ENHANCEMENT OF THE CODED SPEECH USING FILTERING

SALİH SİNAN TAYLAN

IŞIK UNIVERSITY
2017

ENHANCEMENT OF THE CODED SPEECH USING FILTERING

SALİH SİNAN TAYLAN
B.S., Electronics Engineering, IŞIK UNIVERSITY, 2014

Submitted to the Graduate School of Science and Engineering
in partial fulfillment of the requirements for the degree of
Master of Science
in
Electronics Engineering

IŞIK UNIVERSITY
2017

IŞIK UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

ENHANCEMENT OF THE CODED SPEECH USING FILTERING

SALİH SİNAN TAYLAN

APPROVED BY:

Assoc. Prof. Ümit Güz (Thesis Supervisor)	Işık University	_____
Assoc. Prof. Hakan Gurkan (Thesis Co-advisor)	Işık University	_____
Assist. Prof. Sakip Önder	Işık University	_____
Prof. Dr. Serdar Özoğuz	Istanbul Technical University	_____

APPROVAL DATE: 14/04/2017

ENHANCEMENT OF THE CODED SPEECH USING FILTERING

Abstract

The processing and storage of speech signals are widely implemented in modern communication systems. Decreasing the amount of information for modeling the reconstruction of speech signal enhances the transmission and storage capacity of the system.

It is important to compress speech without losing its important properties during transmission or reconstruction independently from the speaker and speech signals itself. However, some losses inevitably occur in every compression process. Increasing the compression ratio results in increased losses. Speech enhancement algorithms may be used to enhance strongly compressed speech signals for better intelligibility and quality. The purpose of this study is to enhance speech with healing algorithms that compress speech signals while reducing background noise.

The SYMPES [1][2][4] algorithm used in this study compresses data resulting in lesser loss than other known compression algorithms. As a result of the compression, noise occurs in the background. The type of the noise cannot be classified. Attempts have been made to reduce these background noises (distortions) by using different methods of speech enhancement algorithms. More than ten speech enhancement algorithms have been investigated and implemented. Two algorithms with the best-enhanced sound output were determined and compared. One of them, Spectral Subtraction Algorithm, was applied via a geometric approach, which was investigated in 2008 by Yang Lu and Philipos C. Loizou [3].In this algorithm, a noise spectrum is subtracted from the noisy speech signal and then a clean signal spectrum is obtained. Moreover, in the absence of the signal, the noise spectrum can be updated and predicted. This approach expressed that the noise spectrum is not significantly different between update periods and is a noisy cum stationary or slowly changing process. Forward and inverse Fourier transforms are used in the algorithm; hence, the algorithm is quite simple. However, the simple subtraction algorithm is a costly operation. Subtraction must be done with extreme caution to avoid any speech distortion.

If too many subtractions are made, some speech information may be removed from the center; if too little is subtracted, it can be observed that a clear majority of the intervening noises are still present. The other speech enhancement method is a statistical model based algorithm. This statistical speech enhancement method involves predicting the statistic of a clean and noisy signal for a sample. In other words, if a speech signal is distorted with a statistically independent noise, the marginal probability distributions of the clean speech and noise signal must be clearly known. In this model-based statistical method, signal and noise statistics are estimated primarily from the speech and noise content. An optimal solution is obtained using statistical models and it is then used in conjunction with distortion measures to solve the existing speech enhancement problem. In this approach, different techniques have been applied to parameterize speech signals such as autoregressive moving average (ARMA), autoregressive (AR), or moving average (MA). Three prediction rules known as the maximum probability (ML), maximum posterior (MAP), and minimum mean square error (MMSE) are used in this approach and have many desirable features to estimate the parameters of the speech signal. ML is used for the maintenance of non-random parameters. The estimation methods MAP and MMSE are used for known parameters of the previously known density function, which can be examined in advance as a random variable. For the speech signal, this model uses the MAP estimation approach, assuming a time-varying AR model for speech enhancement in which both the model and signal are estimated from the noisy signal.

However, since the waveform of the speech signal is distorted as a result of the signal improvement, the SNR results are not found very healthy. Therefore, the results are evaluated by the Mean Opinion Score (MOS) test. A subjective test based on MOS is also carried out on some selected utterances. The results of the subjective test are also compared with those of the objective test to determine the most appropriate objective measure for the evaluation of speech enhancement algorithms. The strengths and weaknesses of the various algorithms are analyzed and compared. Quality has been shown in detailed graphs that can be measured and smoothed using the MOS, which defines the quality of speech by a listener on a scale of 1 to 5.

Keywords: Speech enhancement, spectral subtraction, statistical model based

FİLTRELEME KULLANARAK KODLANMIŞ SESİN İYİLEŞTİRİLMESİ

Özet

Bu çalışma konuşma işaretini sıkıştırıp, arka plan da yer alan gürültünün indirgenmesini sağlayan iyileştirme algoritmaları sayesinde arka plandaki gürültü temizlenmesi hedeflenmiştir. Kullanılan sıkıştırma algoritması SYMPES' in temel amacı ifade edilmek istenirse; konuşma işaretlerinin işlenmesi, depolanması modern iletişim sistemlerinde oldukça önem taşımaktadır.

Özellikle konuşma işaretlerinin modellenmesi ya da yeniden oluşturulması sonucunda, gerekli bilgi miktarının azaltılması, sayısal konuşma işaretlerinin depolanmasını ve iletilmesini sağlayan sistemlerin kapasitesi ses verisi sıkıştırması sayesinde belirgin bir şekilde artmaktadır. Ancak bundan dolayı bir takım veri kaybı olmakta ya da arka plan da gürültü oluşmaktadır.

Bu sıkıştırma algoritmalarında temel amaç; konuşma iletiminin ya da konuşmanın yeniden oluşturulmasında konuşma işaretinin kendisinden ve konuşmacıdan bağımsız ve bilginin önemli özelliklerini kaybetmeden yüksek sıkıştırma oranları ile oluşturmasıdır. Bu çalışma da kullanılan SYMPES algoritması da diğer bilinen sıkıştırma algoritmalarına oranla daha az veri kaybı ile sıkıştırma yapmaktadır. Yine de sıkıştırma sonucunda, arka planda birtakım gürültüler olmaktadır. Bu gürültü diğer bir adı ile bozulmalar ses iyileştirme algoritmalarının farklı metotları kullanarak minimize edilmeye çalışılmıştır. Araştırılan bu ses iyileştirme algoritmalarından gürültü kaynığının belli olmadığında arka plan gürültüsü için en sağlıklı sonuçları veren iki algoritma önerilmiş: Spektral Çıkarma Algoritması ve İstatistiksel Tabanlı Model metodudur. Bu metotlar arasındaki karşılaştırmalar yapılmıştır.

Spektral Çıkarma Algoritması'nı özetlemek istersek; ses sinyaline karşılık, ek gürültü düşünüldüğünde, gürültülü ses spektrumundan bir gürültü spektrumu çıkartılarak, temiz bir sinyal spektrumun bir tahminin elde edilir. Gürültü spectrumu yokluğunda sinyal güncellenebilir ve tahmin edilebilmektedir. Bu yaklaşım, gürültü spektrumunun güncelleme dönemleri arasında önemli ölçüde farklı olmadığını ve gürültülü durağan veya yavaş yavaş değişen bir süreç olduğunu özetler

niteliktedir. İleri ve ters Fourier dönüşümleri algoritmada kullanılır. Bu nedenle algoritma oldukça basittir. Basit çıkarma algoritması maliyetli bir işlem olduğudan dolayı çıkarma işlemi, herhangi bir konuşma bozulmasını önlemek için çok dikkatli yapılmalıdır. Çok fazla çıkarma yapılırsa, bazı konuşma bilgileri ortadan kaldırılabilir. Ancak çok az çıkarılırsa, araya giren gürültünün büyük çoğunluğu kalabildiği gözlemlenmiştir. Buna ek olarak, bazı durumlarda spektral çıkarmadan kaynaklanan konuşma bozukluklarının çoğu kaldırılmıştır.

Diğer bir yaklaşım ise istatistiksel model tabanlı algoritmalardır. Bu istatistiksel konuşma geliştirme metodu, temiz ve gürültülü sinyalin ortak istatistiklerini net bir şekilde bilinmesini ve konuşma sinyalleri için algısal bir bozulma önlemi gerektiren belirli bir örnek fonksiyonu için gürültülü bir sinyalin temiz bir sinyalinin tahmin edilmesi yaklaşımıdır.Diğer bir ifadeyle, eğer konuşma sinyalleri istatistiksel olarak bağımsız bir gürültüyle bozulursa, temiz konuşma ve gürültü sinyalinin marjinal olasılık dağılımları açıkça bilinmesi gereklidir.

Bu model tabanlı istatiksel metotta, sinyal ve gürültü istatistikleri öncelikle konuşma ve gürültü içeriğinden tahmin edilir. Optimal çözüm, istatistiksel modeller kullanılarak elde edilir ve daha sonra mevcut konuşma geliştirme problemini çözmek için bozulma önlemleri ile birlikte kullanılır. Bu yaklaşımda, otoregresif hareketli ortalama (ARMA), otoregresif (AR) veya hareketli ortalama (MA) gibi konuşma sinyallerini parametreleştirmek için farklı teknikler bu yaklaşımda uygulanmıştır. Ayrıca, maksimum olasılık (ML), maksimum posteriori (MAP) ve minimum ortalama karesel hata (MMSE) olarak bilinen üç tahmin kuralının, konuşma sinyalinin parametrelerini tahmin etmek için birçok istenen özelliklere sahip olduğu için bu yaklaşımda kullanılmıştır. ML rasgele olmayan parametrelerin bakımı için kullanılmıştır. Tahmin yöntemleri olan MAP ve MMSE, önceden rastgele değişken olarak incelenebilen önceden bilinen yoğunluk fonksiyonunun bilinen parametreleri için kullanılmıştır. Konuşma sinyali için, bu model hem gürültülü sinyalden hem modelin hem de sinyalin tahmin edildiği konuşma geliştirme için zamanla değişen bir AR modeli varsayarak, MAP tahmini yaklaşımı kullanılmıştır.

Bununla birlikte, sinyal gelişiminin sonucu olarak ses sinyalinin dalga biçimi bozulduğundan, SNR sonuçları çok sağlıklı bulunmadığından, elde edilen sonuçlar

Mean Opinion Score (MOS) testi ile değerlendirilir. Bazı seçilmiş konuşmalar üzerinde MOS temelli öznel bir test gerçekleştirilir.

Konuşma geliştirme algoritmalarının değerlendirilmesi için en uygun objektif önlemi belirlemek için öznel testin sonuçları da objektif test ile karşılaştırıldı. Çeşitli algoritmaların güçlü ve zayıf yönleri analiz edilir ve karşılaştırılır. Kalite, bir dinleyicinin konuşmanın kalitesini 1'den 5'e çıkardığı 'Mean Opinion Score' testi (MOS) kullanılarak ölçülebilir ve gürültünün temizlendiğine dair ayrıntılar grafiklerle gösterilmektedir.

Anahtar kelimeler: ses iyileştirme, spektral çıkarma, istatistiksel model

Acknowledgements

Many people make our university extremely valuable because of their studies. Firstly I want to thank my supervisor Assoc. Prof. Ümit Güz who accepted me as a student for guidance and counseling. Having the opportunity to work with him was intellectually rewarding and fulfilling. I also thank Assoc. Prof. Hakan Gürkan, my project co-advisor, who contributed much to the development of this work.

Finally, I am grateful to my family. I thank my father Uluğ Tuğrul Taylan, my mother Güner Taylan, my brother Emre Giray, and my sister Gülizar Gülin for their patience and encouragement.

To my mother

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Speech Enhancement

Speech enhancement involves developing some perceptual aspects of reduced speech with additional noise. The purpose of implementing speech enhancement is to ensure better quality speech with increased intelligibility.

The requirement to improve the quality of speech signals occurs in many cases. For instance, when communication in a channel is interrupted by noise or a listeners sensation level is affected by long-term, high-level noise or when a speech signal is compressed, as in this study. Speech enhancement algorithms sometimes referred to as noise suppression algorithms, reduce or prevent background noise to a certain degree[6][7][8].

Some general areas exist wherein speech enhancement algorithms are employed. Voice communication through cellular telephones processed by preprocessors in speech coding systems employed in mobile phone standards to prevent background noise from the environment, such as cars and restaurants. Furthermore, noise recognition devices nowadays cell phones use speech enhancement algorithms usually on a remote server. Moreover, in an air-ground communication scenario, a pilots speech that has been distorted by exceptionally high levels of cockpit noise must be processed using speech enhancement techniques to increase intelligibility. Similarly, it is desirable to improve the intelligibility of the speech quality in military operations. Finally, hearing impaired listeners who use hearing assistances have great difficulty in communicating in noisy areas, and speech

enhancement algorithms are used to pre-process and thus "clear" the noisy signal before amplification.

It is possible to decrease background noise with speech enhancement algorithms, but it may also distort speech intelligibility as well as induce speech corruption. Therefore, the primary challenge in designing efficient speech enhancement algorithms is to prohibit noise without introducing any obvious distortions in the signal. The performance of speech enhancement algorithms may be affected by the number of microphones that are used. Typically, when the number of microphones increase, speech development tasks become easier. Adaptive noise cancellation techniques can be used when at least one microphone is located near the noise source. This method focuses on speech enhancement signals reduced by statistically uncorrelated additive noise. The enhancement algorithms defined in this work can be used in a variety of noise conditions; thus, the signals are not restricted to any particular noise signal. The purpose of this study is to enhance speech with healing algorithms that compress speech signals and reduce background noise.

