29) can be written in the form of equation (5.30) to clarify the previous statement. Also it is easily seen that the window function is a $2N-1$ point triangular window and is depicted as $w[n]$ in the next equation.

$$x_{ap}[n] = \begin{bmatrix} 0 \\ w[-N+1]x[-N+1] \\ w[-N+2]x[-N+2] \\ w[-N+3]x[-N+3] \\ \vdots \\ w[-2]x[-2] \\ w[-1]x[-1] \end{bmatrix}^T + \begin{bmatrix} w[0]x[0] \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}^T + \begin{bmatrix} 0 \\ w[1]x[1] \\ w[2]x[2] \\ w[3]x[3] \\ \vdots \\ w[N-2]x[N-2] \\ w[N-1]x[N-1] \end{bmatrix}^T \tag{5.30}$$

Using equation (5.30) the structure in Figure 5.2 can be implemented. This structure is also given in [2].

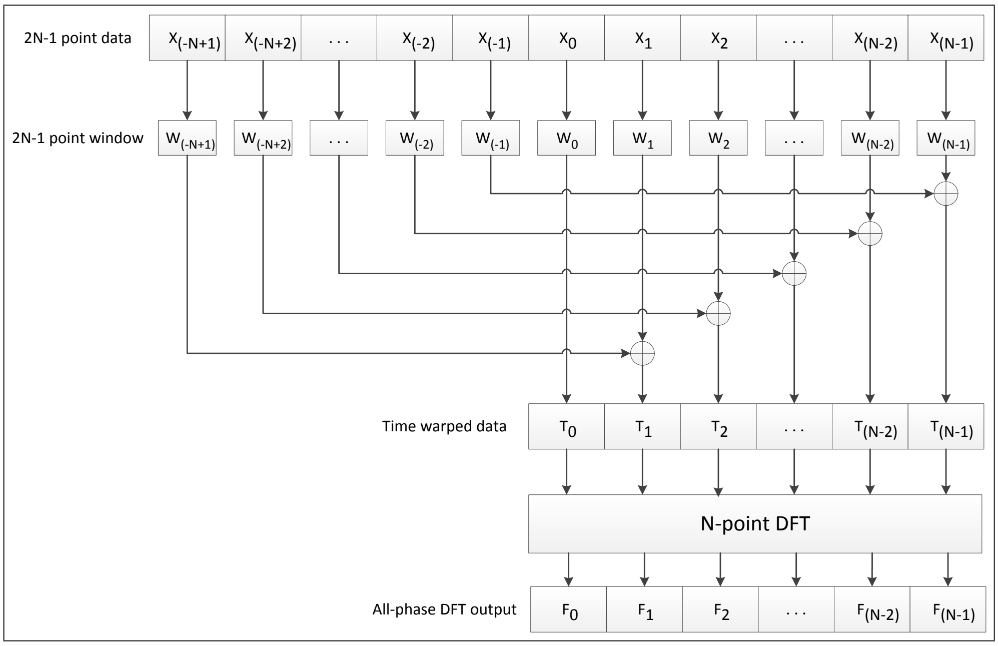


Figure 5.2: Efficient structure for calculating all-phase DFT, (taken from [2])

As a summary, the outputs of all-phase DFT and traditional DFT for a single complex exponential are given in Table 5.1.

When Table 5.1 is examined it might be concluded that all-phase DFT has superior characteristics than the traditional DFT. However there are some drawbacks which may be summarized as follows;

Table 5.1: Comparison of the traditional DFT and all-phase DFT

Method	**Time domain signal**	**Transform magnitude**	**Transform phase**
DFT	$A_0 e^{j\left(2\pi \frac{f_0}{f_s} n + \phi_0\right)}$	$\lvert A_0 \rvert \frac{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)}$	$\phi_0 + \pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)(N-1)$
all-phase DFT	$A_0 e^{j\left(2\pi \frac{f_0}{f_s} n + \phi_0\right)}$	$\frac{\lvert A_0 \rvert}{N} \frac{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)N\right)}{sin^2\left(\pi\left(\frac{f_0}{f_s} - \frac{k}{N}\right)\right)}$	ϕ_0

- To compute N-point all-phase DFT, one has to have $2N-1$ samples. Although all-phase DFT requires $2N-1$ samples, it combines these data samples in way that it ends up with an N-point hybrid data set and it outputs the DFT of this hybrid data set. Hence the frequency resolution is the same as the N-point DFT while the time resolution is about half the N-point DFT's time resolution. A solution to this problem is given in [33]. In this solution the frequency resolution is increased through some some additional procedures and made equal to $2N-1$ point DFT's frequency resolution. The problem might also be partially solved by using overlapping frames; however increasing overlap ratio will increase the computational load as well as the complexity of synthesis procedure if required.

- The computational complexity of all-phase DFT can be calculated for $2N-1$ point data, as follows; Computational load = $(2N-2)$ multiplication + $(N-1)$ addition + N-point FFT

- Although the all-phase DFT is a powerful analysis tool, a well-defined modification and synthesis procedure is not available for a general class of signals. Due to this reason the method might be undesirable in some applications.

5.3.2.2 Simulation results

The performance of all-phase DFT is tested with some simulations in MATLAB. The outputs of the traditional DFT and all-phase DFT are compared.

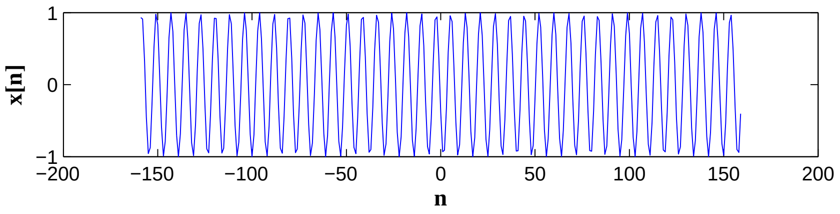


Figure 5.3: Input signal for the first simulation (Single tone)

Case study: Single tone input

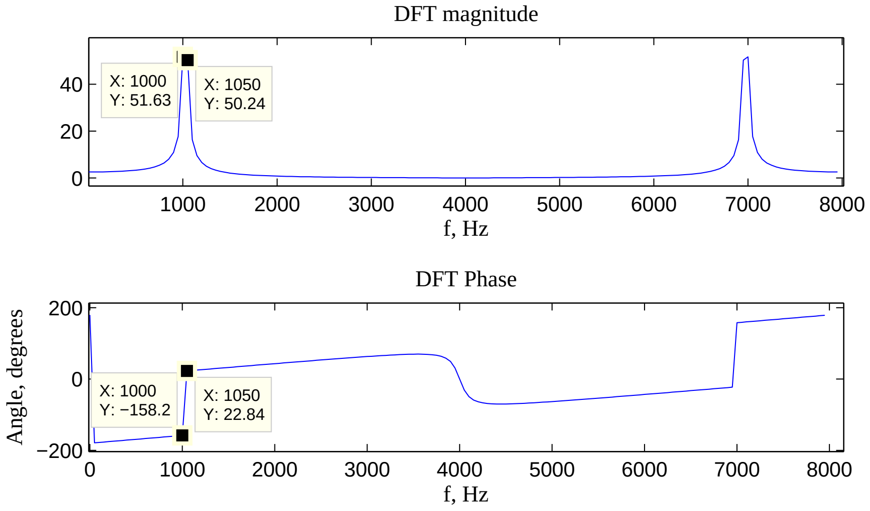


Figure 5.4: DFT output for the first simulation

The first simulation is carried out with a cosine input in the following form;

$$x[n] = cos\left(2\pi\frac{f_0}{f_s}n + \phi_0\right), \qquad \phi_0 = 112^{\circ}, 3 \tag{5.31}$$

For the first simulation f_0 is taken as 1025 Hz and window length N is taken as 160 points with a sampling frequency of 8 kHz. The initial phase ϕ_0 is chosen to be 112.3 degrees. Notice that all-phase DFT will be computed for $2N - 1$ points. Hence the signal is computed for $n = -N + 1, \ldots, N + 1$ for the all-phase DFT process. The reason for this selection of time index 'n' (starting from $-N + 1$ instead of zero) is to observe the phase value of the input signal directly. Because the phase of the all-phase DFT output corresponds to the

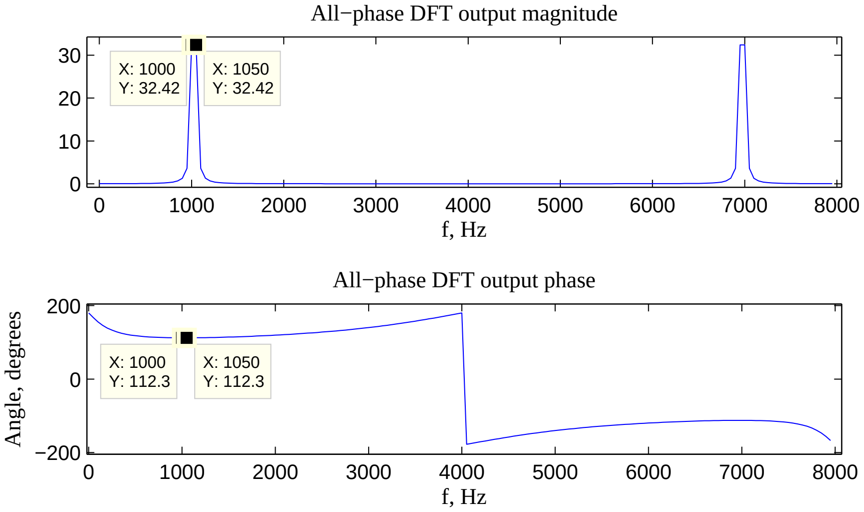


Figure 5.5: All-phase DFT output for the first simulation

instantaneous phase value of the middle point in the frame and when ‘n’ is chosen in the aforementioned way, the instantaneous frequency of the point in the middle of the frame is; $2\pi\frac{f_0}{f_s}n+\phi_0$ evaluated at $n = 0$, which simply yields ϕ_0. For the evaluation of DFT, the signal is evaluated for $n = 0, \ldots, N-1$ in order to have the same frequency resolution as the all-phase DFT and to observe the instantaneous phase estimate of DFT for $n = 0$, since DFT phase corresponds to the instantaneous phase of the signal at the beginning of the frame. The input signal can be observed in Figure-5.3.

With the parameters selected as above, the condition which was elaborated in equation-(5.15), $\left(f_0 - k\frac{f_s}{N}\right)$ is not equivalent to zero in mod 2π for integer values of k. Actually the window length is the worst possible selection, since the frequency resolution is $8000Hz/160 = 50Hz$ and the signal frequency falls just into the middle of two consecutive frequency bins; 1000 Hz and 1050 Hz. As a result, most of the signal energy is divided between these two frequency bins.

The output of traditional DFT is given in Figure-5.4. As indicated in the figure the phase values corresponding to 1000 Hz and 1050 Hz bins are -158.2 and 22.84 degrees respectively. Either of these values can be used to obtain ϕ_0 using equation-(5.14), with the additional knowledge of signal frequency as follows;

$$
\begin{aligned}
\angle X[k] &= \phi_0 + \frac{\pi}{f_s}\left(f_0 - k\frac{f_s}{N}\right)(N-1) \\
22.84\frac{\pi}{180} &= \phi_0 + \frac{\pi}{8000}(1025-1000)(N-1) \\
\phi_0 &\equiv 112^\circ.3
\end{aligned}
$$

There is a method in literature for estimating the frequency and phase of such tonal signals using DFT output [34]. However the method has a drawback as it is also stated in [34] that; ”the method will not work well when the distance between two spectral peaks is less than 5 frequency resolutions”. Unfortunately this is the case when the application is narrowband (f_s=8kHz) speech enhancement, as the fundamental frequency can be as low as 100 Hz and its harmonics will be very closely packed when the traditional 20-40 msec window duration is used. Hence one needs to increase the frequency resolution, but it is not possible without decreasing the time resolution. As a result the problem needs to be solved by other means.

The output of the all-phase DFT on the other hand, is given in Figure-5.5. As indicated in the figure, maximum amplitude is observed both in 1000 Hz and 1050 Hz frequency bins due to the reason mentioned above. The phase of the output however, gives the same result which is 112.3 degrees and is the same as the instantaneous phase of the input signal at time $n = 0$. Notice that no additional information about signal content was used to predict the signal phase.

Case study: Sum of harmonics

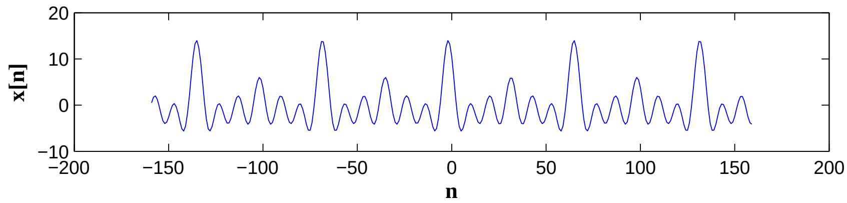


Figure 5.6: Input signal for the second simulation (Sum of 6 harmonics)

In order to observe the performance of the two methods when the input is a speech-like signal, the second simulation is conducted with an input signal of the following form; $x[n] =$

$\sum_{k=1}^{M} A_k cos(2\pi k \frac{f_0}{f_s} n + \phi_k)$. In this simulation f_0 was chosen as 120 Hz (a typical pitch frequency for a male speaker) and 6 harmonics are used. The phase values are selected as $\phi_i = 10i$ degrees. The amplitudes A_i are chosen as [1, 3, 2, 3, 1, 4] for i=1,. . . ,6 respectively. The sampling frequency, f_s, is taken as 8 kHz again. With these parameters the generated input signal can be observed in Figure-5.6.

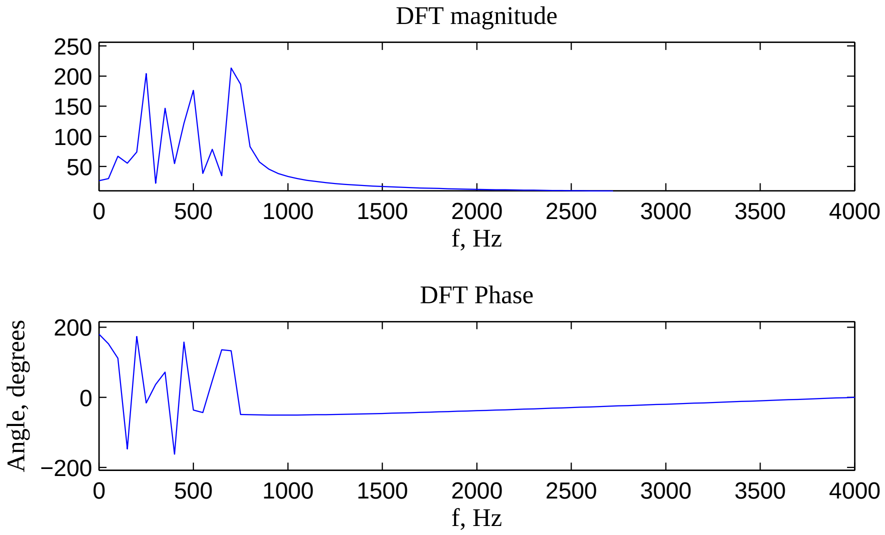


Figure 5.7: DFT output for the second simulation

The output of traditional DFT for the second simulation is given in Figure-5.7. Notice that in the figures the spectrum is zoomed to [0, $\frac{f_s}{2}$] frequency band for better visualization, since the signal under consideration is real hence the spectrum is conjugate symmetric. In this case since the tones are closely packed, spectral leakage phenomenon becomes a serious problem, both for phase estimation and magnitude estimation. The frequency bins of the observed peaks and the corresponding phase values are given in Table-5.2. The ”interpolated phase” tab in this table indicates the operation of estimating the actual phase of the corresponding component, using equation-(5.14) and assuming the actual frequency of that component is known.

An important problem in STFT analysis can be observed in Table-5.2. Although the amplitudes of the first (120 Hz) and fifth (600 Hz) harmonic components are the same, the DFT magnitudes of the corresponding components differ significantly. One of the underlying rea-

Table 5.2: Attributes of the observed peaks in DFT magnitude spectrum in Figure-5.7

Actual frequency	**DFT bin center**	**Magnitude**	**DFT Phase**	**Interpolated phase (eq.5.14)**	**Actual phase**
120Hz	100Hz	67.17	112°.20	39°.65	10°
240Hz	250Hz	204.2	−15°.78	19°.99	20°
360Hz	350Hz	146.7	71°.98	36°.20	30°
480Hz	500Hz	176.3	−36°.16	35°.39	40°
600Hz	600Hz	78.53	47°.10	47°.10	50°
720Hz	700Hz	213.3	132°.70	61°.15	60°

sons for this problem is the fact that there are stronger tones in the neighboring frequency bins and the energy spread of these components modifies the content of the nearby frequency bins. The DFT phase is also modified by the same reason and it is not possible to obtain the actual phase of each component by using the knowledge of the exact frequency and DFT phase, using only equation-(5.14). That was possible in the single tone case where the amplitude of the tone was not affected by other components. As seen in Table-5.2 the difference between the actual phase and interpolated phase values is larger for the weak components. Also notice that the interpolated phase value is equal to DFT phase for 600Hz component, since the signal frequency is exactly equal to the center frequency of the corresponding frequency bin. In other words the condition in (5.15) is satisfied for the 600Hz component.

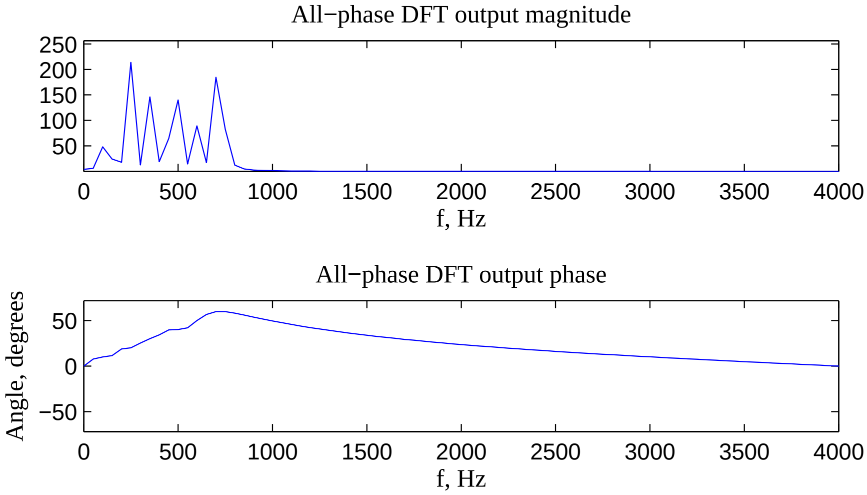


Figure 5.8: All-phase DFT output for the second simulation

The output of all-phase DFT for the second simulation is given in Figure-5.8. If the figure is observed closely the fact that the phase spectra remains locally constant around the frequency bins where the tones reside in the spectrum. The frequency bins of the observed peaks in this figure and the corresponding phase values are given in Table-5.3. It is clear that the obtained results are superior to those obtained via DFT.

Table 5.3: Attributes of the observed peaks in all-phase DFT magnitude spectrum in Figure-5.8

Actual frequency	**All-phase DFT bin center**	**Magnitude**	**All-phase DFT Phase**	**Actual phase**
120Hz	100Hz	48.24	10°.12	10°
240Hz	250Hz	213.80	20°.07	20°
360Hz	350Hz	145.90	30°.06	30°
480Hz	500Hz	140.10	40°.01	40°
600Hz	600Hz	89.23	49°.89	50°
720Hz	700Hz	184.60	59°.78	60°

5.3.3 Reconstruction from all-phase DFT spectrum

One of the major drawbacks of all-phase DFT analysis is the lack of a well-defined synthesis procedure. Although in [35] the signal is reconstructed by simply summing up the sinusoids, after estimating their frequency, amplitude and phase values; it would not be possible to use the same procedure when the input signal has wideband components. The method in [35] is already designed for power systems in which the signal of interest is quite stationary and consists of the sum of harmonic signals. However, when the input signal is speech, the spectral characteristics change rapidly in time. Even in the voiced parts of the speech, there are small fragments of wideband components.

In this section, the proposed synthesis structure for the all-phase DFT will be explained. Before going into the details of the proposed structure, some manipulations on the all-phase DFT analysis equations will be made. After shortening the notation, proposed synthesis structure will be explained.

Consider the i^{th} $2N-1$ point input frame $\boldsymbol{x_i^m}$, centered at the point $x[m]$;

$$\boldsymbol{x_i^m} = \begin{bmatrix} x[m-N+1] \\ x[m-N+2] \\ \vdots \\ x[m-1] \\ x[m] \\ x[m+1] \\ \vdots \\ x[m+N-2] \\ x[m+N-1] \end{bmatrix} = \begin{bmatrix} \boldsymbol{H_{i-1}^m} \\ x[m] \\ \boldsymbol{H_i^m} \end{bmatrix} \tag{5.32}$$

The reason for naming the $N-1$ point data blocks as $\boldsymbol{H_{i-1}^m}$ and $\boldsymbol{H_i^m}$ will be clear in the following steps. Next, let the $2N-1$ point triangular window function $\boldsymbol{w}$ be defined as follows;

$$\boldsymbol{w} = \frac{1}{N}\begin{bmatrix} 1 & 2 & \dots & (N-1) & N & (N-1) & \dots & 2 & 1 \end{bmatrix}^T = \begin{bmatrix} \boldsymbol{w_u} \\ 1 \\ \boldsymbol{w_d} \end{bmatrix} \tag{5.33}$$

With the above definitions of the i^{th} input frame and the window function, equation-(5.30) can be written in the following form;

$$x_{ap}[n,i] = \begin{bmatrix} x[n] \\ \boldsymbol{H_{i-1}^n} \otimes \boldsymbol{w_u} + \boldsymbol{H_i^n} \otimes \boldsymbol{w_d} \end{bmatrix} \tag{5.34}$$

The notation is changed to $x_{ap}[n,i]$ to indicate that the sequence is generated using the i^{th} frame. The operator '$\otimes$' is used to indicate 'elementwise multiplication', i.e.;

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \otimes \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} a_1b_1 \\ a_2b_2 \\ a_3b_3 \end{bmatrix} \tag{5.35}$$

If the $2N-1$ point frames are taken with $N-1$ point overlaps, the frames will have the structure in Figure-5.9. Notice that when the specified overlap amount is used, consecutive data frames will have the following property;

$$\boldsymbol{H_i^{iN}} = \boldsymbol{H_i^{(i+1)N}} \triangleq \boldsymbol{H_i} \tag{5.36}$$

$\boldsymbol{H_i^n}$ is defined in (5.32). Using the notation in (5.32) and the property in (5.36) the structure can be reduced to the form shown in Figure-5.10.

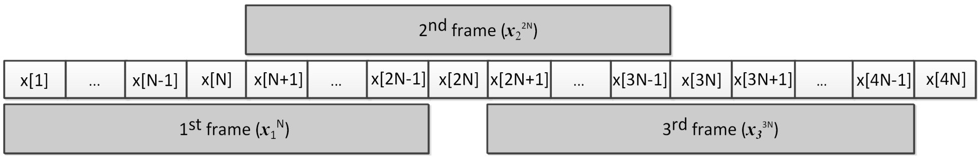


Figure 5.9: (2N-1) point frames with (N-1) point overlap

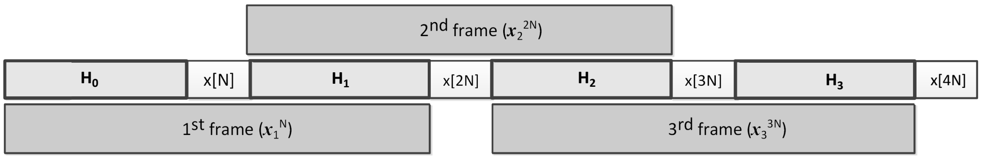


Figure 5.10: Renaming of the data blocks in Figure-5.9

As a result, when $2N - 1$ point frames with $N - 1$ point overlap are used, the all-phase DFT sequences will have the following form;

$$\boldsymbol{x_{ap}}[iN, i] = \begin{bmatrix} x[iN] \\ \boldsymbol{H_{i-1}^{iN}} \otimes \boldsymbol{w_u} + \boldsymbol{H_i^{iN}} \otimes \boldsymbol{w_d} \end{bmatrix} = \begin{bmatrix} x[iN] \\ \boldsymbol{H_{i-1}} \otimes \boldsymbol{w_u} + \boldsymbol{H_i} \otimes \boldsymbol{w_d} \end{bmatrix} \tag{5.37}$$

Where $\otimes$ symbol represents elementwise multiplication, as explained in equation-(5.35). After obtaining $x_{ap}[iN, i]$, the all-phase DFT output is obtained by taking the DFT of these sequences;

$$\boldsymbol{X_{ap}}[k, i] = \mathcal{DFT}\{\boldsymbol{x_{ap}}[iN, i]\} \tag{5.38}$$

The problem arises when one needs to reconstruct the signal using all-phase DFT outputs, namely $X_{ap}[k, i]$. The inverse DFT will yield $x_{ap}[iN, i]$ sequences in time domain and it is clear that these sequences are the time-aliased form of the input signal. To solve this aliasing problem the following algorithm can be implemented;

- Take the inverse DFT of the given frequency domain signals, $X_{ap}[k, i]$;
  $\boldsymbol{x_{ap}}[iN, i] = IDFT\{X_{ap}[k, i]\}$

- Assume that the block H_0 is known, and calculate H_1 as follows;

$$\begin{bmatrix} x[N] \\ \boldsymbol{H_1} \end{bmatrix} = \left(\boldsymbol{x_{ap}}[N,1] - \begin{bmatrix} 0 \\ \boldsymbol{H_0} \otimes \boldsymbol{w_u} \end{bmatrix} \right) \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_d} \end{bmatrix} \tag{5.39}$$

$$= \begin{bmatrix} x[N] \\ \boldsymbol{H_1} \otimes \boldsymbol{w_d} \end{bmatrix} \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_d} \end{bmatrix} \tag{5.40}$$

$$= \begin{bmatrix} x[N] \\ \boldsymbol{H_1} \end{bmatrix} \tag{5.41}$$

Notice that the symbols $\otimes$ and $\oslash$ are used to indicate 'elementwise multiplication' and'elementwise division' respectively. Also $\boldsymbol{x_{ap}}[N,k]$ is defined in equation-(5.37).