One of them, Spectral Subtraction Algorithm, was applied via a geometric approach which was investigated in 2008 by Yang Lu and Philipos C. Loizou[3][6]. In this algorithm, a noise spectrum is subtracted from the noisy speech signal and then a clean signal spectrum is obtained. Moreover, in the absence of signal, the noise spectrum can be updated and predicted. This approach expressed that the noise spectrum is not significantly different between update periods and is a noisy plus stationary or slowly changing process. Forward and inverse Fourier transforms are used in this algorithm; hence, the algorithm is quite simple. Nevertheless, the simple subtraction algorithm is a costly operation; subtraction must be done with extreme caution to avoid any speech distortion. If too many subtractions are made, some speech information can be extracted from the center; if too little is subtracted, it can be observed that a clear majority of the intervening noises are still present.

The other speech enhancement method is a statistical model based algorithm. This statistical speech enhancement method involves predicting the statistic of a clean and noisy signal for a sample. In other words, if a speech signal is distorted with a statistically independent noise, the marginal probability distributions of

the clean speech and noise signal must be explicitly known. In this model-based statistical method, signal and noise statistics are estimated primarily from the speech and noise content. An optimal solution is obtained using statistical models and it is then utilized in conjunction with distortion measures to solve the existing speech enhancement problem.

In general, the intention of this research is to examine, compare, and improve speech enhancement algorithms with respect to quality and comprehensibility.

1.2 Literature Review: Speech Enhancement

Since the 1970s, several single microphone DSP strategies have been put forward in the literature to enhance noisy speech and even eliminate the noises that are possible in the rear. Loizou [6], a valued scientist who uses his work on speech development to a significant extent, has made a detailed study of these algorithms; he explores and presents compressed speech enhancement techniques in two major works. By testing the most appropriate approach, the compressed audio signal is optimally enhanced.

The following two important works are mentioned:

- Spectral Subtraction algorithms [3]
- Statistical model-based algorithms

Furthermore, in addition to these speech enhancement algorithms, the other two algorithms mentioned are also named as follows:

- Wiener filtering algorithms
- Subspace algorithms

Briefly, the Spectral Subtraction Algorithm estimates the background noise spectrum and tries to extract it from the noisy speech frequency spectrum.

Statistical model-based algorithms work on speech enhancement by providing statistical strategies to estimate and improve the speech frequency spectrum. In

Wiener filtering algorithm, the main goal is to search for an optimal filter that decreases the mean square error (MSE) between the output and desired signal. The subspace algorithm dissociates the distorted signal into signal and noise subspaces, separating the noisy subspace as invalid.

Hu and Loizou [3] compare the performance of these different algorithms. Furthermore, there are many more algorithmic variations associated with sound enhancement. For example, the algorithm used in the harmonic properties of speech components. Active noise cancellation techniques are also available depending on the dual microphone multiple microphones (microphone array). In this work, noise cancellation methods have been investigated and applied only for the signal recorded in a single microcontroller.

R. Martin et al.[9] pointed out in their research that the Gaussian statistical model provides a good approximation of DFT coefficients for speech noises. However, this assumption, which is used in mobile communications for speech signals whose specific DFT frame dimensions are too short (10-40 ms), is not fully implementable. Only the approach of R. Martin and colleagues is valid if the DFT frame size is much longer than the correlation interval of the considered signal.

Cohen et al.[10] have been working on methods that cannot take place at all frequencies and at all times. The authors estimated the likelihood that speech will not be seen in an individual frequency coefficient. In this approach, under the Laplace model, the minimum mean square error (MMSE) dimension estimator and the ambiguity of speech presence were defined and also considered as two case models for speech conditions.

R. Martin et al.[9] suggest an estimator based on the real and imaginary parts of the noisy signal observed, where the real and imaginary parts of the clean signal in the MMSE are estimated. Nevertheless, this predictor is modeled with a combination of clean signal and noise, Gaussian, Gamma and Laplace distributions, although the optimum spectral amplitude estimate is not.

C. Breithaupt et al.[11] have indicated that the speech coefficients of Laplace and Gamma using the intensities of the real and imaginary parts are greatly modeled. This observation led to the recreation of the most appropriate MMSE short-time spectral amplitude (STSA) estimator parallel to the investigator, but it was found to depend on different models, i.e., Laplace and/or Gamma for more

accurate predictors. However, it is difficult for some people to look for alternative techniques to calculate the MMSE STSA estimator to derive such an estimator.

Malah et al.[9] modeled speech and noise signal spectrum components as independent Gaussian random variables and statistically reproduced the MMSE STSA estimator. The researchers compared the STSA estimator derived from the Wiener estimator and analyzed the performance of the proposed STSA estimator. Researchers have also examined the MMSE STSA estimator under the uncertainty of the presence of a noisy signal.

Y. Ephraim et al.[29] used the STSA estimator to derive speech signals that are minimized by minimizing the mean square error of the log-spectra (i.e., the original STSA and its estimator) and investigated the development of noisy speech. This estimator was additionally compared with the related minimum mean square error STSA estimator reproduced earlier.

Xuchu et al.[13] proposed an improved algorithm (fast noise tracking algorithm) using the MMSE-LSA algorithm. In these approaches, it is observed that it conforms better to individual sound environments than different traditional algorithm from other speech enhancement approaches. The major part of this method generates the exponential estimate, which is updated using the time-frequency correction factors calculated based on the probability corresponding to the speech in each frequency range of the noisy speech spectrum.

Israel Cohen et al. Get.[14] suggested a minimum recursive recursion average (MCRA) for estimating noise signals. The noise estimate is established by taking the average of the spectral power values in the past and using the smoothing parameter set by the signal existence probability in the sub-bands. The noise estimation in this study can be explained through its efficiency regarding computation, its ability to follow up the change in input signal to noise ratio (SNR), and its rapid contribution to noise, which is robust and noise spectrum

Gustafsson et al.[15] proposed to obtain a lower resolution spectrum to handle the first number and divide the existing analysis frame into smaller subframes. The individual spectra in each subframe are then averaged to obtain a lower variance spectrum. Gustafsson et al.[15] suggested using the adaptive exponential averaging to smooth out the gain function over time to account for the second number. Moreover, because of the utilization of the zero-phase gain function to

avoid noncausal filtering, Gustafsson and et al.[15] have proposed the introduction of a linear phase in the gain function

1.3 Aim and Outline of Thesis

1.3.1 Aim of Thesis

Using of SYMPES algorithms for compression of speech output with a noisy signal. This signal is post-processed with speed enhancement algorithms to explain the noise and increase the quality of the speech and intelligibility. This thesis aims to compare several speech enhancement algorithms: Spectral Subtraction Algorithm and Statistical based model to improve the distortion of the output of SYMPES.

1.3.2 Outline of Thesis

In Chapter 2, the speech coding method that is expressed a new systematic procedure for modeling speech signals on "Envelopes and Signature Sequences" as named SYMPES that is defined general method definition and systematic process is expressed step by step. In Chapter 3, the understanding the enemy as named noised and type of noises are clarified. In Chapter 4, the speech enhancement algorithms used in this study are specified and summarized in this section. In chapter 5, the spectral subtraction algorithm is described in detail and explained theoretically. In chapter 6, statistical model based method is given, and the methods in it are mentioned. Furthermore, in estimation theory, necessary of the techniques for obtaining nonlinear estimators; for instance, Bayesian estimators (e.g., MMSE and maximum posterior estimators) and maximum likelihood estimators (ML) are explained. In Chapter 7, briefly, information has been given about subjective listener test that is named Mean Opinion Score Test. Also, the algorithms were compared with used MOS test, and the averages obtained from the test result is explained. Finally, Chapter 8 includes discussion and conclusion of the thesis.

Chapter 2

Speech Compression Modeling

New speech modeling methods under the name SYMPES (A new systematic procedure for modeling speech signals on "Envelopes and Signature Sequences") have been examined and compared to the methods described. It is stated that the method achieves significantly better hearing quality for the same compression rate or better for SYMPES.

In these methods, after examining the signals of the physical properties first, special waveforms were found. These are the best-defined waveforms called Signature Base Functions. Signature Base Functions of speech signals are provided by utilizing the energy suppression feature of principal component analysis (PCA) [19]. PCA responds optimally to the processing used to least squares mean (LMS) regarding minimizing the error. The method has introduced in this research, considerably improved the results of by introducing the concept of Signal Envelope in the representation of speech signals. In this way, the frame signal is shown as the new mathematical form, $X_i \approx C_i E_i S_R$, where $\mathrm{C_i}$ is a real constant called the gain factor, $\mathrm{S_R}$ and $\mathrm{E_K}$ are derived correctly. The previously defined signature set and envelope set, or short names respectively, are named PSS and PES. This method of PSS and PES, which are formed as research results, is independent of the speaker and the spoken language. When the proposed modeling technique is used for communication, the transmission bandwidth is significantly reduced. When this method is used for digital recording, it is beneficial to store a wider area.

2.1 The Method of SYMPES

Two classes of images can be distinguished, analog and digital images. Both types fall into non temporal multimedia type. The speech signal was examined on a frame-by-frame basis.

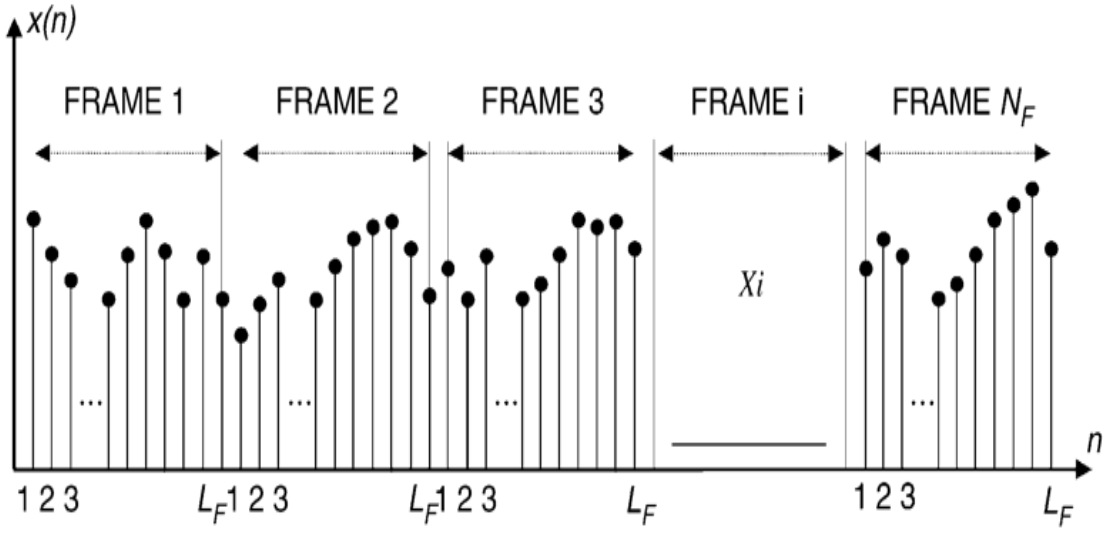

Figure 2.1: Division of speech signals into frames

As shown in Figure 2.1, in this approach represents a selected frame. With reference to to Figure 2.1, at any time for frame i, the sampled speech signal that is given by the vector X_i of length of L_F, can be approximated as $X_i \approx C_i E_i S_R$ where

- The gain factor C_i is denoted as a real constant, $K \in \{1, 2, ..., N_E\}$, $R \in \{1, 2, ..., N_S\}$; K, R, N_E, and N_S are expressed as integers.
- The signature vector $S_T^R = [S_{R1} S_{R2} ... S_{RF}]$ consists of R using the statistical instance of the speech signals and broadly includes the properties of the original frame. Furthermore, it has been shown that the amount quantity $C_i S_R$ carries almost max X_i energy, which means LMS.
- The E_K is (L_F by L_F) is a diagonal matrix as described as follows:

$$E_K = diag[e_{K1} e_{K2} \dots e_{KL_F}] \tag{2.1}$$

- E_K is defined as an envelope term on the quantity $C_i S_R$.

- L_F as an integer which gives the total number of samples in the ith frame.

A long sampled speech signal sequence *x(n)* is written as:

$$x(n) = \sum_{i=1}^{N} x_i \delta_i(n-i) \tag{2.2}$$

In the above equation, *i(n)* represents the unit example, and x_i describes the width of the sequence *x(n)* for the *ith* sample. The boundary condition is thought to be $N \to \infty$ for very long parts of speech, *x(n)* can also be determined in vector notation.

$$X^T = [x(1)\, x(2)\, x(3)\, ...\, x_(N)] = [x_1\, x_2\, ...\, x_N] \tag{2.3}$$

X have is named the Main Frame Vector and it is has divided frames with equal lengths., , such as 16, 24, 64, or 128 samples, etc. in this representation. The Main Frame Matrix that is indicated by M_F is get obtained by means of the frame vectors.

$$M_F = [X_1\, X_2\, X_3\, \cdots\, X_{N_F}] \tag{2.4}$$

where

$$X_i = \begin{bmatrix} x_{(i-1)}L_F + 1 \\ \vdots \\ x_i L_F \end{bmatrix} \tag{2.5}$$

$N_F = N/N_F$ is denoted by the total number of frames in X. Specifically, the integers N and L_F are chosen to be an integer in N_F. It is stated that in a vector space formed by orthonormal vectors $\{\phi_{ik}\}$, each frame vector or sequence X_i can be spanned. It is explicitly stated that $X_i = \sum_{k=1}^{L_F} c_k \phi_{ik}$ and the frames coefficients are acquired as $c_k = \phi_{ik}^T X_i$, *k=1, 2, 3,...*L_F. Let it be $1 \leq l \leq L_F$ so that $X_{il} = \sum_{k=1}^{l} c_k \phi_{ik}$ is a truncated version of X_i. Then, the approximate error (ϵ_l) is dedicated by $(\epsilon_l) = X_i - X_{il} = \sum_{k=1}^{L_F} c_k \phi_{ik}$.