- After calculating $\boldsymbol{H_1}$, the procedure can be continued in a similar way; as $\boldsymbol{H_1}$ and $\boldsymbol{x_{ap}}[N,2]$ are enough to calculate $\boldsymbol{H_1}$.
- For mathematical completeness, the calculation of $\boldsymbol{H_k}$ with the use of $\boldsymbol{H_{k-1}}$ and $\boldsymbol{x_{ap}}[N,k]$ is as follows;

$$\begin{bmatrix} x[kN] \\ \boldsymbol{H_k} \end{bmatrix} = \left(\boldsymbol{x_{ap}}[kN,k] - \begin{bmatrix} 0 \\ \boldsymbol{H_{k-1}} \otimes \boldsymbol{w_u} \end{bmatrix} \right) \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_d} \end{bmatrix} \tag{5.42}$$

$$= \begin{bmatrix} x[kN] \\ \boldsymbol{H_k} \otimes \boldsymbol{w_d} \end{bmatrix} \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_d} \end{bmatrix} \tag{5.43}$$

$$= \begin{bmatrix} x[kN] \\ \boldsymbol{H_k} \end{bmatrix} \tag{5.44}$$

Where $\otimes$ symbol represents elementwise multiplication, as explained in equation-(5.35). With the above algorithm $\boldsymbol{H_k}$ and $x[kN]$ can be evaluated and the original signal can be formed as X_{rec} =[$\boldsymbol{H_0}$ x[0] $\boldsymbol{H_1}$ x[1] $\boldsymbol{H_k}$ $\ldots$].

However, there are two major problems with the algorithm defined above. One of these problems is the availability of $\boldsymbol{H_0}$. It might be possible to know $\boldsymbol{H_0}$ by imposing a condition to not to modify the first frame and simply using the data acquired in the first frame. Another option might be to add a frame of zeros in front of the signal. By doing so, the elements of $\boldsymbol{H}_0$ are known to be zeros.

The second problem with the above algorithm is the accumulating error. In a standard DFT-IDFT procedure there will be a small amount of numerical error. For instance, using 'double' precision in Matlab, if the FFT of a sequence is calculated and then the inverse FFT of the result is subtracted from the original sequence, the maximum difference between these two sequences will be in the order of 10^{-15} to 10^{-14}. This error can be ignored in many cases since it does not increase in time. However, in the algorithm defined above, since the previous frame is used to calculate the current frame, the numerical error caused by IDFT procedure is accumulated. Unfortunately, the elementwise divisions with the window functions amplify the error in each step since the elements of the window function are smaller than one. The accumulated error can be observed for a test signal (a sinusoid) in Figure-5.11. The frame length is used as 1000 points for better visualization.

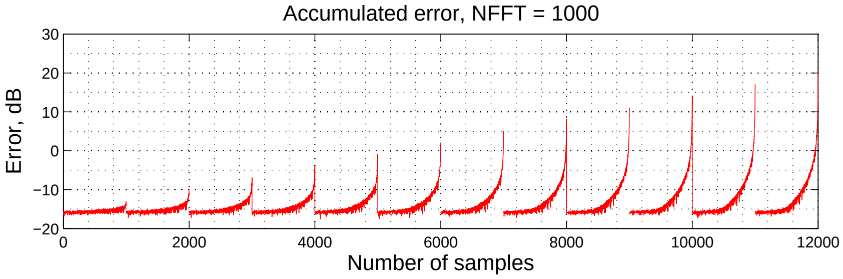


Figure 5.11: Difference between the original and reconstructed signals (Reconstruction started from the first frame)

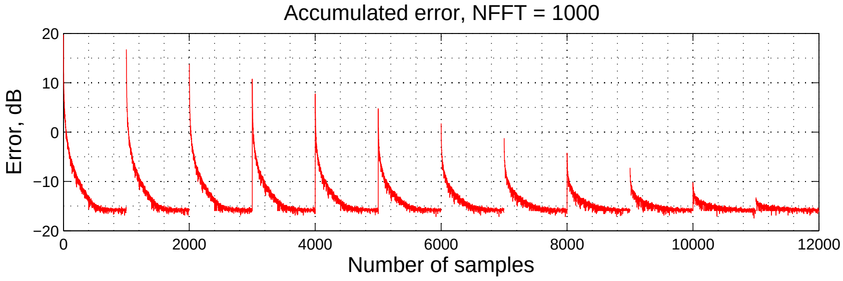


Figure 5.12: Difference between the original and reconstructed signals (Reconstruction started from the last frame)

It is clearly seen in Figure-5.11 that the error is amplified in the second half of the frame.

Taking this fact into account, the algorithm can be executed backwards, i.e. instead of starting from $\boldsymbol{H_0}$, one can start from the last frame $\boldsymbol{H_M}$ and go backwards, finding $\boldsymbol{H_{M-1}}$, then $\boldsymbol{H_{M-2}}$ and so on. This 'backwards' process can be formulated as follows;

$$\begin{bmatrix} x[MN] \\ \boldsymbol{H_{M-1}} \end{bmatrix} = \left(\boldsymbol{x_{ap}}[MN, M] - \begin{bmatrix} 0 \\ \boldsymbol{H_M} \otimes \boldsymbol{w_d} \end{bmatrix} \right) \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_u} \end{bmatrix} \tag{5.45}$$

$$= \begin{bmatrix} x[MN] \\ \boldsymbol{H_{M-1}} \otimes \boldsymbol{w_u} \end{bmatrix} \oslash \begin{bmatrix} 1 \\ \boldsymbol{w_u} \end{bmatrix} \tag{5.46}$$

$$= \begin{bmatrix} x[MN] \\ \boldsymbol{H_{M-1}} \end{bmatrix} \tag{5.47}$$

In this case the error between the original signal and the reconstructed signal will be as in Figure-5.12 and is just the symmetric version of Figure-5.11. This time the error is amplified in the first half of each frame.

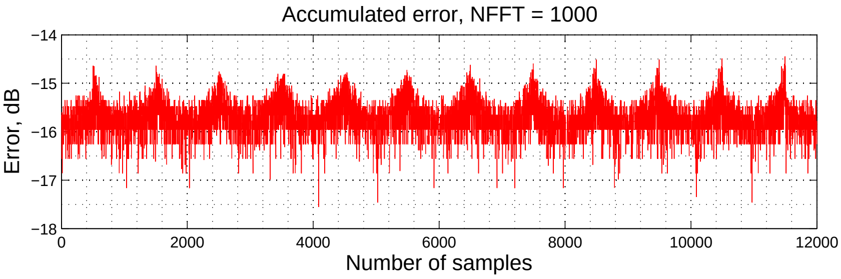


Figure 5.13: Accumulated error, when two approaches are combined

A much better result can be obtained by combining the results of the two aforementioned methods. Using the defined algorithm the first halves of the frames are retrieved with only the numerical error of the FFT process. Running the algorithm starting from the last frame, the second halves of the frames are retrieved in the same fashion. Applying this procedure, the error between the original signal and the reconstructed signal becomes as in Figure-5.13.

It is at last possible to reconstruct the signal with an acceptable degree of distortion. The method can not be directly applied in real time, as a backwards reconstruction procedure is required. However it might be possible to implement it in real time with some delay if short durations of input signals are used.

Although the proposed reconstruction scheme is capable of perfect reconstruction, when some modifications are done on the phase spectrum, reconstructed signal suffers from some additional distortions. For instance, consider a single sinusoid whose frequency is not an integer multiple of the frequency resolution of the system($\frac{f_s}{N}$). In this case a modification in the phase spectrum needs to be extended to more than one frequency bin, as most of the signal energy will be divided between more than one frequency bin. Actually to avoid any kind of amplitude modulation, the modifications must be extended to the entire spectrum even for a single sinusoid. Also the modifications in the phase spectra together with the modifications in the magnitude spectra may correspond to a filtering operation in time domain which is equivalent to a convolution operation. Due to this convolution operation, resulting signal is longer than N-points; however the frequency domain modifications and N-point inverse DFT will result in an N-point sequence and the obtained sequence will be aliased. This problem also existent in STFT based enhancement applications. In these applications the problem is solved by padding N zeros to the end of the N-point analysis frame. By doing so, even if the modifications correspond to an N-tap filter, the output signal will be a $3N - 1$ point long sequence whose last N-points are zeros. As a result, aliasing will not occur in the first N-points of the output. After the modifications and $2N$-point inverse DFT operations the last N-points of the output is simply discarded. The same approach can be used in all-phase DFT reconstruction process to avoid such problems.

5.4 Pitch estimation

In this section, some of the classical pitch estimation methods are presented. A more detailed overview on the subject can be found in [36].

5.4.1 Overview

Pitch determination is one of the fundamental problems in speech processing and a major research topic. There are many algorithms proposed for this problem. However in the context of this work, only the basic methodology is explained here.

5.4.2 Time domain waveform similarity methods

The most obvious feature of a periodic signal is, by definition, the similarity between different segments of it. Hence a time domain self-similarity measure can be employed for pitch detection purposes. One of the most popular self-similarity measure is the auto-correlation function (ACF), which is defined for a stationary signal x(t) as follows;

$$r_x(\tau) = \int_{-\infty}^{\infty} x(t)x(t+\tau)dt \tag{5.48}$$

As its name implies the auto-correlation function (ACF) is a similarity measure. As a result the ACF must take its maximum value for zero lag ($r_x(0)$), since the similarity is maximized when two of the signals are the same. If the signal is periodic, then it means that it will repeat itself with a certain frequency. As a result the signals x(t) and x($t + nT_0$) will be the same for all integer values of n, for a periodic signal with period T_0. In this case the auto-correlation function takes maximum values as well for the lags of length (nT_0). Hence for a periodic signal with period T_0, the following observation is true;

$$r_x(nT_0) = r_x(0), \qquad \forall n \epsilon \mathbb{Z} \tag{5.49}$$

If there are no global maxima except $\tau = 0$, then there might still be some local maxima. In this case, if the highest of the local maxima, say $r_x(\tau_{max})$, is large enough (comparable to $r_x(0)$) then the signal is said to have a periodic part.

The signal of interest for pitch detection, is generally speech and it is not a stationary signal for long durations. Hence the short-term auto-correlation function makes more sense for speech signals, which is defined as follows;

$$R_x[\tau] = \sum_{n=0}^{N-1} s[n]s[n+\tau] \tag{5.50}$$

The value of τ that maximizes $R_x(\tau)$ would yield the pitch period of the periodic signal. Another time domain similarity measure is the so called average magnitude difference function (AMDF) defined as;

$$E(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} |s[n] - s[n-\tau]| \tag{5.51}$$

It is clearly seen that the AMDF measures the dissimilarity unlike ACF which measures similarity. Because of this propoerty AMDF is also referred as anti-correlation function and to find the pitch period one needs to search for the minimum values instead of peaks.

Aforementioned methods are two of the most fundamental methods for pitch detection. Their performance significantly degrades under noise. Hence more sophisticated algorithms are needed for better pitch estimates under noise. There are many algorithms proposed for this purpose (See [36] for a review). The pitch detection block in the proposed method (Chapter-5) is implemented by using the pitch output of PRAAT software [37] which uses an autocorrelation based pitch detection algorithm [38].

CHAPTER 6

VALIDATION AND IMPLEMENTATION OF THE ALGORITHM

In this chapter, the algorithm proposed in Chapter-5 is tested with a clean speech signal to show that speech signals possess the aforementioned property. The phase distortion introduced by the additive noise is also illustrated using a noisy input. The implementation of the proposed method (in Chapter-5) and obtained results are presented.

6.1 Validation of the proposed structure

Based on the development in Chapter-5, the phase difference between two consecutive frames (of the frequency bin of interest) is expected to be $2\pi\frac{f_0}{f_s}N$, where the input sinusoid has the frequency f_0 and the observation window is N-point long. This is expected to be true, if the input signal is stationary during the observation window or if f_0 estimate represents the average value of the fundamental frequency over the analysis frame.

When the input signal is narrowband speech (f_s=8kHz), the analysis problem becomes more difficult as the harmonics are closely packed in the spectrum and the effect of spectral leakage starts to dominate. Nevertheless, the signal retains the same property, as the following simulations indicate.

An ‘a-C-a’ word (a-consonant-a) is used as input in the first simulation, spoken by a female speaker. Such words (aCa, VCV (vowel-consonant-vowel) or CVC etc.) are commonly used in intelligibility tests, in order to prevent the listener using his/her vocabulary (training data) to fill in the unperceived parts, as these words are generally meaningless. Such tests are called ‘nonsense syllable tests’ and they are first introduced by Fletcher and Steinberg [39]. The selection of this ‘a-C-a’ word in the first simulation has a completely different reason however. These words are spoken in a rather prolonged and calm manner, hence making it a very stationary signal over the voiced segments. As a result, this input signal can almost be considered as the ‘best case scenario’.

Figure-6.1 illustrates the characteristics of the clean input signal.

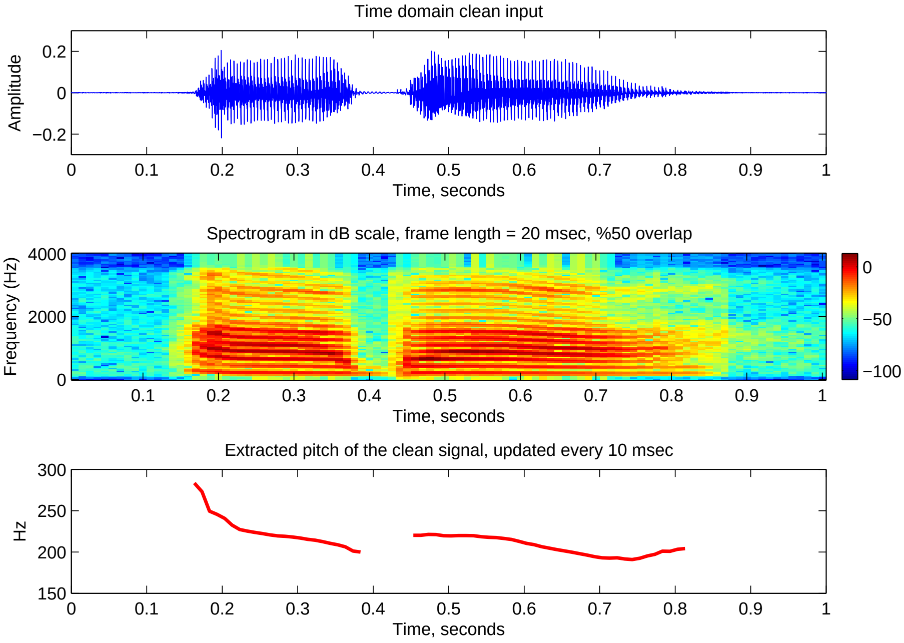


Figure 6.1: Time domain signal, spectrogram and extracted pitch of the word ‘a-b-a’, spoken by a female speaker

The pitch of the signal is extracted using the Praat software [37]. When the output of the pitch estimation algorithm is not in the range of 75-400 Hz the output is simply discarded. Using this pitch estimate and equation-5.4, ($\Delta\theta = 2\pi\frac{f_0}{f_s}N$), the phase difference between consecutive frames is calculated and sketched in Figure-6.2. The phase differences of the consecutive frames (of the frequency bins corresponding to f_0) are also drawn on the same plot for comparison. All-phase DFT analysis is used instead of traditional DFT as it has better side-lobe suppression. One other reason is the fact that the phase value obtained from the traditional DFT will be a function of the difference between the fundamental frequency and the corresponding frequency bin. Since this difference will change in each frame, so does the phase estimate of DFT. As a result the difference between consecutive phase estimates will deviate more than it will if all-phase analysis were used.

As seen in Figure-6.2, the estimated phase differences almost coincide with the analysis results. The small differences between the estimated values and the analysis results arise because of the following reasons;

- As the pitch estimates clearly indicate, the fundamental frequency is not constant and the estimate does not exactly correspond to the average value of f_0 over the analysis frame.
- Spectral leakage problem introduces some amount of distortion to phase estimates.
- The pitch estimate is not 100% accurate, hence this estimation error may also cause some deviations.

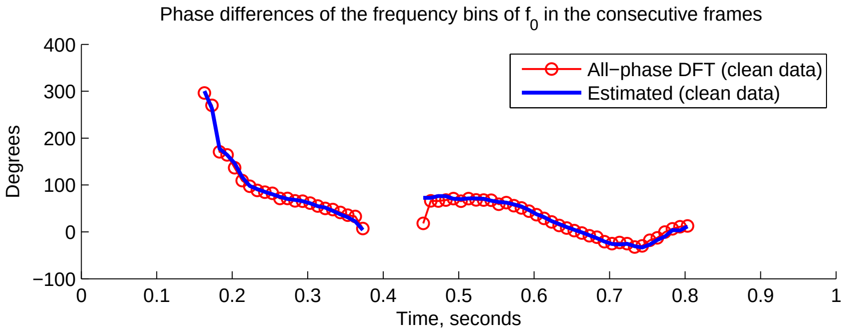


Figure 6.2: Comparison of the all-phase DFT analysis result and the proposed estimation

The analysis can be extended to include the harmonic components as well. By multiplying the extracted pitch with the harmonic number, the frequency of the k^{th} harmonic can be estimated. Hence the phase-difference of the k^{th} harmonic, between consecutive frames can be estimated as well.

To observe the phase distortion introduced by additive noise, the previous analysis is repeated by adding a white Gaussian noise (using Matlab) to the clean signal such that the SNR is approximately 3 dB. The estimation results on the noisy signal as well as the analysis results of the clean signal are demonstrated in Figure-6.3.

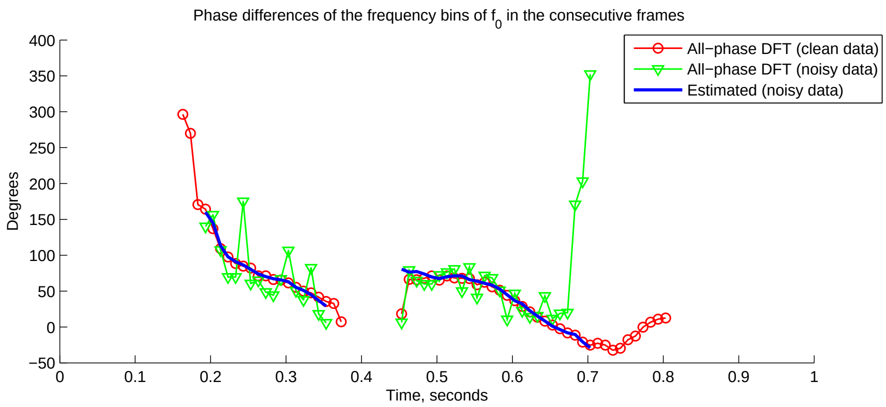


Figure 6.3: Comparison of the all-phase DFT analysis result and the proposed estimation

6.2 Implementation of the proposed structure

The proposed structure in Chapter-5 is implemented in MATLAB© environment. All-phase DFT is employed for phase estimation. The pitch exraction is executed using the Praat software. The details of the used methods are given in Chapter-5.3 and [37], [38]. The implemented version of the proposed system can be seen in Figure-6.4.

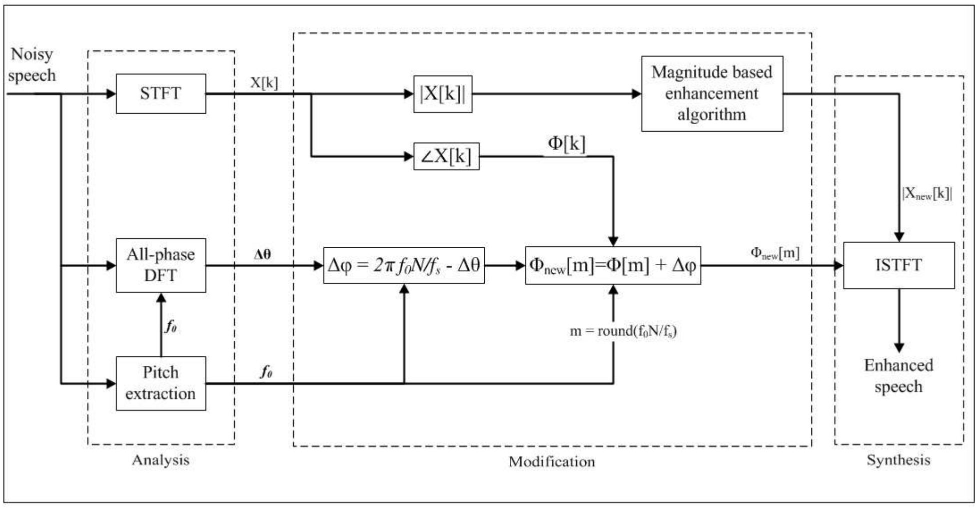


Figure 6.4: Implemented version of the proposed method

Notice that instead of STFT, all-phase DFT could have been used. However, when some modifications are done in all-phase DFT spectra, the proposed reconstruction method in Chapter-

5.3 becomes very unstable due to the time-aliasing and accumulating error problems. Hence, a more stable and well defined method (STFT) is used instead of all-phase DFT analysis in spectral analysis block. To exploit the advantages of all-phase DFT over STFT, the all-phase DFT analysis is run in parallel with STFT for phase estimation purposes. By doing so, the algorithm benefits from the better phase estimates of all-phase DFT without introducing distortions in the reconstruction process.

The algorithm is implemented with 3 different magnitude based algorithms, namely; MMSE [16], log-MMSE [25] and spectral subtraction using oversubtraction [18].

6.2.1 Parameters of analysis blocks

It is important to synchronize analysis blocks, as the proposed method combines the outputs of these blocks. To achieve this, the following parameters are selected for analysis blocks;

- 2N point STFT length with N point overlap.
- 2N-1 point all-phase DFT length with (N-1) point overlap.
- Pitch estimate update at every $(kN)^{th}$ point for k=1,2,3 . . ..

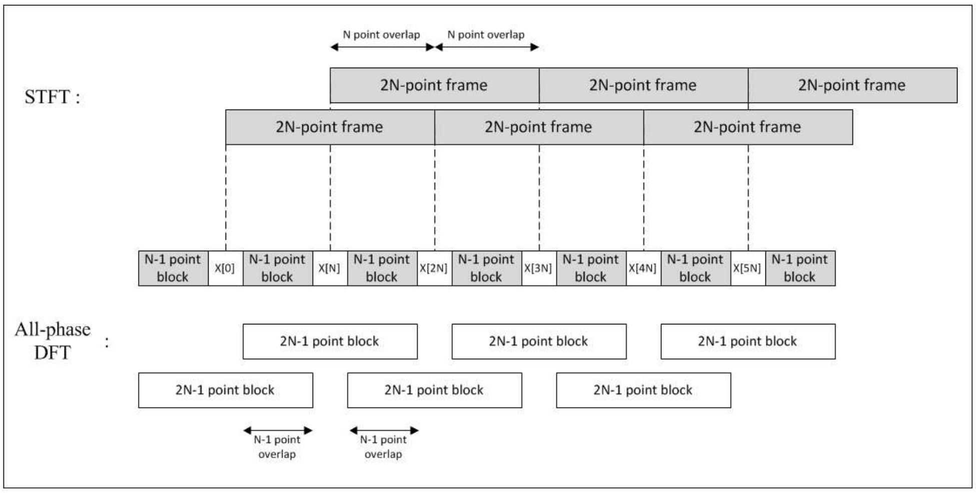


Figure 6.5: Frame lengths and overlap ratios for STFT and all-phase DFT

The analysis blocks use the frames shown in Figure-6.5, with the above parameters. Notice that all-phase DFT output phase corresponds to the instantaneous phase at the middle point of the frame. Hence the STFT frames are aligned to the middle points of the all-phase DFT frames.