Here, ϕ_{ik} is defined by decreasing the expected value of the error , which is expressed expresses the LMS giving the following eigenvalue problem [19].

$$R_i\phi_{ik} = \lambda_k\phi_{ik} \tag{2.6}$$

In (2.6), the matrix $R_i = \{r_i(k,l); k,l = 1,2,3,...L_F\}$ is expressed as the correlation matrix. The diagonal elements are real, symmetric in terms of positive, semi-definite and toeplitz. Its inputs and are stated as follow as,

$$R_i = \begin{bmatrix} r_i(1) & r_i(2) & r_i(3) & \dots & r_i(L_F) \\ r_i(2) & r_i(1) & r_i(2) & \dots & r_i(L_{F-1}) \\ r_i(3) & r_i(2) & r_i(1) & \dots & r_i(L_{F-2}) \\ \vdots & \vdots & \vdots & \ddots & \dots \\ r_i(L_F) & r_i(L_{F-1}) & r_i(L_{F-2}) & \dots & r_i(1) \end{bmatrix} \tag{2.7}$$

Obviously, λ_{ik} and ϕ_{ik} are specified as eigenvalues and eigenvectors of the problem being considered. It is also known that the eigenvalues of R_i are non-negative, real, and distinct. Furthermore, the eigenvectors ϕ_{ik} seem to be entirely orthonormal.

$$r_i(d+1) = \frac{1}{L_F} \sum_{j=[(l-1)iL_F+1]}^{[iL_F-d]} x_j x_{j+d} \tag{2.8}$$

The eigenvalues are sequenced in decreasing order with the corresponding eigenvectors ($\lambda_{i1} \geq \lambda_{i2} \geq \lambda_{i3} \; ... \; \geq \lambda_{iL_F}$). The total energy of the frame i is then expressed as follows:

$$X_i^T X_i = \sum_{k=1}^{L_F} x_{ik}^2 = \sum_{k=1}^{L_F} c_{ik}^2 = \sum_{k=1}^{L_F} \lambda_{ik} \tag{2.9}$$

As can be seen, Equation (2.9) is cut by taking the main components of the first p with the highest energy of the original signal,

$$X_i \cong \sum_{k=1}^{p} c_k \phi_{ik} \tag{2.10}$$

It is shown as follows; (2.10) is obtained by putting *p=1* in its simplest form. The eigenvalue vector ϕ_{ik} is called the signature vector. In other words, it is the highest energy signature vector in the sense of LMS. It is specified that each frame belongs to the original speech signal. It looks like the following,

$$X_i \cong c_1 \phi_{ik} \tag{2.11}$$

In this case, the frame length L_F has been changed to a parameter where almost all of the energy (2.10) is caught in the first period and the rest becomes insignificant. Hence, (2.11) is obtained. For this reason, ϕ_{ik} contains the original signal frame, which is called the signature vector of the most useful information. Once (2.11) is acquired, (2.11), an envelope matrix for each frame is transformed into an equation through E_i that is a diagonal matrix. Thus, X_i is calculated as:

$$X_i = C_i E_i \phi_{i1} \tag{2.12}$$

(2.12) is expressed by a simple division of the diagonal inputs e_{ir} of the matrix E_i, the entries of the signature vector ϕ_{i1} and the entries ϕ_{i1r} of the frame vector X_i,

$$e_{ir} = \frac{x_i r}{c_1 \phi_{i1r}} (r = 1\ , 2\ , 3, ...L_F). \tag{2.13}$$

Therefore, e_{ir} which is in fact the property of (2.13), absorbs some of the energy of terms that are eliminated by truncating (2.10). When e_{ir} (the frame index against e_{ir} ; n = 1, 2, 3, , L_F) and ϕ_{i1r} (the frame index against to ϕ_{i1r}; n= 1, 2, 3, ..., L_F) are plotted, it is seen that the patterns obtained are iterative similarities. These similar speech signals have been inferred predicted as being obtained due to the quasi-stationary behavior of the forms. Therefore, similar patterns can be removed and a set of predefined "Envelope (PES) and Signature (PSS) Sequence" with some kind or unique patterns are created.

This method is a systematic process for modeling speech signals in four main stages defined as follows:

Step 1: ***Choosing of speech parts to create signature and envelope sequences:***

To examine the diversity of speech parts on a frame by frame, the basic features of the speakers and languages that the signature and envelope sequences have been assigned, are defined and investigated. This section has resulted in hundreds of thousands of envelope and signature sequences for different languages. In addition, these sequences exhibit a large number of similar models that need to be removed.In this first step of the method, hundreds of thousands of envelopes and signatures were created, and in addition, there are a similar number of models that should be removed.

Step 2: ***Extraction of similar patterns:***

Afterward,After similar the predefined Envelope (PSS) and Signature (PSS) forms were created,. Then remove similar envelopes and similar layouts are removed to and get obtain unique patterns. The Pearson Correlation correlation Coefficient coefficient (PCC) was used to remove similar patterns as described in [6]. It is determined by the PCC and is given as:

$$\rho_{yz} = \frac{\sum_{i=1}^{L}(y_i z_i) - [\sum_{i=1}^{L} y_i \sum_{i=1}^{L} z_i]/L}{(\sqrt{\sum_{i=1}^{L}(y_i^2) - (\sum_{i=1}^{L} y_i)^2/L[\sum_{i=1}^{L}(z_i^2) - (\sum_{i=1}^{L}(z_i)^2]/L]} \tag{2.14}$$

In this equation, is always between -1 and +1, a nd while $Y = [y_1\ y_2\ ...\ y_{L_F}]$ and $Z = [z_1\ z_2\ ...\ z_{L_F}]$ are two sequences subject to comparison. $\rho_{yz} = 1$ specifies that the two vectors are the same. $\rho_{yz} = 0$ corresponds to vectors completely unrelated to each other. If $0.9 \leq \rho_{yz} \leq 1$, it is assumed that the two series are almost identical. Therefore, similar signature and envelope sequences have been removed accordingly and unique signature and envelope sequences have been obtained.

Step 3: ***Using the speech enhancement method to reduce noises in the background at the end of the reconstruction:***

The frame is ready to synthesize a particular speech segment *x(n)* for N length, frame by frame, after PSS and PES are generated. In this case, it should be divided into frames of length L_F to form the main frame vector of (2.5). Then for each frame X_i calculates the best approximation $X_{Ai} = C_i E_K S_R$ by subtracting S_R from the PSS, and E_K from the PES and then by computing the actual

coefficient C_i to reduce each frame error, which is defined by $\varepsilon \approx \varepsilon_i(n) = x_i(n) - C_i E_K S_R$; it is made with the sense of LMS. At the end of this, approximate frame vectors (arrays) X_{Ai} are gathered under the approximate main frame matrix $M_{AF} = [X_{A1}\ X_{A2}\ X_{A3}\ ...\ X_{AN_F}]$ to reconstruct the speech as $x_A \approx x(n)$.

Step 4: ***Reconstruction of speech frame by frame:***

At the end of the third step, the reconstructed signal as a result of combining the speech frames in consecutive order contains unexpected spikes. These spikes can produce unexpected background sounds that can be qualified as musical noise. The purpose of this study implies is to use and examine the speech enhancement methods, and by comparing these methods, it will be possible to experience that the audio signal compressed with the best performance can be significantly reduced by speech enhancement methods. In this approach, it is explicitly stated that the gain factor C_i of frame i is reconstructed with three important parameters,i.e., the index R of the predefined signature vector S_R drawn from PSS and the index K obtained from the predefined envelope parameter E_K pulled from PES. The S_R and E_K parameters are specified to reduce the LMS error described by the difference between the original frame part X_i and the model $X_{Ai} = C_i E_K S_R$.

Finally, in this method, frame speech signals are presented to the frame by means of predefined "Signature and Envelope" patterns. In this process, the reconstructed speech frame X_{Ai} is defined by the multiplication of three main quantities, the gain factor C_i, the diagonal envelope matrix E_K, and the frame signature vector or briefly, $X_{Ai} = C_i E_K S_R$. Signature and envelope samples were selected from the relevant PSS and PES generated using a speech model variation included in the IPA. The PSS and PES array sets are independent of the speaker and the underside, that which is almost universal. During the synthesis process, each speech frame is indicated by the gain factor C_i and is the R and K indices, respectively, of the predefined signature and envelope patterns.

Subjective test evaluations show that SYMPES improves quality at lower compression ratios ($CR \ll 8$) to better than ADPCM (16, 24, 32 and 48 kbps). In other respects, SYMPES results in excellent hearing quality at higher compression ratios ($CR \gg 8$) than ADPCM and LPC techniques.

Chapter 3

Understanding the Enemy: Noise

In generally, the Spectral spectral Subtraction subtraction is used to designing algorithms to before struggle counter additive noise. It is critical to understand the attitude of different kinds of noises fully; the variation between the noise sources in conditions of spectral and temporal features, and the range of noise levels that can be met in real life.

3.1 Noise Sources

Wherever we go, noise is always in our surroundings. Noise is available, for instance, in the car (e.g., engine noise, the wind), in the street (horn sound, street construction work, etc.), on the plane in the car (the wind, engine noise, etc.), and the office (PC fan noise, keyboard sound, etc.), and the shopping centers (e.g., people talking, sales representatives talking, etc.). As these examples explained, noise appears in different versions and shapes in daily life.

The noisy signal can be stationary, so it does not change over time like the fan sound from PCs. The noisy signal can also be nonstationary, as like such as the restaurant noise, that is, the noise of various spoken people talking in the background noise of mixed with noise spreading from the kitchen.

Spectral and temporal features of the restaurant noise are frequently changing as people keep going speaking in neighboring tables and as the waiters interact and converse with people. Evidently, the duty of preventing noise that is nonstationary as well as always changing ever changing is harder more difficult than the preventing from stationary noise.

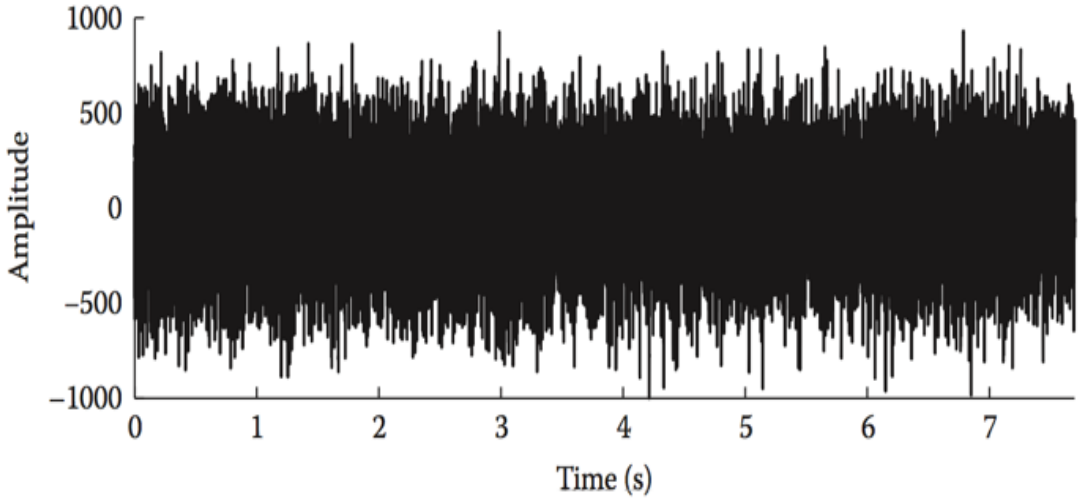

Figure 3.1: Example noise from a car

The other particular characteristic of the different types of noises is the form of their spectrum, especially as it depends on the distribution of noisy signal energy in the frequency domain.

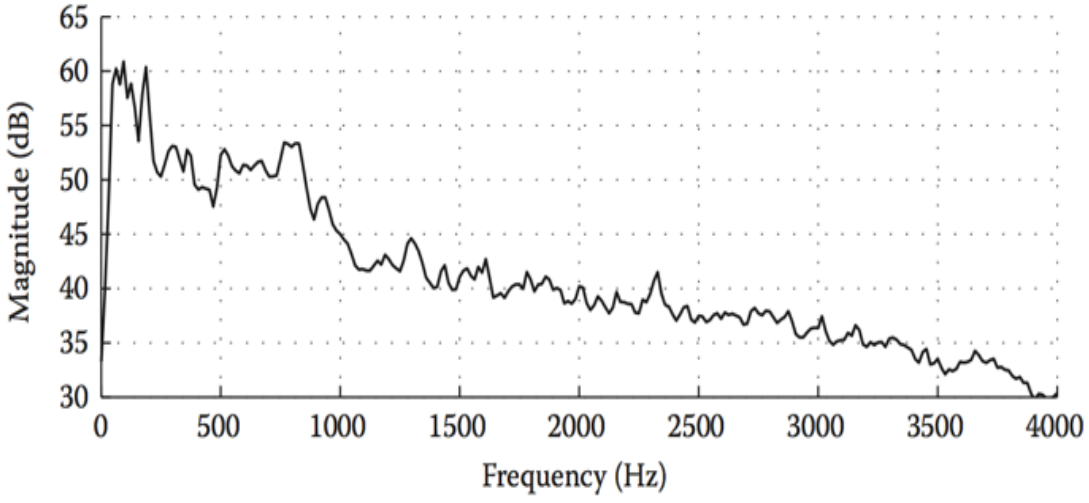

Figure 3.2: Long-term average spectrum

For example, the noisy energy of the wind noise is at low frequencies, such as below 500 Hz. In restaurant, noisy signal, otherwise, holds a wide frequency range. Figures 3.1 to 3.5 show example time waveforms of car noise, train noise, and restaurant noise. The corresponding long time mean spectra of the noise sources that are given in the examples are also shown. In the three noisy signal sources, the car noise (Figure 3.1) is relatively stationary, but there are not train

and restaurant voices. Figures 3.1 to 3.5 show that the differences between these three noisy signal sources are clearer in the frequency domain than in the time domain. The noise of the noisy signal in the vehicle is concentrated at very low frequencies because it is low pass in nature. However, since it has a wide frequency range, the train noise is wideband.