6.2.2 Phase modification

If pitch estimation block is able to output reasonable estimates (between 75-400 Hz range) for two consecutive analysis frames, phase modification process is triggered for the latter frame. Using the extracted pitch, the corresponding frequency bin of the fundamental frequency is identified for STFT and all-phase DFT as follows;

$$m_{STFT} = round\left(\frac{f_0}{f_s}2N\right) \tag{6.1}$$

$$m_{apDFT} = round\left(\frac{f_0}{f_s}N\right) \tag{6.2}$$

Although all-phase DFT uses $2N - 1$ point long data, it combines these points to form an N-point dataset and outputs the DFT of this N-point dataset (see Chapter-5.3 for details). As a result, the frequency resolution of $2N - 1$ point all-phase DFT is $\frac{N}{f_s}$ and $2N$ point STFT is $\frac{2N}{f_s}$.

After calculating the frequency bins of the extracted pitch for each frame, the phase difference between the frames are evaluated using all-phase DFT and the deviation from the expected phase difference value, $2\pi\frac{f_0}{f_s}N$, is calculated. This error signal must be subtracted from the phase of the frequency bins of interest in STFT spectrum. As explained in Chapter-5.3 the phase of the DFT coefficients is in the form of equation-6.3, for a single complex exponential $(e^{j}(2\pi\frac{f_0}{f_s}n + \phi_0))$;

$$\angle X[k] = \phi_0 + \frac{\pi}{f_s}\left(f_0 - k\frac{f_s}{N}\right)(N - 1) \tag{6.3}$$

This result was obtained for a rectangular window. The STFT analysis on the other hand is carried out using Hamming window, to benefit from its better side-lobe suppression and better phase estimation. The phase of the DFT coefficients give the following expression for a single complex exponential when Hamming window is used;

$$X[k] = \sum_{n=0}^{N-1} e^{j\left(2\pi \frac{f_0}{f_s} n + \phi_0\right)} e^{-j\frac{2\pi}{N}kn} w[n], \qquad w[n] = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \tag{6.4}$$

$$= e^{j\phi_0} \sum_{n=0}^{N-1} e^{j2\pi n\left(\frac{f_0}{f_s} - \frac{k}{N}k\right)} w[n] \tag{6.5}$$

$$= e^{j\phi_0} \boldsymbol{R}(N, f_0, f_s, k) e^{j\boldsymbol{\theta}(N, f_0, f_s, k)} \tag{6.6}$$

$$\angle X[k] = \phi_0 + \theta(N, f_0, f_s, k) \tag{6.7}$$

It is hard to analytically contain $\angle X[k]$ for Hamming windowed complex exponential. Fortunately this is not needed. As seen in equation-6.7, the phase is equal to the actual value (ϕ_0) plus a function of other parameters (N, f_0, f_s,). As a result; if the window length N, signal frequency (f_0) and sampling frequency (f_s) are unchanged, then the phase of the tonal signal can be changed by adding the desired offset value to the phase of the related frequency bins. And this is exactly what is tried to be done in the proposed algorithm.

Although the signal of interest is a single tone, due to spectral leakage effect (see Chapter-5.3) its energy spreads over the entire spectrum. Hence phase corrections must be extended to more than one frequency bin. The modification must also be limited to those frequency bins where the signal of interest dominates, in order not to distort the phase of the other strong components in the neighboring frequency bins. In other words, phase modification should not be extended to the bins where the energy spread of the interested component becomes negligible.

The window length for STFT in the implemented structure was selected as 160 samples at 8 kHz sampling rate ($\equiv 20msec$), with Hamming window. In this case the phase modification is done at the frequency bin of the tone of interest and its neighboring bins as well; because of the reason explained above. To justify this operation, consider the DFT of 160 point Hamming window which is shown in Figure-6.6, with discrete pulses for better visualization.

It is clearly observable that the energy of a single pulse will spread among the entire spectrum, however most of the energy (more than %99.99) will be contained in the 3 frequency bins. As a result, utilization of the aforementioned phase modification to the center frequency bin and two neighboring bins seems an acceptable approximation.

After the phase modification is done on the target bins, the signal is reconstructed using the

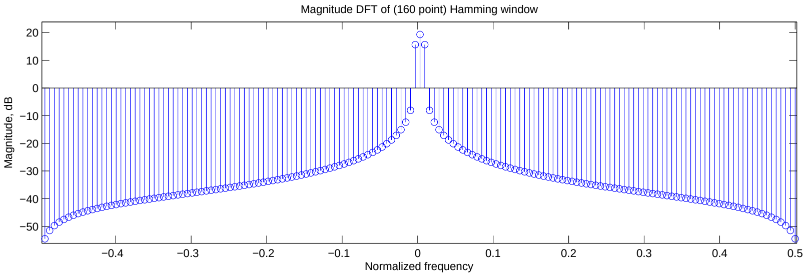


Figure 6.6: Magnitude response of Hamming window

modified phase spectra and the modified magnitude spectra (using a spectral subtraction based enhancement algorithm). The synthesis procedure is carried out with the overlap-add (OLA) method, as explained in Chapter-2.

6.2.3 Test results

The implemented structure is tested using the NOIZEUS database ([27], [28]). The database consists of 30 sentences, spoken by 3 male and 3 female speakers. The audio-files are degraded with 8 different colored noises at 4 different SNR levels. In addition to these SNR levels (0dB, 5dB, 10dB and 15dB) -5dB level is also generated, by subtracting the clean signals from their degraded versions and amplifying the residual with an appropriate constant. Then adding these amplified noise signals to the clean signals the new set of audio files are obtained. Also another noise type (white Gaussian) is artificially generated and added to clean audio signals, to extend the database with a stationary noise type as well. As a result, 9 different noise-degraded audio files became available in the test database as seen in Table-6.1.

As seen in Table-6.1, the database includes highly non-stationary noise types (airport, babble, exhibition, restaurant). Only the artificially generated white Gaussian and in a weaker sense car noise can be considered as stationary.

As mentioned before the algorithm is implemented using 3 different magnitude based enhancement algorithms for magnitude modification, namely; MMSE [16], log-MMSE [25] and spectral subtraction using oversubtraction [18]. The new algorithm is named for each of the different method used, as ‘phase modified MMSE’, ‘phase modified log-MMSE’ and ‘phase modified spectral subtraction’. The PESQ scores (see Chapter-2) of these classical

Table 6.1: Noise types in the database

Noise type
Airport
Babble
Car
Exhibition
Restaurant
Station
Street
Train
White Gaussian

(magnitude based) algorithms are evaluated in parallel with the PESQ scores of the phase modified versions (the proposed method) of these algorithms for 30 different noise degraded speech signals. The difference between these PESQ scores (Proposed - classical) are plotted and labeled as ‘improvement’. Figures 6.7 to 6.15 show the improvement of the proposed method over the classical magnitude based methods. The improvements in these figures are obtained by correcting only the phase of the fundamental component (f_0). The algorithms are also tested when phase of more than one harmonic components are corrected. Table-6.2 and Table-6.3 summarize the average quality improvements of 30 sentences for different noise types and different magnitude based enhancement algorithms.

The results indicate significant improvements especially for phase modified spectral subtraction over the classical spectral subtraction. Considering the fact that about 0.1 PESQ difference can be differentiated by the listener, even on the average the performance of the spectral subtraction algorithm is improved about a perceivable amount for each of the noise types. For phase modified MMSE and phase modified log-MMSE significant improvements are obtained for certain noise types. The performance is degraded in many cases due to poor pitch estimates. These estimates generally fail to differentiate the fundamental component and higher harmonics and sudden jumps are observed in the extracted pitch. As a result the desired continuity of phase values can not be maintained in such sections. As a future work, additional constraints can be imposed on the extracted pitch to maintain its smoothness. Also a pitch tracking algorithm can be implemented for further improvements.

The improvements are more prominent when signal to noise ratio (SNR) is low. This can be explained by the trade-off between phase correction and the induced distortion while modifying the phase spectra. As explained before, the phase modification is done for certain frequency bins with an approximation that causes an amplitude modulation in the reconstructed signal. Hence for high SNR values, the effect of phase correction for a single tone is dominated by the effect of the induced distortion. As a result the algorithm works better for low SNR values. As a future work, phase modification scheme can be revised.

Table 6.2: Average PESQ improvements of 30 sentences (only the phase of the fundamental component (f_0) is corrected)

Noise type	**SNR**	**Phase modified MMSE**	**Phase modified logMMSE**	**Phase modified spectral subtraction**
Airport	-5 dB	-0.0168	0.0222	0.0521
Babble	-5 dB	-0.0040	-0.0256	0.0267
Car	-5 dB	0.0067	0.0057	0.0382
Exhibition	-5 dB	0.0170	0.0184	0.0317
Restaurant	-5 dB	0.0247	0.0075	0.0305
Station	-5 dB	0.0132	0.0068	0.0412
Street	-5 dB	-0.0073	0.0028	0.0321
Train	-5 dB	-0.0016	-0.0207	0.0455
WGN	-5 dB	0.0635	0.0627	0.0386
Airport	0 dB	0.0008	-0.0097	0.0342
Babble	0 dB	0.0032	-0.0093	0.0336
Car	0 dB	0.0204	0.0046	0.0361
Exhibition	0 dB	-0.0044	0.0015	0.0258
Restaurant	0 dB	-0.0052	0.0246	0.0167
Station	0 dB	0.0111	-0.0060	0.0438
Street	0 dB	0.0029	0.0093	0.0245
Train	0 dB	0.0089	-0.0016	0.0363
WGN	0 dB	0.0271	-0.0005	0.0365

Table 6.3: Average PESQ improvements of 30 sentences (phase of the fundamental component (f_0) and second harmonic ($2f_0$) are corrected)

Noise type	**SNR**	**Phase modified MMSE**	**Phase modified logMMSE**	**Phase modified spectral subtraction**
Airport	-5 dB	-0.0555	-0.0339	0.0458
Babble	-5 dB	-0.0285	-0.0133	0.0439
Car	-5 dB	0.0160	0.0162	0.0368
Exhibition	-5 dB	0.0124	-0.0118	0.0398
Restaurant	-5 dB	0.0514	-0.0174	0.0149
Station	-5 dB	0.0155	-0.0065	0.0301
Street	-5 dB	0.0089	0.0259	0.0314
Train	-5 dB	-0.0008	-0.0118	0.0397
WGN	-5 dB	0.0654	0.0610	0.0431
Airport	0 dB	-0.0156	-0.0253	0.0283
Babble	0 dB	-0.0101	-0.0260	0.0321
Car	0 dB	0.0245	0.0091	0.0367
Exhibition	0 dB	-0.0194	-0.0266	0.0273
Restaurant	0 dB	-0.0395	-0.0716	0.0011
Station	0 dB	0.0087	-0.0208	0.0441
Street	0 dB	-0.0107	-0.0021	0.0257
Train	0 dB	-0.0018	-0.0096	0.0327
WGN	0 dB	0.0177	-0.0096	0.0290