3.1.1 Noisy Signal and Speech Levels in Various Environments

The information of speech and noise intensity levels are is critical to the design of speech enhancement algorithms. The signal-to-noise ratio (SNR) levels encountered can be estimated in realistic environments, which that is important, because speech enhancement algorithms must be effective in preventing from noise and increasing speech quality in the range of SNR levels.

An exhaustive analysis and estimate of speech and noise levels in real-world environments was done by Pearsons [20]. They considered a variety of environments encountered in daily life, which included classrooms, urban and suburban houses (inside and outside), hospitals (nursing stations and patient rooms), department stores, trains, and airplanes.

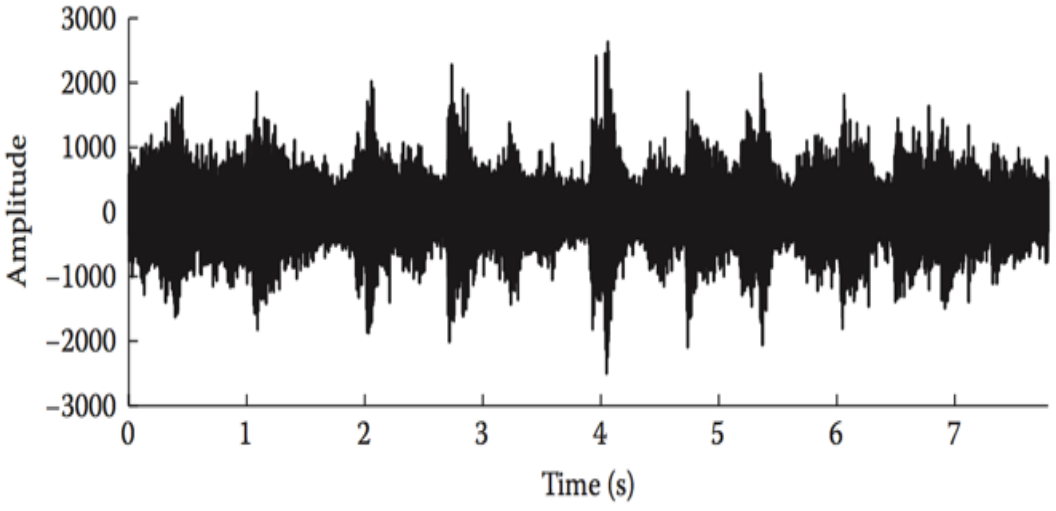

Figure 3.3: Example noise from a train

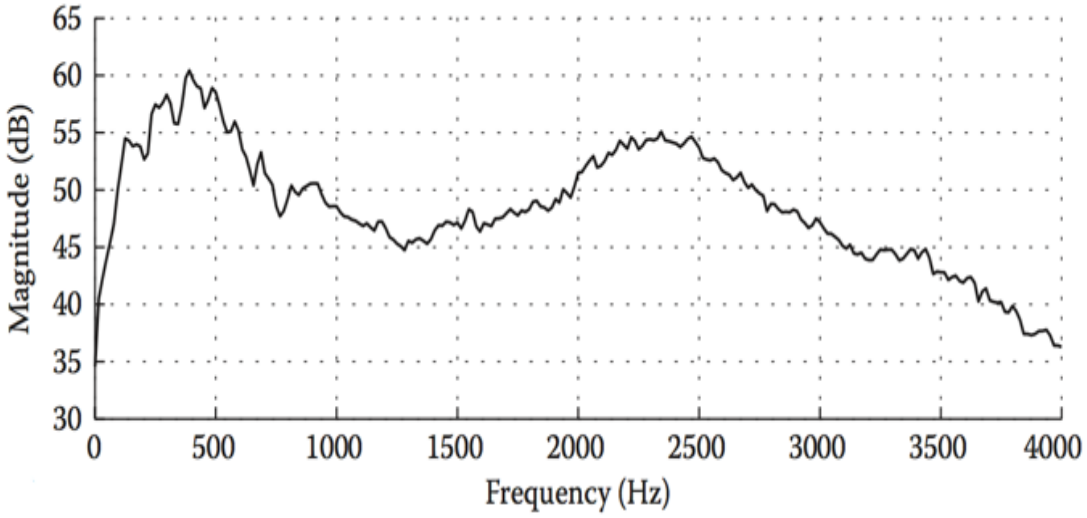

Figure 3.4: Long-term average spectrum.

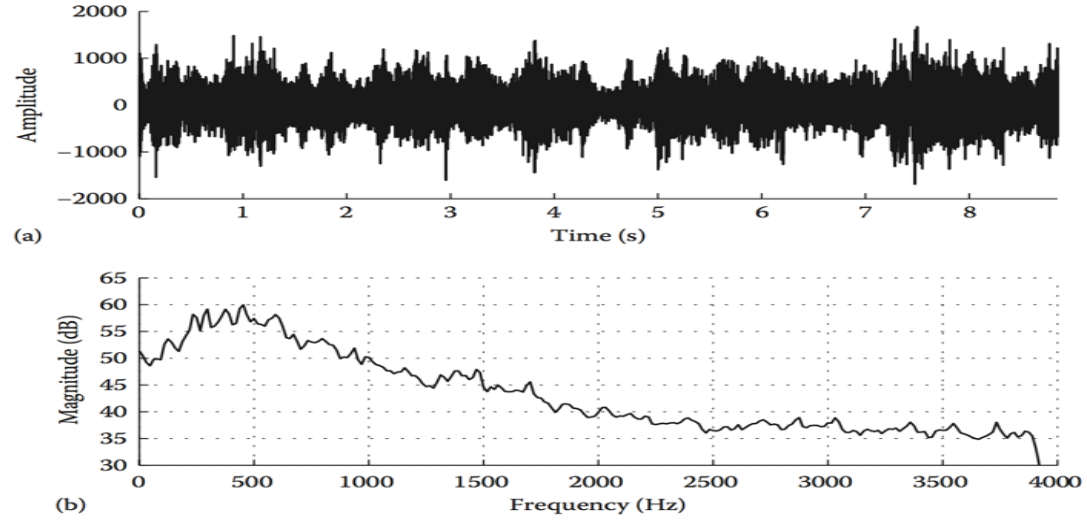

Figure 3.5: Example noise from a restaurant and (b) its long-term average spectrum.

Chapter 4

Classification of Speech Enhancement Algorithms

Classifying the methods of speech enhancement with many methods is possible. In general, it is quite difficult for a particular algorithm to work homogeneously across all types of noise. For this reason, speech enhancement systems are based on certain hypotheses and constraints, which are dependent on the environment and dependent on the application. In general, the performance of the speech enhancement algorithm may depend on the following factors:

- The determinative limitations on the number of noise sources that are used may provide for different uses of previously known information about the signal of interest and the corrupting signal,
- Model-based conditions such as (non-stationary) limitations on the permissible time variation for the corrupting signal; a restriction of the algorithm to uncorrelated noise.

The number of input channels (one / two / multiple) coming from the speech enhancement systems can be classified according to the field of application (time or frequency), depending on the type of algorithm (Adapted or Adaptable) [23][24][25]. Especially In particular, in speech development literature, various speech processing methods are separated into single and multi-channel development techniques.

This system uses the most common real-time implemented algorithms; mobile communication, hearing aids, etc. It is usually not available to for use it as a second channel.

Single microphone systems: This system uses the most common real-time implemented algorithms, mobile communication, hearing aids, etc. It is usually

not available to be used it as a second channel. These systems are easy to set up and are much cheaper the multiple input systems. One of the most difficult situations of difficult speech convergence is that there is no reference signal to noise and clean speech is not pre-processed before being affected by the speech voice. Single channel systems use different speech statistics and undesired noises. It can be said that the performance of these methods is commonly limited to non-stationary noise because most of them are noisy at speech intervals and the performance is assumed to be considerably reduced at low SNRs.

Multiple microphone systems: A noise reference in a noise cancellation device that suffers from multiple signal input to the system, and noise cancellation is utilized in these systems. Specific limitations for single-channel systems are clearly marked by taking into account the spatial characteristics of the signal and noise source. Multiple microphone systems are intended to be more complex. Adaptive noise cancellation is one such powerful speech enhancement technique based on the availability of an auxiliary channel, known as source path, where a correlated sample or reference of the contaminating noise is present. This kind of powerful speech enhancement technique is called adaptive noise cancellation, which occurs within the presence of a correlated example or a reference channel that is known to have a reference noise. Delayed and summed beamforming is the most direct approach. For instance, multi-sensor beamforming [26] is performed via microprocessor arrays reproduced from radar and sonar applications. The underlying idea behind this system is based on the assumption that the direction of arrival of the desired signal. It is known and also the improvement of the reflections is small. Further, for each sensor to correctly align the phase function, the desired signal can be increased, and all the noisy components not in phase are ignored. From this point on, the analysis in this section will focus on single-channel enhancement techniques, since this constitutes the most common use of enhancement algorithms.

4.1 Single Channel Enhancement Systems

Over the last few decades, the problem of enhancing noise-reduced speech in the background has only been a useful research topic when there is noisy speech. As mentioned earlier, if there is only one microphone available, speech is one of the most demanding conditions for enrichment because no specific reference signal is available for noise. It cannot be prepared before the clean speech is exhibited

to noise. Single-channel systems clearly show limited performance, because they are enhanced despite the loss in the signal [27][28]. For this reason, there is a contrast between quality and intelligibility. Therefore, if we want a quality signal, we should not compromise on clarity.

The existing single-channel speech enhancement systems are separated into three groups:

1. Model based speech enhancement.
2. Short-time spectral spectral amplitude amplitude estimation based speech enhancement.
3. Enhancement based on Perceptual perceptual criteria.

4.1.1 Statistical Model Based Algorithms

This speech enhancement method involves predicting a clean signal of a noisy signal for a particular sample function that requires explicit knowledge of the clean signal. Noisy signal statistic requires a perceptual distortion measure for speech signals. For this reason, if statistically, independent noise distorts the speech signals, the marginal probability distributions of the clean speech and noise signal should be precisely known.

The problem of speech enhancement is explained in a statistical estimation approach. The series of measurements corresponding to the Fourier transform coefficients of the noisy signal is implemented. It is necessary to find a clean signal conversion coefficient, for instance; a linear or non-linear estimation of the corresponding parameter.

For example, minimum mean square error (MMSE) algorithm are lies in this category along with others. The work on this subject have been started with McAulay and Malpass [28], and the maximum likelihood approach has been recommended in predicting the Fourier transform coefficients of the clean signal. This approach has been followed by studies by Ephraim and Malah [29] as the MMSE estimator.

Hendriks' theoretical approach can be applied as a two-step procedure:

1. The signal and noise statistics are estimated from the speech and noise content. The optimal solution is obtained by utilizing statistical models and then used together with distortion measures to solve the available speech enhancement problem. Different techniques are applied to parameterize the speech signal such as autoregressive moving average (ARMA), autoregressive (AR) or moving average (MA). It is known that the three estimation rules are known as maximum probability (ML), maximum posterior (MAP) and minimum mean square error (MMSE) have many desirable properties to estimate the parameters of a speech signal. Regularly, ML is utilized for non-random parameter maintenance. MAP and MMSE, which are estimation methods, are widely used for known parameters of a priori density function which can be thought as random variables.

   For the speech signal, this model is proposed by using the MAP estimation approach. It is a supposed a time-varying AR model for the speech enhancement in which both the model and the signal are estimated from the noisy signal [32]. The maximization of the likelihood probability function is carried out once in the AR model by supposing the clean signal. It is present and additionally by repeating. On the clean signal using a known estimate of the estimated model and the noise power spectral density accepted.

2. As another class entering this category, examples of Hidden Markov Model (HMM) as a development technique is are introduced. HMM essential speech models implement well to better represent the different spectrums of speech signals as well as the correlation of the time frequency of the signal (i.e., second-order statistics of speech signals.) Time-frequency correlation can be useful to significantly influence the robustness of the signal estimator for the required smoothness restriction in enhancement implementations.

4.1.2 Enhancement Based on Short-Time Spectral Amplitude Estimation

The short-term spectral amplitude estimate is depended dependent on most commonly used speech enhancement techniques. Specifically, it is easier to calculate the original spectral amplitude. When when comparing, the spectral amplitude associated with the original clean speech and estimating both the amplitude and

the phase. As stated in [33], it is known that the short-term spectral amplitude is more important for the intelligibility and quality of speaking.

Speech enhancement techniques can be divided into two main groups [33]. This method is indicated on the principle of spectral subtraction as a first group. It is based on the removal of the noise, by shifting sequential short speech segments to the frequency domain, and by removing the noise estimate estimated during speech pauses. The method usually is sets the particular frequency "coefficients" of the words on a frame basis. In the second group, the impaired speech that includes the methods previously. It used to obtain a filter applied to the impaired speech first.

Different methods (differentiated by noise estimation, suppression rules, and other details) are all known as Short Term Spectral Amplitude (STSA) methods. Moreover, the majority are known as subtractive type algorithms. Speech and noise signal are commonly supposed to be uncorrelated. Spectral subtraction method is one of the most comprehensive speech enhancement techniques on short-term amplitude estimation.

4.1.3 Speech Enhancement According to Perception Criteria

The detection of voice or speech signal is the result of various physiological and psychological influences not fully understood. For this reason, although the clarity of speech is strongly related to the perception of the human being, these aspects are not involved in the processing of noisy speech. In contrast to the above described methods, in which only signal and speech features are applied. Various various improvisations can be made taking advantage of the characteristics of the human ear in the area of noise suppression and speech enhancement.

These speech enhancement techniques aim to overcome the problem of classical trade noise reduction and speech impairment, where the speech is masked against further suppression of the noise. The possibility of further distortion [35] of speech by this means is significantly reduced.

Chapter 5

Spectral Subtraction Algorithms

5.1 Introduction

The spectral subtractive algorithms were firstly recommended by Weiss et al. [35] in the correlation domain and subsequently by Boll [36] in the Fourier transform domain.

The spectral subtraction algorithm is one of the earliest researchers related to noise reduction in history. Moreover, further research has been performed, and more articles have been written to describing the differences of this algorithm compared to other algorithms. When additional noise is considered, a noise spectrum is neglected in the noisy signal spectrum, and a clean signal spectrum is taken. In the absence of signal, the noise signal spectrum can be updated and estimated.

This approach implies that the noise spectrum is not considerably different between the update periods and a noisy stationary or slowly changing process. The inverse discrete Fourier transform of the estimated signal spectrum is computed to obtain the enhanced signal by using the phase of the noisy signal. The forward and inverse Fourier transforms are used in the algorithm; therefore, the algorithm is quite simple. The simple subtraction algorithm is a costly process. The subtraction process must be performed with much care to prevent any speech distortion. If too many subtractions are done, some speech information may be reduced, but if too little is subtracted, the vast majority of intervening noise may stay.

Many methods have been recommended for interferences due to noise residues; moreover, in some cases, most of the speech impairments result from spectral subtraction, and thus, noise residues have been removed.

5.2 Basic Principles of Spectral Subtraction

Assuming that $y(n)$ is the input signal causing the noise, which is created from a clean speech signal $x(n)$ and an additional noisy signal, $d(n)$, given by

$$y(n) = x(n) + d(n) \tag{5.1}$$

The discrete-time Fourier transform of both sides is given as

$$Y(w) = X(w) + D(w) \tag{5.2}$$

The polar form can be expressed $Y(w)$ as:

$$Y(w) = |Y(w)|x^{j\phi_y(w)} \tag{5.3}$$

where

$|Y(w)|$ is defined the magnitude spectrum

$\phi_y(w)$ is defined the spectrum of the distorted noisy signal

The noisy signal spectrum, i.e., $D(w) = |D(w)|e^{j\phi_d(w)}$ defines magnitude and phase spectrum. The noisy signal magnitude is an unknown parameter. Additionally, it can be modified by the mean value of the calculated non-speech activity. Likewise, the noisy speech phase $\phi_y(w)$ can be replaced by the noise phase $\phi_d(w)$. The part that affects the comprehension of speech influences the intelligibility quality of speech [31]. After these modifications are given to (5.2), a prediction is obtained for the clean signal spectrum as follows:

$$\hat{X}(w) = \Big[\big|Y(w)\big| - \big|\hat{D}(w)\big|\Big]e^{j\phi_y(w)} \tag{5.4}$$

This symbol "ˆ" will be used to represent the estimated spectrum or estimated interest parameters. Exclusively, the inverse Fourier transform (IFT) of is utilized to obtain the enhanced speech signal. The basic principle of spectral subtraction is summarized in Equation 5.4. The noisy speech signal, which is the amplitude spectrum, is calculated by fast Fourier transform (FFT) and the noise spectrum is estimated without speech.

The spectrum of the noisy signal magnitude is subtracted from the noisy speech magnitude spectrum, and thus, the IFT of the difference spectra is taken to generate the enhanced speech signal. It has been noted that the spectrum of the enhanced signal magnitude, $\hat{X}(w) = \big|Y(w)\big| - \big|\hat{D}(w)\big|$, may be negative due to errors in estimating the noise spectrum. The magnitude of the spectra cannot be negative; for this reason, care must be taken to provide that $|\hat{D}(w)|$ is not always negative when excluded from the two spectra. An alternative approach to correct the difference spectra in the half-wave direction so that the negative spectral components are set to zero, is given as:

$$\big|\hat{D}(w)\big| = \begin{cases} -x, & if\ |Y(w)| - |\hat{D}(w)| \\ x, & else \end{cases} \tag{5.5}$$

Half-wave rectification is only one way of ensuring non-negative $|\hat{X}(w)|$ in many ways. The previous derivation makes it easy to expand in the power spectrum domain, which is the magnitude spectral subtraction algorithm. In some cases, it may be useful to work with power spectra instead of magnitude spectra as an alternative. To obtain the short-term power spectrum without a noisy speech signal, when multiplied encode the value in (5.2) by conjugate $Y^*(w)$, we get the following equation:

$$\begin{aligned} |Y(w)|^2 &= |X(w)|^2 + |D(w)|^2 + X(w) \cdot D^*(w) + X^*(w)D(w) \\ &= |X(w)|^2 + |D(w)|^2 + 2Re\big\{X(w) \cdot D^*(w)\big\} \end{aligned} \tag{5.6}$$

The terms, $|D(w)|^2$ and $X(w) \cdot D^*(w)$, which are the terms of the above equation, are not clearly obtained and they are approximated as, $E\big\{|D(w)|^2\big\}, E\big\{X^*(w) \cdot D(w)\big\}$ and $E\big\{X(w) \cdot D^*(w)\big\}$ where the expectation operator is denoted $E[\cdot]$. In particular, $E\big\{|D(w)|^2\big\}$ is predicted throughout non-speech activity and it is expressed in $|\hat{D}(w)|^2$ shape. A clean signal $x(n)$, assuming $d(n)$ is uncorrelated

and zero mean. In this case, the elements of the equation, which are $E\{X^*(w) \cdot D(w)\}$ and $E\{X(w) \cdot D^*(w)\}$ are degraded to zero. In this manner, the estimate of the clean speech power spectrum can be determined as follows after using the previous assumptions:

$$|\hat{X}(w)|^2 = |\hat{Y}(w)|^2 - |\hat{D}(w)|^2 \tag{5.7}$$

In the preceding equation, the power spectrum extraction algorithm is explained. As previously mentioned, it is not guaranteed in (5.7) that the power spectrum $\hat{X}(w)|^2$ is positive, but can be a half-wave rectifier, specified in (5.5). Eventually, an enhancement signal was acquired by calculating the IFT by using the phase of the noisy speech signal. If the IFT of either side of (5.7) is taken, a similar equation is obtained in the autocorrelation domain, as described below:

$$r_{\hat{x}\hat{x}} = r_{yy}(n) - r_{\hat{d}\hat{d}}(n) \tag{5.8}$$

$r_{\hat{x}\hat{x}}(n)$, $r_{\hat{y}\hat{y}}(n)$ and $r_{\hat{d}\hat{d}}(n)$ are the autocorrelation sequences of the estimated clean signal, noisy speech signal, and estimated noise signal, respectively, given in the above equation. Accordingly, subtraction can be accomplished in the autocorrelation domain. Thus, a technique was recommended by Weiss et al. [35]. He also recommended implementing the subtraction in the cepstrum domain. Equation 5.7 can be written as follows:

$$|\hat{X}(w)|^2 = H^2(w)|Y(w)|^2 \tag{5.9}$$

where

$$H(w) = \sqrt{1 - \frac{|\hat{D}(w)|^2}{|Y(w)|^2}} \tag{5.10}$$

The system transfer function is known as $H(w)$, which is also indicated as the gain function in speech enhancement. It should be noted that $H(w)$ in (5.10) is always positive and correct in principle. Moreover, the values are in the interval of $0 \leq H(w) \leq 1$. Sometimes negative values are obtained due to incorrect estimates of the noise spectrum. To obtain the enhanced power spectrum $|\hat{X}(w)|^2$,

suppression is applied to a frequency noise power spectrum $|\hat{Y}(w)|^2$ that provides $H(w)$, which is called the suppression function.

There is a specific speech enhancement algorithm according to the shape of the suppression function, i.e., each algorithm has a special suppression function. For this reason, it is often possible to compare the corresponding suppression functions by comparing different algorithms. The fact that $H(w)$ is a real value and when the IFT of is taken it is equal to zero, and hence, it is symmetrical and non-causal. Equation 5.9 corresponds to the time domain for a non-causal filtering operation by changing the suppression function. This method is recommended.

Another general version of the spectral subtraction algorithm is as follows:

$$|\hat{X}(w)|^p = |Y(w)|^p - |\hat{D}(w)|^p \tag{5.11}$$

P is the power of the power specified by the above equation. The original magnitude power spectral subtraction [3] is obtained with p = 1 and p = 2 with the spectral subtraction algorithm. The generic form of the spectral subtraction algorithm is represented in Figure 5.1. It was noted that (5.7) and (5.11) are only approximate values because of the inclusion of cross terms. The cross terms in (5.6) are statistically insignificant when using sufficient data and assuming that the signals are stationary.

However, speech signals are nonstationary. In many implementations, the speech signal cross-expectation terms may not necessarily be zero and are processed on a per-frame basis. As shown, the cross terms are not insignificant, at least at low frequencies, near the values of the power spectrum of the noisy speech signal. The values of small and perhaps insignificant cross terms at extremely high frequencies can be compared to noisy speech magnitudes at low frequencies. In most spectral subtraction algorithms, it is assumed that the cross terms are zero, although the fact that the entire spectrum cannot be neglected is also zero.

A process for estimating cross terms has been recommended [3]. The effects of subtracting cross terms are discussed using the geometric observe of spectral subtraction.

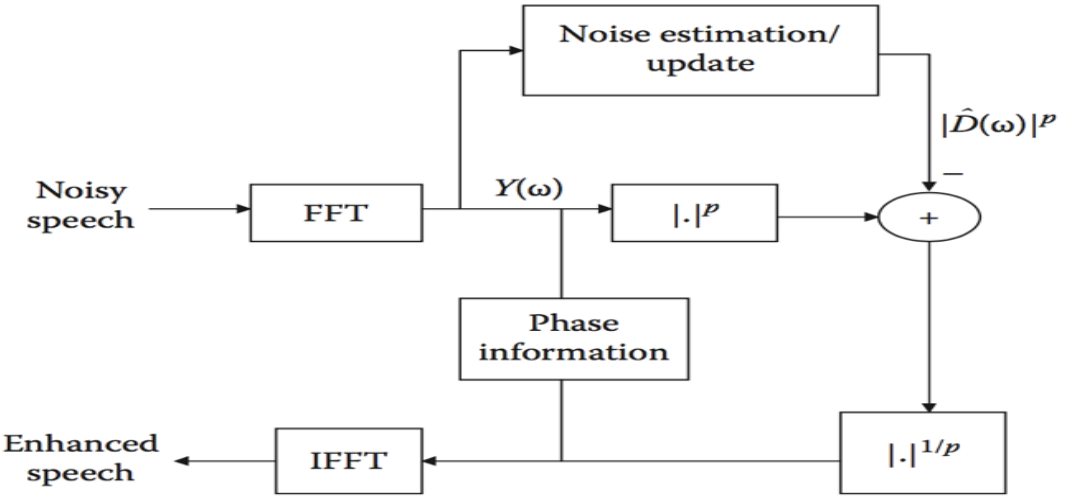

Figure 5.1: General form of the algorithm of spectral subtraction

5.3 Geometric View of Spectral Subtraction

From (5.1), two complex-valued spectra of frequency w_k are summed to obtain the frequency noisy spectrum $Y(w_k)$ at frequency w_k. For this reason, the summation of $X(w_k)$ and $D(w_k)$ complex numbers can be expressed geometrically as $Y(w_k)$ in the complex plane. The illustration of $Y(w_k)$ and a vector in summation to $X(w_k)$ and $D(w_k)$ in the complex plane is shown in Figure 5.2. When the noisy signal spectrum is geometrically represented in the complex plane in Figure 5.2, it can provide important information for the spectral subtraction approach.

On the difference between the phases of the clean and noisy signal spectrum, such a geometric perspective can provide an upper limit [39]; for example, it will explain the conditions that it is safe to take the approach that the noisy speech spectrum can be changed to the phase of the clean signal spectrum phase. $Y(w_k)$ will also indicate whether it is theoretically possible to fully reproduce the noise and the clean signal magnitude $|X(w_k)|$ in the given noisy speech spectrum $Y(w_k)$ and under what conditions. Finally, how the magnitude spectrum affects the estimation accuracy when the (5.6) cross terms are extracted from the center will be described in the next sections.

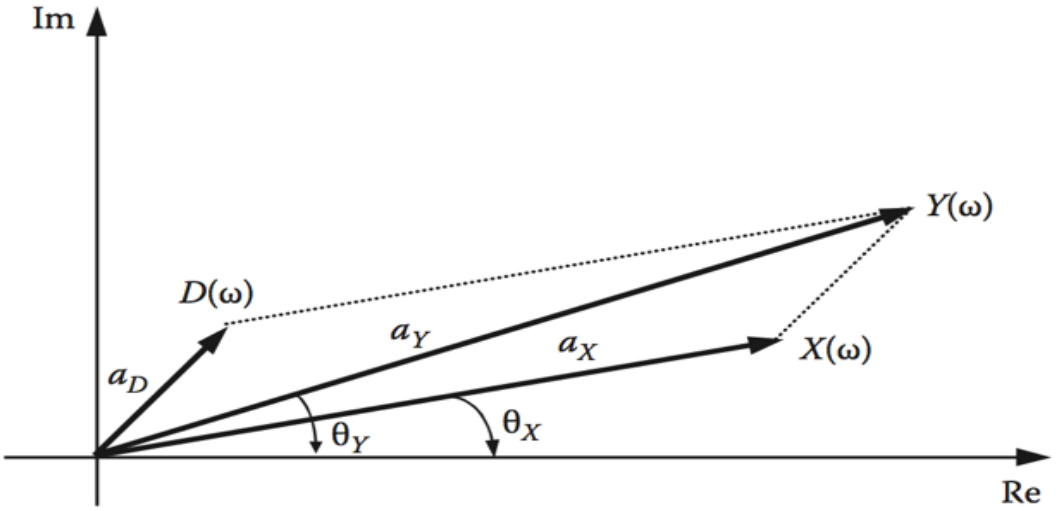

Figure 5.2: Noisy speech spectrum Y(w) that indicates complex vector addition of clean spectrum X(w) with noise spectrum D(w).