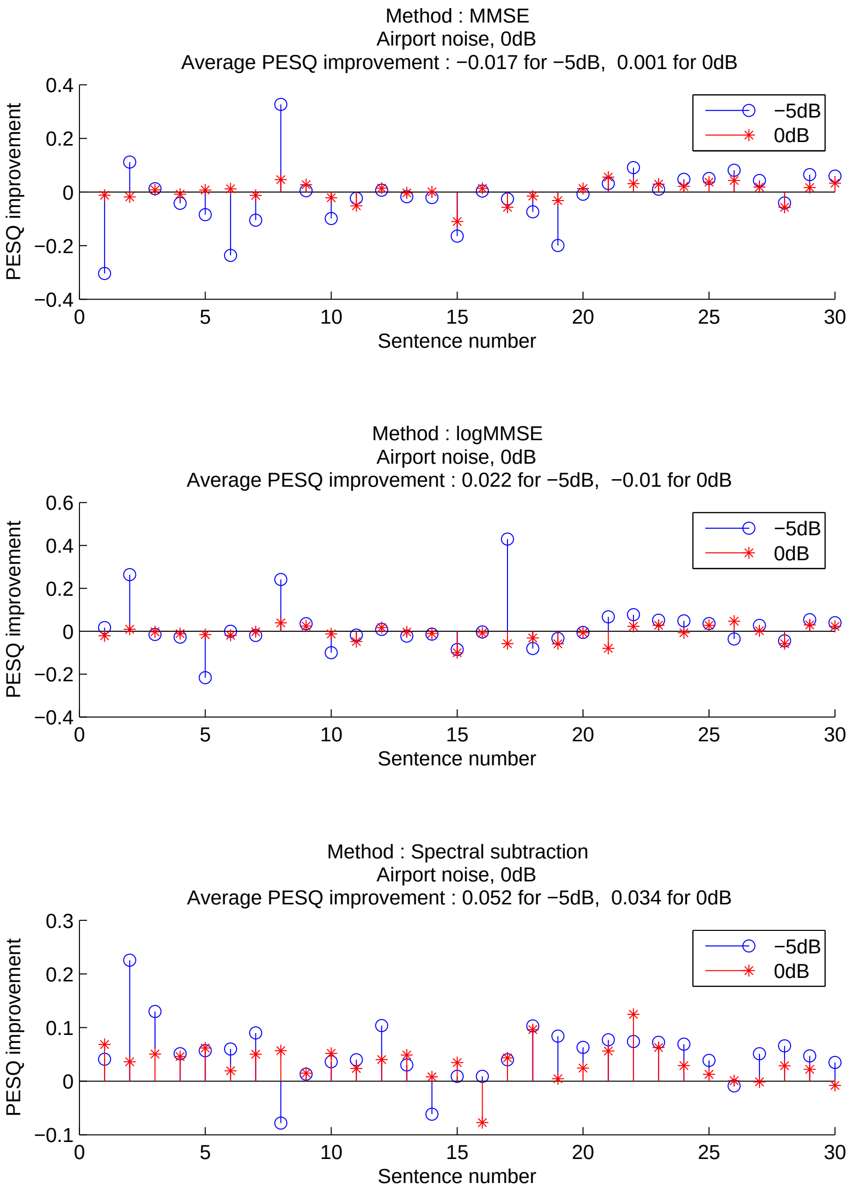


Figure 6.7: Quality improvements under airport noise, only the phase of (f_0) is corrected

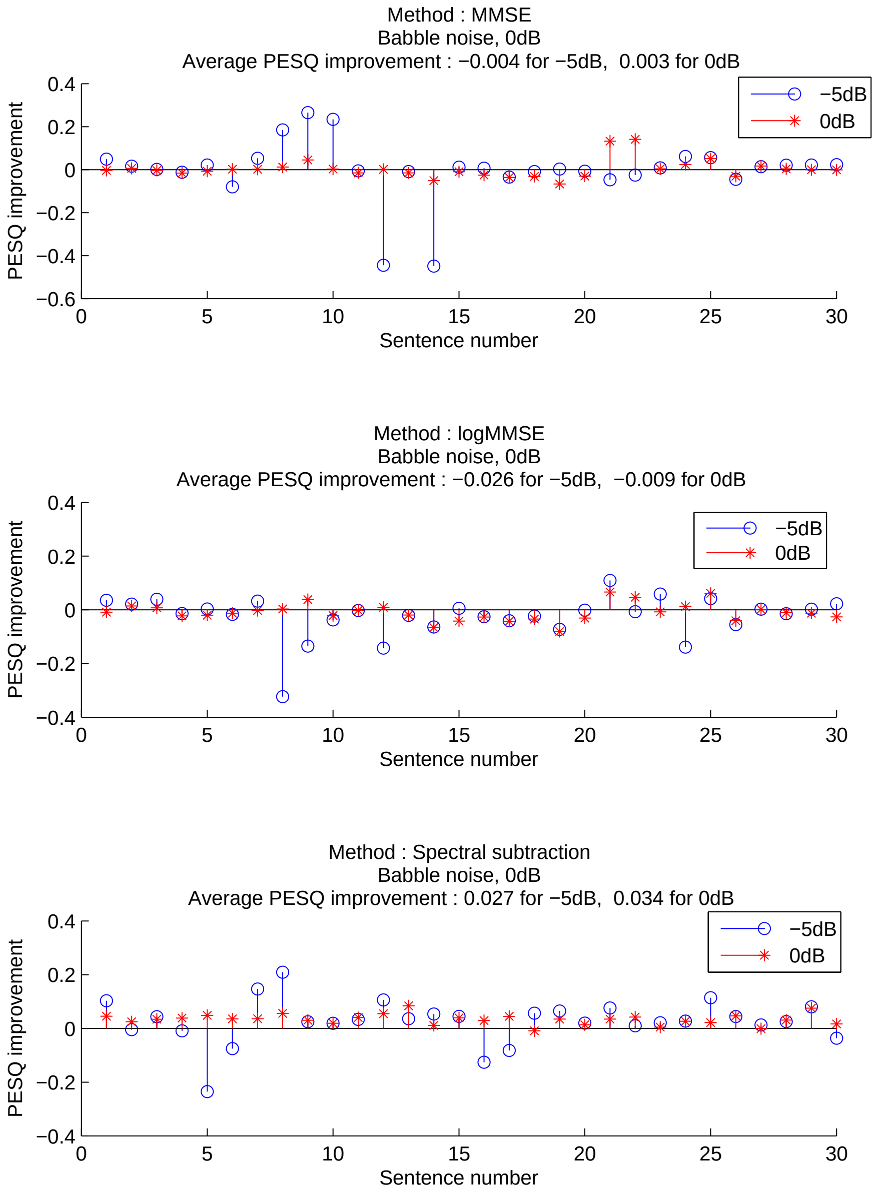
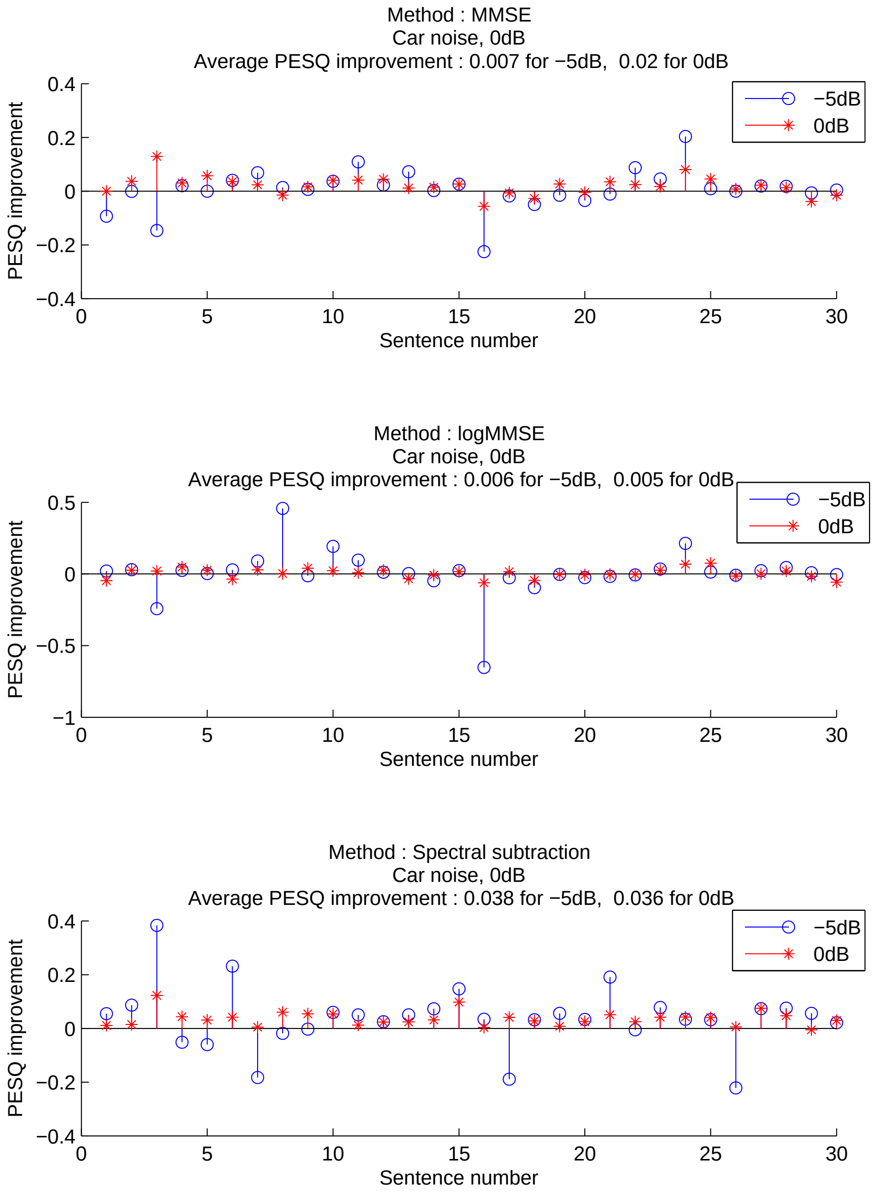


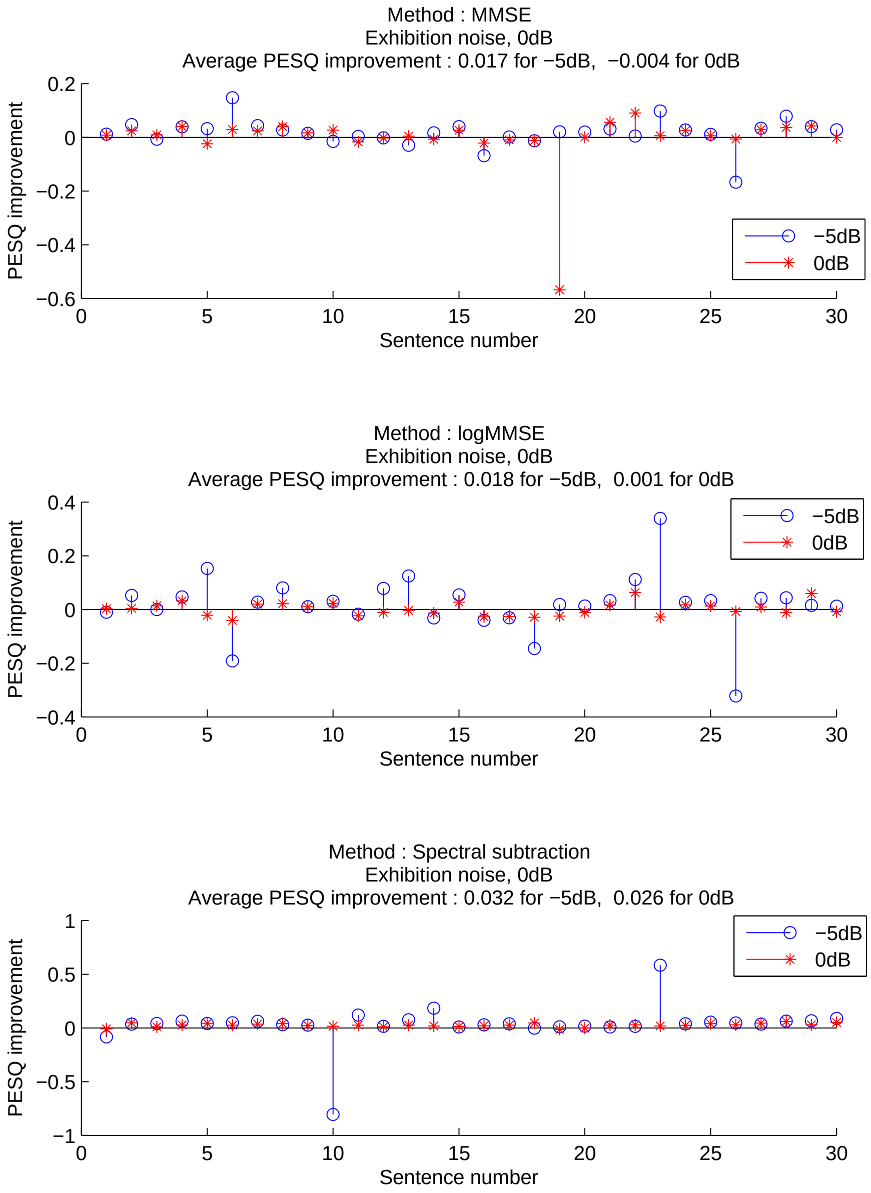
Figure 6.8: Quality improvements under babble noise, only the phase of (f_0) is corrected


Method : MMSE
Car noise, 0dB
Average PESQ improvement : 0.007 for −5dB, 0.02 for 0dB