5.3.1 Upper Limits on the Difference Between the Phases of the Noisy and Clean Signals

$Y(w_k)$ is expressed as the complex spectrum that describes the magnitude and phase in polar form and the real or imaginary parts as follows as:

$$Y(w_k) = a_Y e^{j\phi_y} \tag{5.12}$$

where

The magnitude spectrum is expressed by a_y $(ie., a_y \triangleq Y(w_k))$

The phase spectrum expressed by θ_y

We subtract the frequency variable w_k for simplicity. Likewise, it can express clean signal and noise spectra in polar form as follows:

$$X(w_k) = a_X e^{j\theta_X}, \quad D(w_k) = a_D e^{j\theta_D} \tag{5.13}$$

The upper limit of the difference between the noisy and clean phases is determined by $\theta_y - \theta_x$. The vector diagram shown Figure 5.2 should be considered.

The phase difference $\theta_y - \theta_x$ is expressed by the following equation [39]:

$$tan(\theta_y - \theta_x) = \frac{a_D sin(\theta_D - \theta_X)}{a_X + a_D cos(\theta_D - \theta_X)} \tag{5.14}$$

It is easy to see from Figure.8 that when the noise and clean signal vectors are perpendicular to each other, the phase difference reaches the maximum value, i.e., when $\theta_D - \theta_X = \pi/5$

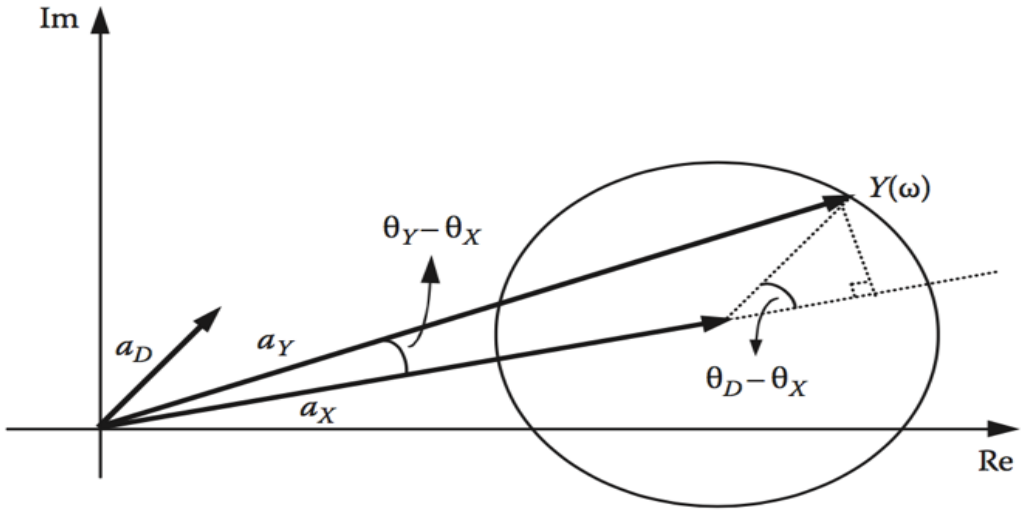

Figure 5.3: Diagram showing the trigonometric relationship of phase difference among noisy and clean signals.

In (5.14) replace $\pi/2$ for $\theta_D - \theta_X$, and when solved for the phase difference we obtain the following upper limit:

$$\theta_D - \theta_X = |_{\theta_D - \theta_X = \frac{\Pi}{2}} \triangleq \theta_{max} = tan^{-1}\frac{a_D}{a_X} \tag{5.15}$$

Note that the upper limit for $\left|\theta_D - \theta_X\right| > \pi/2$ is π, and the upper limit specified in the above equation applies to $\left|\theta_D - \theta_X\right| < \pi/5$

$$\xi = \frac{a_x^2}{a_d^2} \tag{5.16}$$

The frequency bin is expressed as the instantaneous spectral SNR in w_k, so that (5.16) can be written as:

$$\xi = tan^{-1}\frac{1}{\sqrt{\xi}} \tag{5.17}$$

The upper limit of the difference between clean and noisy phases as a function of the SNR is defined in (5.16). As predicted, the larger the SNR, the smaller the difference between clean (θ_X) and (θ_Y) noisy phases.

It has been expressed to be implemented that the noisy spectral phase at a particular frequency coefficient, the phase of the clean signal spectrum, the corresponding SNR is large in this segment. As part of speech enhancement, the following strategy can be implemented: the noisy phase can use the noisy phase as long as the phase difference is not identified by the hearing system.

5.3.2 Alternate Spectral-Subtractive Rules and Theoretical Limits

The previous part also shows that it is not necessary to fully restore the signal phase when the spectral SNR is high enough. In the next part, the signal size is returned into the problem of spectrum estimation.

By estimating the magnitude of the noise, it has been shown whether it is possible to improve the magnitude spectrum of the clean signal with relatively high accuracy. The question arises whether the phase information access of the relevant subject is critical for enhancing the signal magnitude spectrum, if so, is it necessary to have the correct signal size for accurate phase estimation? In response to this question, alternative spectral subtraction rules are acquired from those given in Section 5.1, and the relationship between the phases and magnitudes of the signals is examined. The real and imaginary parts of either side of (5.2) are set equal to each other, utilizing the notation in (5.12) and (5.13):

$$a_X cos(\theta_X) = a_Y cos(\theta_Y) - a_D cos(\theta_D) \tag{5.18}$$

$$a_X sin(\theta_X) = a_Y sin(\theta_Y) - a_D sin(\theta_D) \tag{5.19}$$

These are obtained by taking and adding the squares of (5.18) and (5.19) as follows:

$$a_Y^2 = a_X^2 + a_D^2 + 2a_X a_D cos(\theta_X - \theta_D) \tag{5.20}$$

It is seen that the previous equation is the same as in (5.6)

$$Re\big\{X(w)D^*(w)\big\}a_X a_D cos(\theta_X - \theta_D) \tag{5.21}$$

In (5.20), a_x is a quadratic function as described and the following two solutions are explained:

$$a_X = -a_D a_{XD} \pm \sqrt{a_D^2(c_{XD}^2 - 1) + a_Y^2} \tag{5.22}$$

The c_{XD} in (5.22) is defined as follows:

$$c_{XD} \triangleq cos(\theta_X - \theta_D) \tag{5.23}$$

Equation 5.20 is a function of the clean signal phase and magnitude of the cross terms, which prevents it from enhancing it. However, a different form of the equation, the magnitude of the noisy speech with cross terms, and the functions of the phase can be derived:

$$a_X^2 = a_X^2 + a_D^2 - 2a_Y a_D cos(\theta_Y - \theta_D) \tag{5.24}$$

When $cos(\theta_Y - \theta_D) = 1$ is specified, the previous equation is reduced for the standard subtraction rule given in (5.5). For clean signal magnitude, the different algebraic operations of (5.18) and (5.19) result in different expressions. For example, the clean signal magnitude a_x is shown to utilize the following equation:

$$a_X = a_Y c_{XY} \pm \sqrt{a_Y^2(c_{XY}^2 - 1) + a_D^2} \tag{5.25}$$

where c_{XY} is defined as follows:

$$c_{XD} \triangleq cos(\theta_X - \theta_Y) \tag{5.26}$$

For example, c_{XD} in (5.12), then we obtained $a_X^2 = a_Y^2 - a_D^2$. Geometrically, if $cos(\pm\pi/2) = 0$, when the clean and noise vectors are perpendicular to each other, they are expressed as c_{XD}. Statistically, if the noisy and clean signals have zero mean and the spectrum is orthogonal, they are uncorrelated. It is noted that the common assumption is that the noise and clean signal are uncorrelated to each other.

Therefore, it is very straightforward to use the standard power subtraction rule, (5.5), under conditions where signal and noise vectors are perpendicular to each other. Depending on (5.20), in the same circumstances, the power spectrum extraction rule, (5.7), is also true. Here, when it is assumed that the phase difference $(\theta_X - \theta_D)$ is uniformly distributed in the range $[-\pi, \pi]$. In this case, it can be expressed that $E[c_{XD}] = 0$.

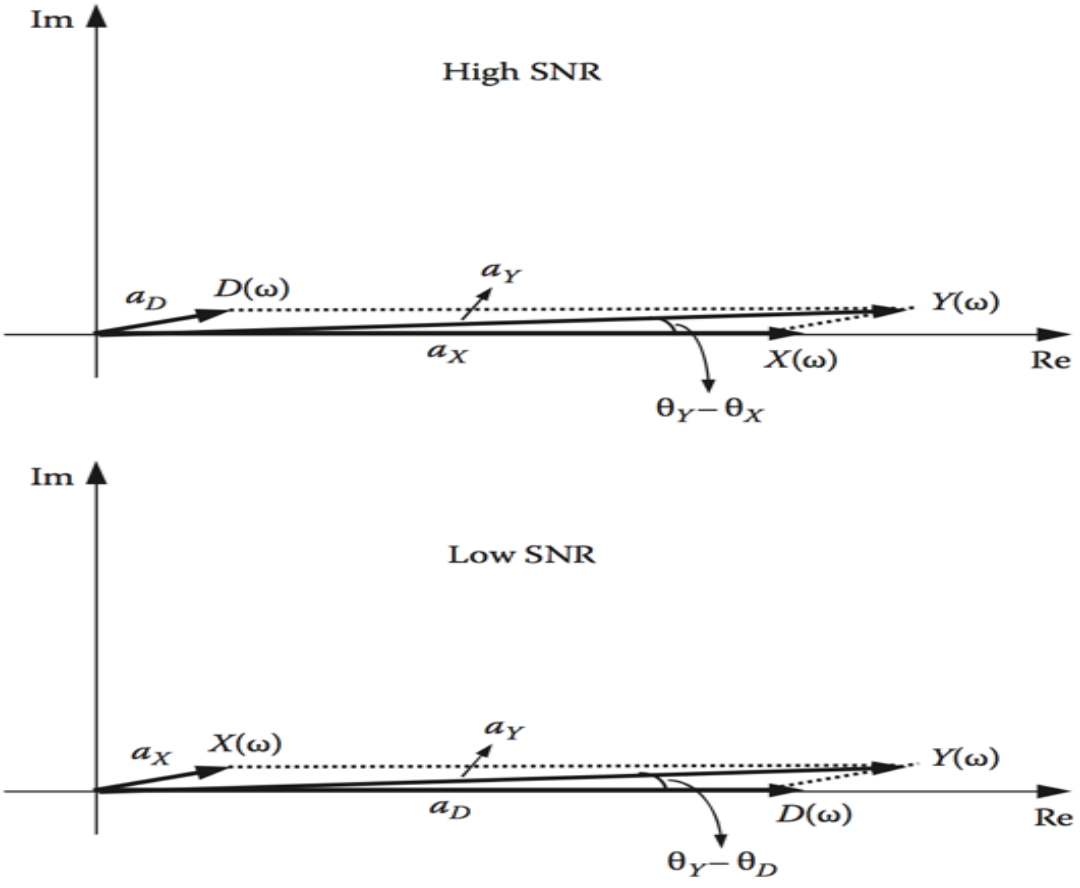

Figure 5.4: Geometric view of high and low SNR conditions

If $c_{XY} = 1$ in (5.25) and take negative sign, after $a_X = a_Y - a_D$ is obtained.

Geometrically, $c_{XY} = 1$ indicates that the noisy and clean signal vectors are co-linear, so, in the same direction. Noisy and clean signal vectors are roughly co-linear when SNR is high, i.e., when $a_X >> a_D$, as shown in Figure 5.4 Therefore, the standard subtraction rule (5.5) is quite accurate under high SNR conditions.

It is clear that the subtractive rules are given in (5.22) and (5.25)which are easier and simpler to use than those given in (5.5)and (5.7) because they do not make statistical assumptions about the relationship between clean signals and noise, the existence of cross terms is considered. However, subtractive rules (5.22) and (5.25) are ambiguous because there is no simple way to determine which sign ($\pm$) to use. Alternative subtractive rules used to avoid quadratic terms in can be derived by algebraically manipulating (5.18) and (5.19) without any sign uncertainty.

One example of these rules is given as below:

$$a_X = a_Y c_{YX} - a_D c_{XD} \tag{5.27}$$

$$a_X = \frac{a_Y c_{YD} - a_D}{c_{XD}} \tag{5.28}$$

$$a_X = \frac{a_Y - a_D c_{YD}}{c_{XY}} \tag{5.29}$$

$$a_x^2 = (a_Y - a_D c_{YD})^2 + a_D^2(1 - c_{YD}^2) \tag{5.30}$$

$$a_x^2 = (a_D - a_Y c_{YD})^2 + a_Y^2(1 - c_{YD}^2) \tag{5.31}$$

$$a_x^2 = \frac{a_D^2(1 - c_{YD}^2)}{(1 - c_{YX}^2)} \tag{5.32}$$

$$a_x^2 = \frac{a_Y^2(1 - c_{YD}^2)}{(1 - c_{XD}^2)} \tag{5.33}$$

Beginning with (5.30), after that we obtained $a_X = a_Y - a_D$. Geometrically, when $c_{YD} = 1$ is noisy and the noise vectors are co-linear in the same direction, they

point to the same direction. As noise and noise vectors are approximately co-linear, as shown in Figure 5.4 The SNR is low, so that, when $a_D \ll a_x$. Therefore, (5.30), under low SNR conditions, decrease the standard subtraction rule expressed in (5.5).