Method : logMMSE
Car noise, 0dB
Average PESQ improvement : 0.006 for −5dB, 0.005 for 0dB

Method : Spectral subtraction
Car noise, 0dB
Average PESQ improvement : 0.038 for −5dB, 0.036 for 0dB

Figure 6.9: Quality improvements under car noise, only the phase of (f_0) is corrected



Method : MMSE
Exhibition noise, 0dB
Average PESQ improvement : 0.017 for −5dB, −0.004 for 0dB

Method : logMMSE
Exhibition noise, 0dB
Average PESQ improvement : 0.018 for −5dB, 0.001 for 0dB

Method : Spectral subtraction
Exhibition noise, 0dB
Average PESQ improvement : 0.032 for −5dB, 0.026 for 0dB

Figure 6.10: Quality improvements under exhibition noise, only the phase of (f_0) is corrected

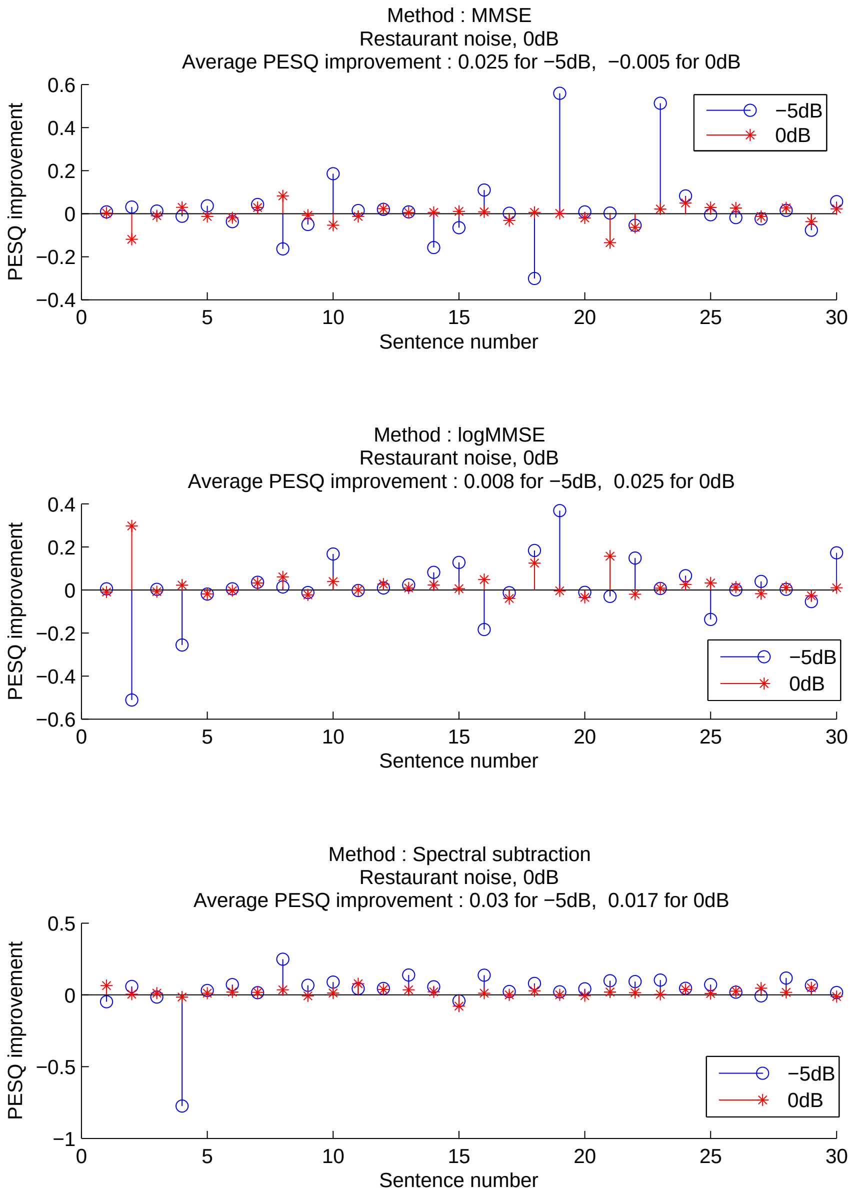


Figure 6.11: Quality improvements under restaurant noise, only the phase of (f_0) is corrected

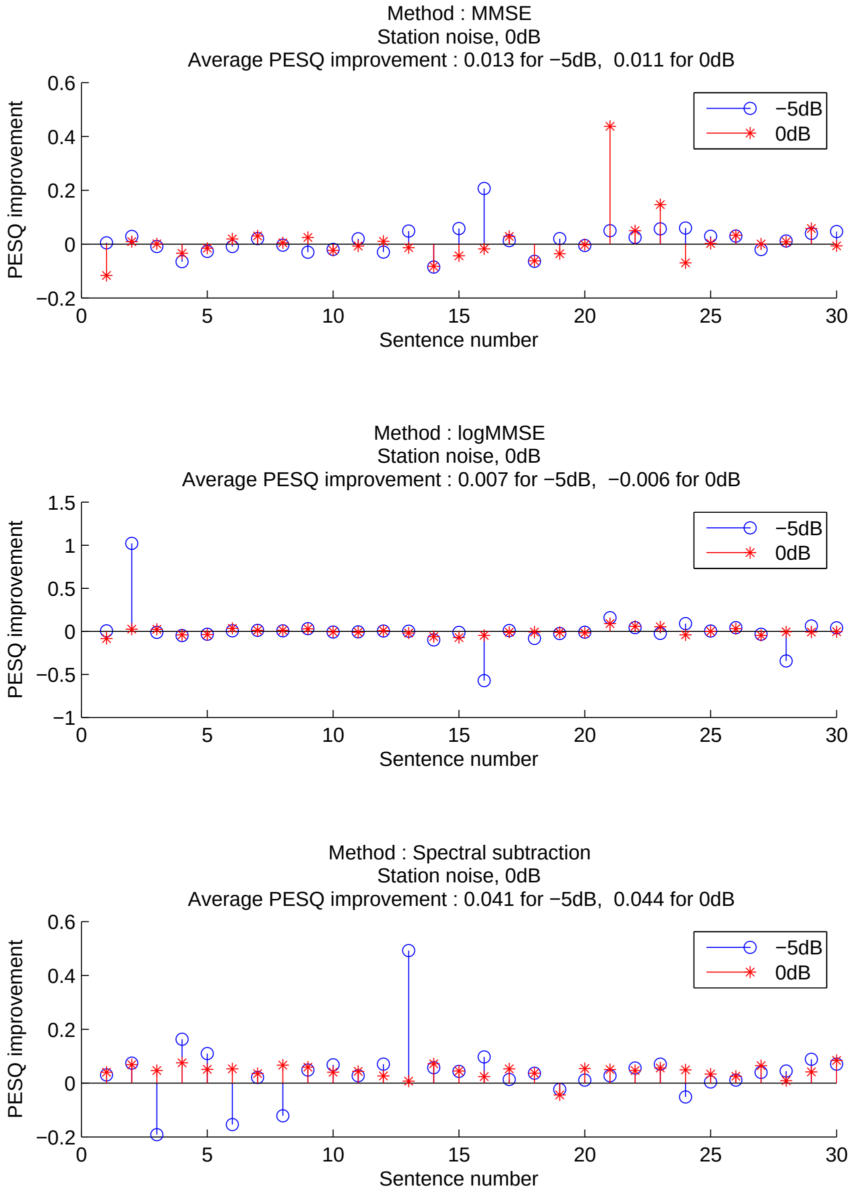


Figure 6.12: Quality improvements under station noise, only the phase of (f_0) is corrected

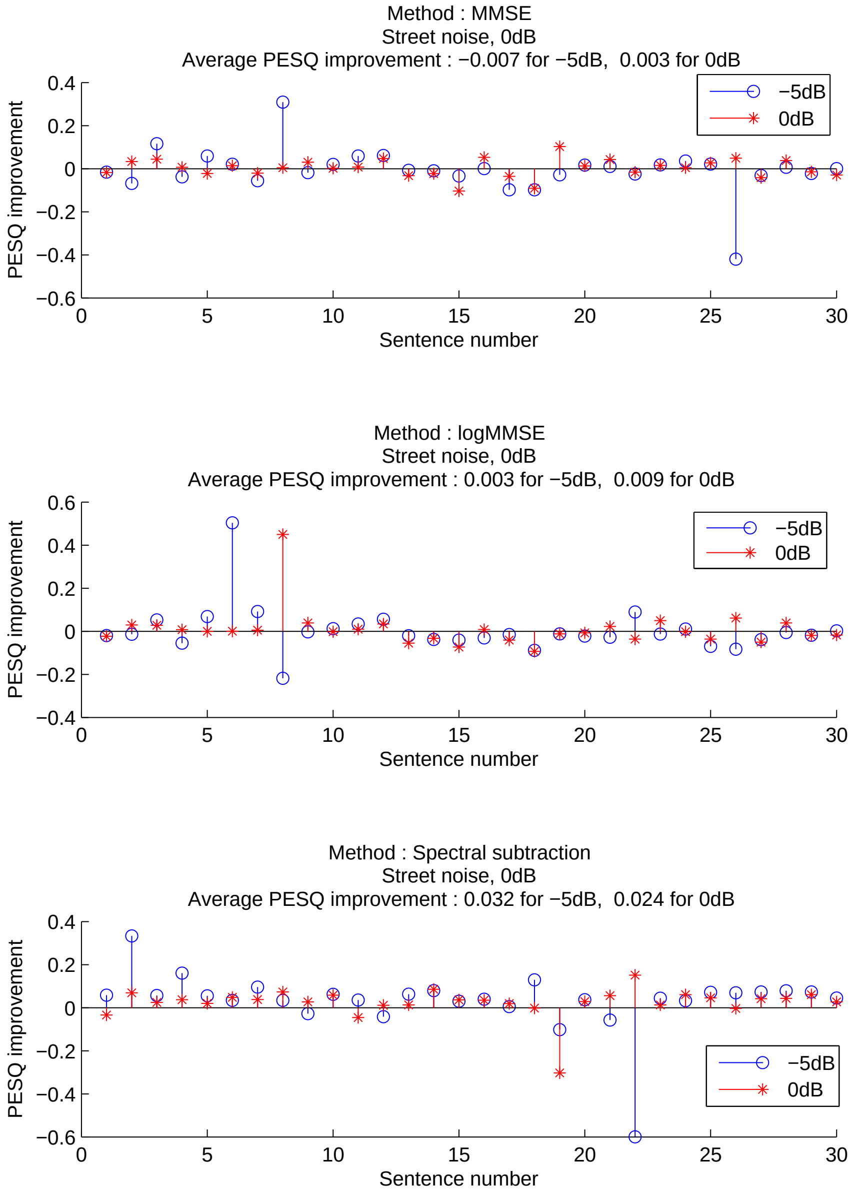


Figure 6.13: Quality improvements under street noise, only the phase of (f_0) is corrected