Where $c_{YX} \approx 1$ corresponds to a high SNR condition seen in Figure 5.4, (5.27) is observed to resemble the over subtraction rule proposed in [40], with the over subtraction factor explicitly indicated by c_{XD}. The value c_{XD} is always equal to value when over-subtraction factor is always equal to it. Only a distinct relation for instantaneous SNR (as well as shown in (5.16)) is given in (5.32) regarding signal/noise phases:

$$\xi \triangleq \frac{a_X^2}{a_D^2} = \frac{(1 - c_{YD}^2)}{(1 - c_{YX}^2)} \tag{5.34}$$

The geometric explication of high and low SNR conditions is consistent as shown in Figure 5.4 and described in the preceding equation. More specifically, if $c_{YX} \approx 1$, the denominator can be small and will result in large values of ξ, i.e., a high SNR case. Similarly, if $c_{YD} \approx 1$, the numerator will be small and the small values of ξ are small in SNR value case. Finally, (5.33) shows that the phase distinction among the noisy signal and the noisy/clean signal is clearly related to the gain function $a_x - a_y$ in the system:

$$h \triangleq \frac{a_X^2}{a_D^2} = \sqrt{\frac{(1 - c_{YD}^2)}{(1 - c_{YX}^2)}} \tag{5.35}$$

Fundamentally, the clean signal amplitude a_X should multiply by the previous gain function of the noisy amplitude a_Y. It is stated that the approach to this gain function is that the power spectrum subtraction method has fallen to the gain function. If $c_{XD} = 0$, is statistically perpendicular, the signal and noise are perpendicular to each other. To confirm this, solve for c_{YD} in (5.28) together with $c_{XD} = 0$ and then obtain the following equation:

$$c_{YD} = \frac{a_D}{a_Y} \tag{5.36}$$

The gain rule in (5.10), assuming $c_{XD} = 0$, because it is assumed that the noisy and clean signal are perpendicular to each other. However, several methods have

been proposed for phase estimation that need to be investigated in previous analyses. One way is to derive and use the phases of noisy and clean signals as shown in (5.14), explicit relationships between them based on trigonometric factors. Another way is to use (5.20) and (5.24), and by making the necessary changes using c_{YD} and c_{XD}, the following equation is obtained as:

$$c_{YD} = \frac{a_Y^2 + a_D^2 - a_X^2}{2a_X a_D} \tag{5.37}$$

$$c_{XD} = \frac{a_Y^2 - a_X^2 - a_D^2}{2a_X a_D} \tag{5.38}$$

It is easy to show that c_{YX} is given with the following equation as:

$$c_{YX} = \frac{a_Y^2 + a_X^2 - a_D^2}{2a_X a_Y} \tag{5.39}$$

Obviously, the principal obstacle to estimating the phase distinctions among signal and noise signals are using previous equations that they are not dependent on the clean signal amplitude. It uses estimated a_x values as a reasonable solution to make assumptions that imply and the spectra of magnitude do not change.

Therefore, when the spectral SNR is sufficiently high, it is safe to use a noisy phase instead of a clean signal phase. The phase difference should not be perceivable, depending on the explanations given in [39]. It has a phase estimation critical significance for accurate signal magnitude estimation. It is not possible to obtain the magnitude spectrum of the clean signal, even though it can reach the noise. It is necessary to acquire phase information.

When the subtractive rule determined in (5.33) is uniformly associated with the phase difference between noisy and clean signals, after that, the signal magnitude can be recovered without the need for significant information about noise magnitude. However, the fact that the phase distinctions between the noisy or clean signal and the noise signal are implicitly embedded is made possible by the knowledge of the magnitude of the noise (in see (5.37)).

5.4 Nonlinear Spectral Subtraction

This method is applied in the spectral subtraction algorithm recommended by Berouti et al. [40]. It is assumed that all spectral components that are noisy are affected at equal levels. As a result, a single over-subtraction factor is used to extracted a robust estimate over the entire spectrum. However, this is not the case with real world noises (e.g., car noise and restaurant noises).

Some disturbing noise may affect the low-frequency region of the spectrum more than the high-frequency region. This suggests using a frequency-dependent subtraction factor to add different types of noises to account. The nonlinear spectral subtraction (NSS) recommended in [41] is a modification of the method suggested in [40], because the subtraction factor frequency is dependent and subtraction is not linear. Larger values occur at frequencies with low SNR levels, while smaller values are derived at frequencies with high SNR levels. The format of the subtraction rule used in the non-linear subtraction algorithm is as follows:

$$\left|\hat{X}(w)\right| = \begin{cases} \left|\bar{Y}(w)\right| - a(w)N(w) & if\ \left|\bar{Y}(w)\right| > a(w)N(w) + \beta.\left|\bar{D}(w)\right| \\ \beta|\bar{Y}(w)| & else \end{cases} \tag{5.40}$$

where

β is the spectral floor (set to 0.1 in [41])

$|\bar{Y}(w)|$ and $|\bar{X}(w)|$ respectively referred to as noisy speech and noisy smoothed estimates.

$a(w)$ is specified as a frequency-dependent subtraction factor

$N(w)$ is specified as a nonlinear function of the noise spectrum

The tenderized estimates of noise (indicated as $|\bar{D}(w)|$) and noisy speech (indicated as $|\bar{Y}(w)|$) obtained as follows:

$$\begin{aligned} \left|\bar{Y}_i(w)\right| &= \mu_y\left|\bar{Y}_{i-1}(w)\right| + (1-\mu_y)\left|\bar{Y}_i(w)\right| \\ \left|\bar{D}_i(w)\right| &= \mu_d\left|\bar{D}_{i-1}(w)\right| + (1-\mu_d)\left|\bar{D}_i(w)\right| \end{aligned} \tag{5.41}$$

$|\bar{Y}_i(w)|$ is the magnitude spectrum obtained in the ith frame of noisy speech

$|\bar{D}_i(w)|$ is the estimate for the ith frame of the spectral magnitude spectrum

The values of the constants , take values in the interval $0.1 \leq \mu_y \leq 0.5$ and $0.5 \leq \mu_d \leq 0.9$. The term $N(w)$ of was obtained by calculating the maximum of $|\bar{D}_i(w)|$ the noise magnitude spectra over the last 40 frames:

$$N(w) = \max_{i-40 \leq j \leq} (|\hat{D}_j(w)|) \tag{5.42}$$

It should be noted that for all frequencies in [40], $a(w)$ is a constant; it changes only from frame to frame based on a posteriori SNR. For the function of $a(w)$ in (5.40), several nonlinear functions are proposed in [41] and different weighting types are specified for SNR. One of the many functions [41] is given in the following form:

$$a(w) = \frac{1}{1 + \gamma\rho(w)} \tag{5.43}$$

In the equation below, $\rho(w)$ is the square root of a posteriori SNR estimate and γ is a scaling factor as follow as:

$$\rho(w) = \frac{|\bar{Y}(w)|}{|\bar{D}(w)|} \tag{5.44}$$

It should be noted that the function in Eq. (5.43) resembles the extreme auction function proposed by Berouti et al. [40], in that it applies a small weight because the posteriori SNR [i.e., large values of $\rho(w)$] values have a large weight with respect to frequencies having high frequencies and low SNR [i.e., for small values of $\rho(w)$] values. In contrast to the over subtraction algorithm, the weighting is implemented separately to each frequency coefficient.

5.5 Minimum Mean Square Error Spectral Subtraction Algorithm

In previous methods, subtraction parameters were specified experimentally and were not optimally selected in any sense. Sim et al. [42] recommended a mean square error method to optimally select subtraction parameters. Their derivation was depended on a generalized spectral subtraction algorithm, which is a parametric formulation. Obviously, a general version of the spectral subtraction algorithm is considered as in the following equation:

$$|\hat{X}(w)|^p = \gamma_p(w)|Y(w)|^p - a_p(w)|\hat{D}(w)|^p \tag{5.45}$$

where

$\gamma_p(w)$ and $a_p(w)$ are related parameters

p is expressed as the power exponent

$\hat{D}(w)$ is expressed as the mean noise spectrum acquired during non-speech activity.

It should be noted that if $\gamma_p(w) = 1$, $a_p(w)$, and $p = 2$, that are set as shown and the parameters in (5.45). Furthermore, when set to $\gamma_p(w) = c_{YX}(w), a_p(w) = c_{XD}$, and $p = 1$, then the inference rule given in (5.27) is obtained, where c_{XY} and c_{XD} are defined in (5.26) and (5.23), respectively. The parameters $\gamma_p(w)$ and are $a_p(w)$ optimally determined by minimizing the mean squared error spectrum:

$$e_p(w) = |X_p(w)|^p - |\hat{X}|^p \tag{5.46}$$

where, an ideal spectral subtraction model is assumed. $|X_p(w)|^p$ is specified as the clean speech spectrum. It is stated here that the noisy speech spectrum is supposed to be composed of the sum of the two independent spectra, the true noise spectrum and $|X_p(w)|^p$. That is, the following equation appears to be valid for some constant p:

$$|Y(w)|^p = |X_p(w)|^p + |D(w)|^p \tag{5.47}$$

where $D(w)$ is shown as the true noise spectrum. Theoretically, in the previous equation it is expressed that the phases of the clean signal spectrum for $p = 1$ are equal to the phases of the noise spectrum.

When the mean square error of the error spectrum $e_p(w)$, i.e., $E\left[\left\{e_p(w)\right\}^2\right]$, with respect to $\gamma_p(w)$ and $e_p(w)$ is minimized by minimizing the mean-square error of the error spectrum $e_p(w)$, that is, $E\left[\left\{e_p(w)\right\}^2\right]$, with respect to $\gamma_p(w)$ and $e_p(w)$. The most appropriate subtractive parameters are obtained with the following:

$$a_p(w) = \frac{\xi^p(w)}{1 + \xi^p(w)} \tag{5.48}$$

$$\gamma_p(w) = \frac{\xi^p(w)}{1 + \xi^p(w)}\left\{1 - 1 + \xi^{-\frac{p}{2}}(w)\right\} \tag{5.49}$$

where

$$\xi(w) = \frac{E\left[\left|X_p(w)\right|^2\right]}{E\left[\left|D(w)\right|^2\right]} \tag{5.50}$$

The previous equations are derived after assuming that the individual spectral components of noise and speech are zero-mean and complex-Gaussian random variables statistically independent. This assumption is necessary to simplify the solution.

Conversion of (5.48) and (5.49) to (5.45) yields the most appropriate parametric estimation:

$$\left|\hat{X}(w)\right| = \left\{\frac{\xi^p(w)}{1 + \xi^p(w)}\left[\left|Y(w)\right|^p - (1 - \xi^{-\frac{p}{2}}(w))\left|\hat{D}(w)\right|^p\right]\right\}^{1/p} \tag{5.51}$$

The previous estimator has been derived without making any assumptions about the relationship between the two parameters $\gamma_p(w)$ and $a_p(w)$, and therefore is called the unrestricted estimator. Sim et al. [42], also assumed that the two parameters are equal to each other, i.e., $\gamma_p(w) = a_p(w)$ and a different estimator is derived. The form of the optimal constrained estimator obtained is as follows:

$$\left|\hat{X}(w)\right| = \left\{\frac{\xi^p(w)}{1 + \xi^p(w)}\left[\left|Y(w)\right|^p - \left|\hat{D}(w)\right|^p\right]\right\}^{1/p} \tag{5.52}$$

where p is a power exponent given here, δ_p is constant (0.2146, 0.5, and 0.7055 for $p = 1$, 2, and 3, respectively). A lower spectral bound was implemented in (5.40) to limit the attenuation of low-energy speech segments. The smoothed lower bound spectrum is obtained by averaging a noise-coded spectrum of the enhanced and smoothed spectrum estimated in the previous frame, an attenuated version of $x(n)$ $(0 < \mu < 1)$;

$$\mu\big|\bar{Y}(w)\big| = 0.5(\mu\big|Y(w)\big|) + \big|\bar{X}_{prev}(w)\big|) \tag{5.53}$$

where the smoothed lower spectral bound is specified as $\mu|\bar{Y}(w)|$. That is, when the enhanced spectral value $\mu|Y(w)|$ in (5.22) was small, and it should be set equal to $\mu|\bar{Y}|$. The form of the last constrained estimator was as follows as:

$$\big|\bar{X}(w)\big| = \begin{cases} \big|\hat{X}(w)\big| & if \quad \big|\hat{X}(w)\big| \geq \mu\big|Y(w)\big| \\ \mu\big|\bar{Y}(w)\big| & else \end{cases} \tag{5.54}$$

It should be noted that the term $|\bar{X}_{prev}(w)|$ used in (5.53) is based on the previous equilibrium and is not included in (5.52). The previous parameter μ (set in the range of $0.05 - 0.2$ in [42]) was used as a spectral flooring constant very similar to the parameter β in (5.41).

The parameter $\xi(w)$ in (5.50) that corresponds to the ratio of signal power to noise, and is often denoted as a priori SNR. Unfortunately, this term cannot be determined because we cannot access the exact clean signal. Following the approach recommended in [29], Sim et al. [42] a priori SNR is calculated as shown in the following equation:

$$\xi(w) \approx (1-\eta)max\Big(\frac{|Y(w)|^2}{|\hat{D}(w)|^2} - 1, 0\Big) + \eta\frac{|\bar{x}_{prev}(w)|^2}{|\hat{D}(w)|} \tag{5.55}$$

where $\eta(n)$ is a smoothing constant (set to 0.96) and $|\bar{x}_{prev}(w)|$ is the enhanced spectrum calculated in the previous frame. Equation 5.55 basically expresses the weighted average of the current instantaneous SNR (first term) and the old SNR (second term). The estimate of the noise spectrum has been corrected and improved in a similar way. The two previous estimators are compared with the standard spectral subtraction algorithm [11] using objective measures, which are

namely log-likelihood evaluates and the SNR [28]. The major enhancements (for the same p value) are observed only for low SNR levels compared to the standard spectrum subtraction algorithm. $p = 0.1$ in (5.54) was used as the constrained estimator. The performance achieved with $p = 1$ and 2 is almost the same.

For the constrained estimator, $\mu = 0.1$ in (5.54) was used. Performance obtained with $p = 1$ and 2 was nearly the same. The analysis of the two estimators shows that the compression function of the unrestricted estimator (calculated as $|\hat{X}(w)|/|Y(w)|$) has a similar bias to the average MMSE estimate [29]. The comparison between the suppression functions of the two estimators reveals that the restrictive estimator provides more noise reduction than the unrestricted predictor, especially for low-energy speech parts.