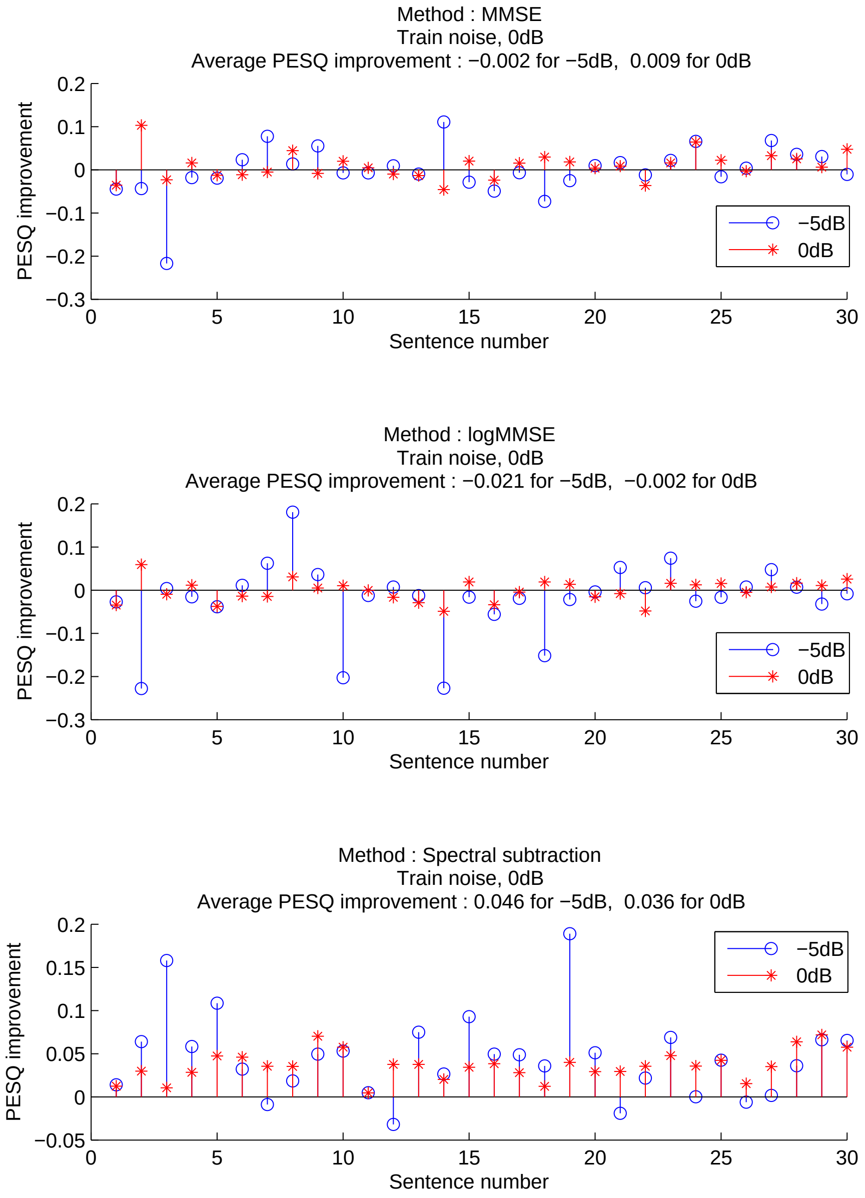


Figure 6.14: Quality improvements under train noise, only the phase of (f_0) is corrected

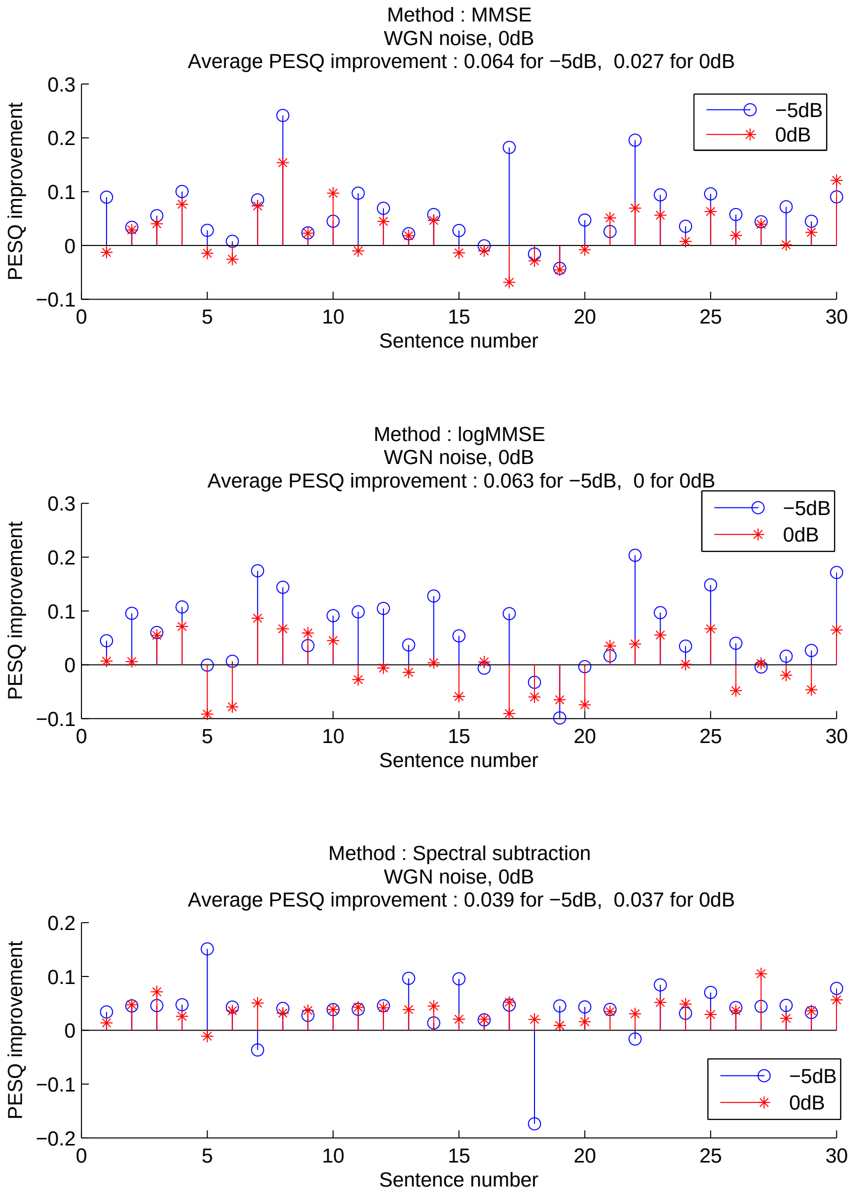


Figure 6.15: Quality improvements under white Gaussian noise, only the phase of (f_0) is corrected

CHAPTER 7

CONCLUSION

7.1 Summary of the thesis

In this thesis study, it is shown that the tonal components in voiced parts of clean speech signals possess an important property, namely the phase continuity. It is also shown that the phase of these tonal components can be predicted with the use of phase continuity assumption and estimated frequency values. Then using this prediction technique, a novel method is presented for single channel narrowband speech enhancement. The method can be applied to wideband speech without any modifications as there are no constraints imposed on the input signal. The proposed method maintains the phase continuity of certain tonal components in voiced parts of the input speech by modifying the DFT phase spectra; unlike most of the classical enhancement algorithms which only modify the magnitude spectra. The proposed method makes use of the detected average fundamental frequency to estimate the phase of the corresponding DFT coefficient. The proposed algorithm uses the classical enhancement algorithms to modify the magnitude spectra in addition to the conducted phase corrections. As the implementation results indicate, the proposed system improves the performance of the classical methods, in terms of speech quality.

It is important to note that the proposed method makes no assumptions about the noise statistics, in the phase modification block.

It is also worth mentioning that the phase estimation block of the proposed structure is implemented by using a recently developed method, namely the ‘all-phase DFT analysis’. Furthermore, a signal reconstruction scheme is presented for all-phase DFT analysis.

The performance improvement of the proposed system is strongly dependent on the quality of pitch detection and signal to noise ratio of the input signal. The estimated fundamental frequency must be a smooth function as it is in clean speech. When SNR is too low, or the noise is highly non-stationary (e.g. bable noise) the quality of pitch estimation degrades. Hence there is a lower bound on the signal SNR for the proposed algorithm to work properly. When SNR is high, there is not much phase distortion on the tonal components and the utilized approximation while modifying the phase spectra causes additional distortions. As a result there is also an upper limit on the signal SNR, as expected. In the conducted tests, it is observed that for SNR values less than or equal to 0 dB, the algorithm produces better results than the classical methods.

7.2 Future work

The performance of the proposed enhancement algorithm strongly depends on the quality of the pitch detection algorithm and input SNR. To improve the performance of the proposed method additional algorithms should be designed to check and interpret the degree of smoothness of the detected pitch and input SNR. Furthermore, the phase modification scheme can be improved as the applied approximation introduces amplitude modulations to the reconstructed signal. The all-phase DFT method is not used in the modification and synthesis procedures of the implemented algorithm, to avoid accumulating error problem. This problem can be investigated and all-phase DFT analysis can be used for spectral modification as well. To do so, classical methods should be derived for all-phase DFT transform.

REFERENCES

[1] R.W. Schafer and L.R. Rabiner. Digital representations of speech signals. *Proceedings of the IEEE*, 63(4):662 – 677, april 1975.

[2] Xiangdong Huang, Zhaohua Wang, Limian Ren, Yifang Zeng, and Xiaoyan Ruan. A novel high-accuracy digitalized measuring phase method. In *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, pages 120 –123, oct. 2008.

[3] P.C. Loizou. *Speech enhancement: Theory and Practice*. Signal processing and communications. CRC Press, 2007.

[4] D. Wang and Jae Lim. The unimportance of phase in speech enhancement. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 30(4):679– 681, August 1982.

[5] P.C. Loizou and Gibak Kim. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):47 –56, jan. 2011.

[6] ITU-T Recommendation P.862. Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, 2000.

[7] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing (3rd Edition)*. Prentice Hall, 3 edition, August 2009.

[8] Rainer Martin, Ulrich Heute, and Christiane Antweiler. *Advances in Digital Speech Transmission*. Wiley Publishing, 2008.

[9] X. Maitre. 7 khz audio coding within 64 kbit/s. *Selected Areas in Communications, IEEE Journal on*, 6(2):283 –298, feb 1988.

[10] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.

[11] Petre Stoica and Randolph L. Moses. *Introduction to Spectral Analysis*. Prentice-Hall, New Jersey, 1997.

[12] Thomas Quatieri. *Discrete-time speech signal processing: principles and practice*. Prentice Hall Press, Upper Saddle River, NJ, USA, first edition, 2001.

[13] J.B. Allen and L.R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558 – 1564, nov. 1977.

[14] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, volume 2, pages 749 –752 vol.2, 2001.

[15] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113 – 120, apr 1979.

[16] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 32(6):1109 – 1121, dec 1984.

[17] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *Speech and Audio Processing, IEEE Transactions on*, 9(5):504 – 512, jul 2001.

[18] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '79.*, volume 4, pages 208 – 211, apr 1979.

[19] Boh Lim Sim, Yit Chow Tong, J.S. Chang, and Chin Tuan Tan. A parametric formulation of the generalized spectral subtraction method. *Speech and Audio Processing, IEEE Transactions on*, 6(4):328 –337, jul 1998.

[20] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Companies, 3rd edition, February 1991.

[21] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. *Noise Reduction in Speech Processing*. Springer Publishing Company, Incorporated, 1st edition, 2009.

[22] Charles W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 1992.

[23] Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1996.

[24] K. Wojcicki, M. Milacic, A. Stark, J. Lyons, and K. Paliwal. Exploiting conjugate symmetry of the short-time fourier spectrum for speech enhancement. *Signal Processing Letters, IEEE*, 15:461 –464, 2008.

[25] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):443 – 445, apr 1985.

[26] Kuldip K. Paliwal and Leigh Alsteris. Usefulness of phase spectrum in human speech perception. In *Proc. Eurospeech*, pages 2117–2120, 2003.

[27] Yi Hu and P.C. Loizou. Subjective comparison of speech enhancement algorithms. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, page I, may 2006.

[28] NOIZEUS database. http://www.utdallas.edu/ loizou/speech/noizeus/ , last accessed on 01.08.2011.

[29] IEEE recommended practice for speech quality measurements. *Audio and Electroacoustics, IEEE Transactions on*, 17(3):225 – 246, sep 1969.

[30] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(4):744 – 754, aug 1986.

[31] Matteo Frigo and Steven G. Johnson. The design and implementation of FFTW3. *Proceedings of the IEEE*, 93(2):216–231, 2005. Special issue on “Program Generation, Optimization, and Platform Adaptation”.

[32] FFTW web page. http://www.fftw.org, last accessed on 01.08.2011.

[33] Jian-Zhong Liu, Zheng-Xin Hou, and Cheng-You Wang. Windowed all-phase dft modulated in time domain and its application in spectrum analysis. In *Wireless Communications, Networking and Mobile Computing, 2009. WiCom ’09. 5th International Conference on*, pages 1 –4, sept. 2009.

[34] D. Kang, X. Ming, and Z. Xiaofei. Phase difference correction method for phase and frequency in spectral analysis. *Mechanical Systems and Signal Processing*, 14(5):835–843, September 2000.

[35] Xiangdong Huang, Haitao Cui, Zhaohua Wang, and Yifang Zeng. Mechanical fault diagnosis based on all-phase FFT parameters estimation. In *Signal Processing (ICSP), 2010 IEEE 10th International Conference on*, pages 176 –179, oct. 2010.

[36] A. Kindoz and A. M. Kondoz. *Digital Speech; Coding for Low Bit Rate Communication Systems*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[37] Paul Boersma and David Weenink. Praat: doing phonetics by computer [computer program]. version 5.2.35, http://www.praat.org/, 2011.

[38] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences 17: 97-110, University of Amsterdam*.

[39] H. Fletcher. Articulation testing methods. *Journal of The Acoustical Society of America*, 1, 1930.