5.6 Spectral Subtraction Using Adaptive Gain Averaging

Gustafsson et al. [15] proposed to obtain a lower resolution spectrum to handle the first number and divide the existing analysis frame into smaller subframes. The individual spectra in each subframe are averaged to obtain a lower variance spectrum. Gustafsson et al. [15] suggested using the adaptive exponential averaging to smooth out the gain function over time to account for the second number, because of the utilization of the zero-phase gain function to avoid noncausal filtering. Gustafsson et al. [15] have proposed the introduction of a linear phase in the gain function. The recommended spectral subtraction algorithm decreases the processing delay to the period of the analysis frame. The block diagram of the proposed spectral subtraction algorithm is shown in Figure 5.5 5.6 [15].

The input signal is divided into sample squares of L and divided into subframes each consisting of $M(M < L)$ samples. The calculated spectra in each subframe, $|\bar{Y}_i^{(M)}(w)|$ and a lower variance size spectrum estimate, which indicates the number of spectrum components of the upper variant (i.e., the size of the FFT), and is expressed as an i frame number. A lower-resolution gain function of $|\bar{Y}_i^{(M)}(w)|$ is constructed as follows:

$$G_i^{(M)}(w) = 1 - k\frac{\left|\hat{D}_i^{(M)}(w_k)\right|}{\left|\bar{Y}_i^{(M)}(w)\right|} \tag{5.56}$$

where k specifies as a subtraction factor (set to k = 0.7 in [15])

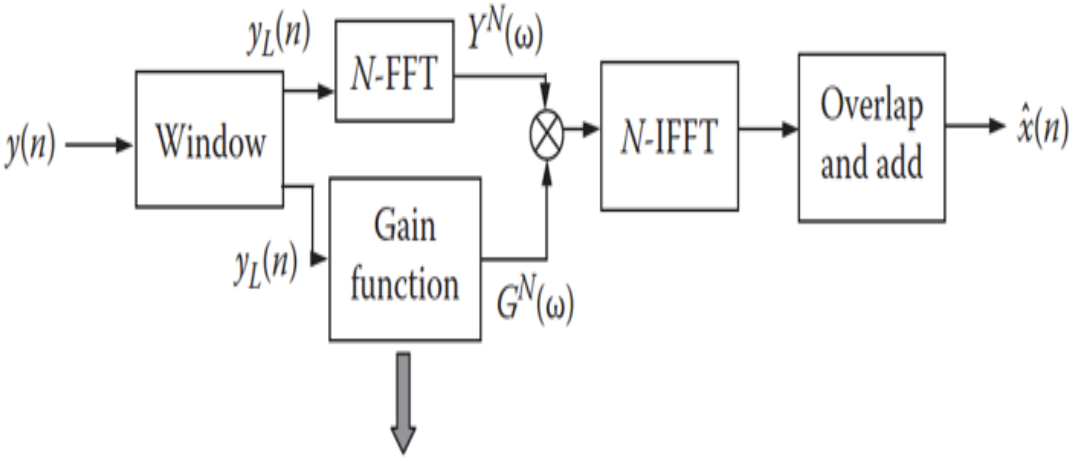

Figure 5.5: Spectral subtraction algorithm with adaptive gain averaging

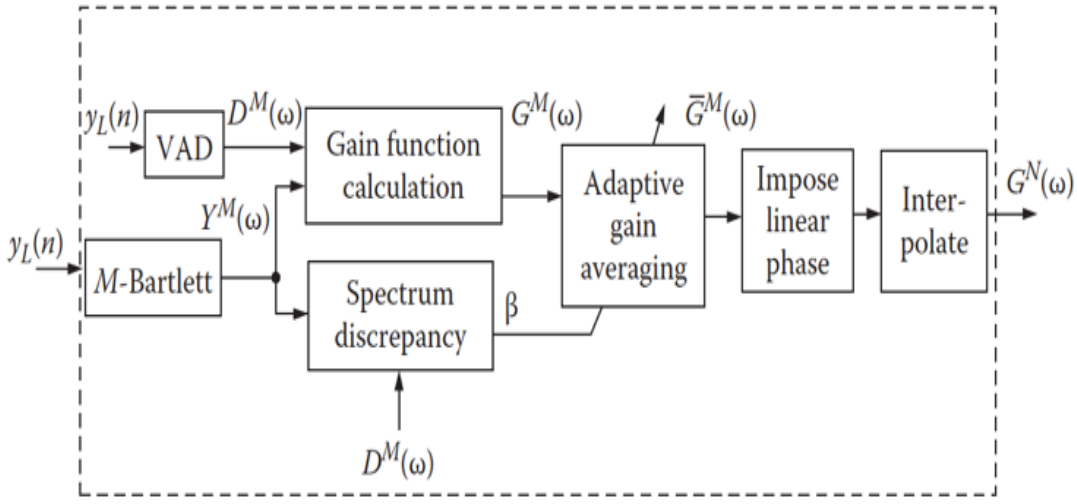

Figure 5.6: Block diagram of the spectral subtraction algorithm with adaptive gain averaging

$\hat{D}_i^{(M)}(w)$ specifies the estimated magnitude spectrum updated during speech segments for which speech is not available. In order to decrease the variable of the gain function, $G_i^{(M)}(w)$ is averaged over time as shown in the following equation:

$$\bar{G}_i^{(M)}(w) = a_i \bar{G}_{i-1}^{(M)}(w) + (1 - a_i) G_i^{(M)}(w) \tag{5.57}$$

where $\bar{G}_i^{(M)}(w)$ indicates the smoothed gain function in frame i and a_i is designated as the adaptive smoothing parameter.

The adaptive mean parameter a_i is extracted from a spectral inconsistency measure identified as β_i as follows:

$$\beta_i = min\Big(\frac{\sum_{w_k}^{M-1}\Big|\big|\bar{G}_i^{(M)}(w_k)\big| - \big|\hat{D}_i^{(M)}(w_k)\big|\Big|}{\sum_{w_k}^{M-1}\big|\hat{D}_i^{(M)}(w_k)\big|}, 1\Big) \tag{5.58}$$

According to the background noise, the previous inconsistency criterion marks a spectral change roughly. Small inconsistency value recommends a relatively stationary background noise while a large inconsistency value suggests situations where highly variable background noise is present or speech. The adaptive averaging parameter a_i is computed from β_i as follows as:

$$a_i = \begin{cases} \gamma a_{i-1} + (1-\gamma)(1-\beta_i) & if \quad a_{i-1} < 1-(1-\beta_i) \\ 1-\beta_i & Otherwise \end{cases} \tag{5.59}$$

where γ is specified as a smoothing parameter ($\gamma = 0.8$ [5]). The rapid reduction of the adaptive parameter allows the gain function to quickly adapt to the new input signal, but allows it to gradually increase. Following a mean operation of (5.56), a linear phase is applied to the gain function to produce a causal filter. These results lead to a filter with the following time-domain symmetry [33]:

$$g_M(n) = \pm g_M(M-1-n) \quad n = 0, 1, \ldots, M-1 \tag{5.60}$$

where $g_M(n)$ denotes the gain function and $\bar{G}_i^{(M)}(w)$ denotes the IFT. The original frame length (L) of the obtained filter has a delay of $(M-1)/2$ on one part. Finally, after creating a linear phase in the gain function, interpolate $\bar{G}_i^{(M)}(w)$ from the function at $M-$point to form the $N-$point function, where N is the size of the FFT. It has not been forgotten that N is chosen to be $N > L + M$ and therefore circular convolution effects are avoided. The resulting output signal is described below, obtained by calculating the inverse FFT:

$$\hat{X}_i^{(M)}(w) = \bar{G}_i^{(N)}(w) \cdot Y_i^{(N)}(w) \tag{5.61}$$

where

$\hat{X}_i^{(M)}(w)$ is the enhanced complex spectrum

$\bar{G}_i^{(N)}(w)$ is the N-point interpolated gain function

$Y_i^{(N)}(w)$ is the complex spectrum of noisy speech, the L-sample input signal obtained after zero padding

In (5.61), the inverse FFT of $\hat{X}_i^{(M)}(w)$ yields the linear convolution of the gain function $g_M(n)$ $(n = 0, 1, , M1)$ of with the noisy signal $y(n)$. Figure 5.7 shows an example of the average gain function with and without adaptive averaging for the fixed frequency fraction. Figure 5.8 indicates the gain function given in (5.56) with no averaging, and Figure 5.9 indicates the gain function given in (5.58) obtained with adaptive averaging.

The gain function indicates in Figure 5.8 has changed considerably, especially in speech segments where speech is not present (for example, the part at $t < 0.5s$ and $t > 5.5s$). In contrast, the gain function in Figure 5.9 is observed to be relatively smooth. The adaptive mean parameter calculated using (5.59) is shown in Figure 5.10. Large values of are obtained when the spectral inconsistency value is small and suggest constant background noise conditions ($t < 0.5s$ and $t > 5.5s$)

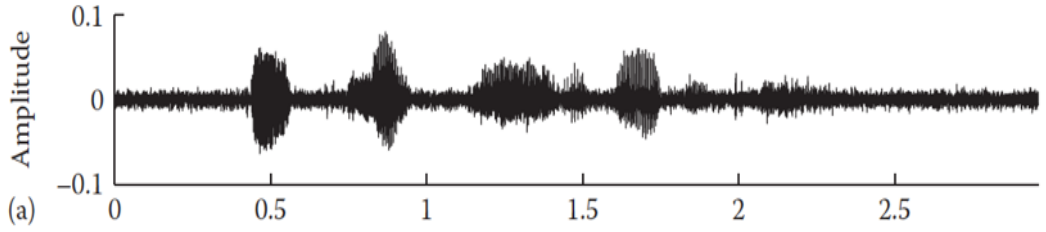

Figure 5.7: A sample noisy sentence. Example gain functions (for a fixed frequency) obtained

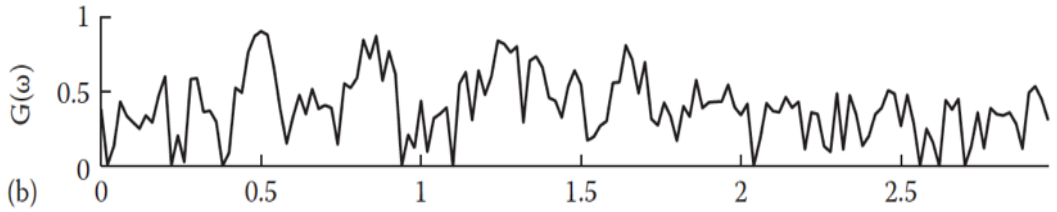

Figure 5.8: Example gain functions (for a fixed frequency) obtained without averaging

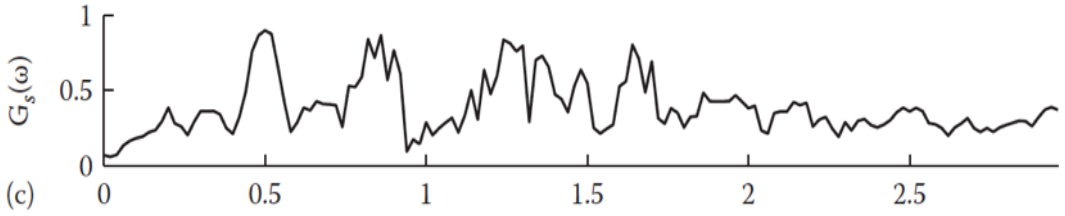

Figure 5.9: Example gain functions (for a fixed frequency) obtained with averaging (Equation 5.57)

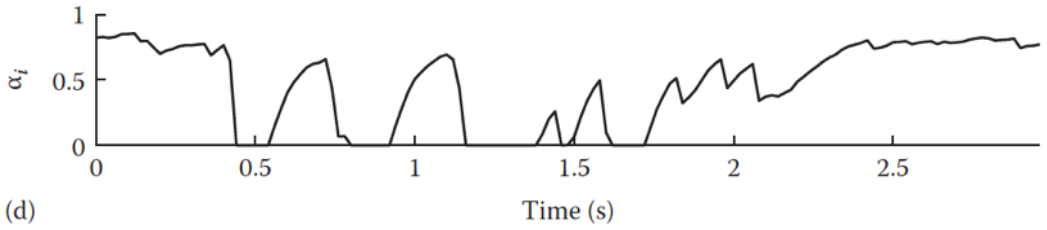

Figure 5.10: The instantaneous values of the smoothing parameter used in the gain averaging.

When speech is present, small values of a_i are obtained (see for example $t \approx 1.25s$ and $t \approx 0.8s$). The following parameters are found in [15] for a good study of speech sampled at 8 kHz: $L = 160$, $M = 32$, and $N = 256$. Although power spectral subtraction can be used in practice, the magnitude spectrum is used in subtraction [15], as shown in (5.56).

Chapter 6

Statistical Model Based Methods

6.1 Introduction

In this section, nonlinear estimators of extreme magnitude (i.e., the module of the DFT coefficients) from the complex spectrum of the signal are discussed using various statistical models and optimization criteria. These nonlinear estimators explicitly add the noise probability density function (PDF) and the speech DFT coefficients to the account, and in some cases the non-Gaussian prior distributions are used.

Usually, by incorporating the probability of the presence of speech into the calculations with soft decision-making gains that add to these estimators, speech enhancement methods have been applied in a statistical estimation framework [30]. When an unknown series of measurements is assigned to a parameter, we wish to find a non-linear estimator of the relevant parameter. In this method, the measurements correspond to a set of DFT coefficients of the noisy signal spectrum, and the relevant parameters form a set of DFT coefficients of the clean signal spectrum. There are many techniques for deriving these nonlinear estimators in the literature of estimation theory [30]. For instance, Bayesian estimators (e.g., MMSE and maximum posteriori estimators) and maximum likelihood estimators (ML) are available. In addition, these estimators differentiate in terms of the assumptions generated, the relevant parameter interest (e.g., deterministic but unknown, random), and the optimization criterion applied.

6.2 Maximum-Likelihood Estimators

The ML [30] approach is also the most popular approach to obtaining practical estimators of statistical estimation theory, and even the most complex estimation problems are often used. This method was first applied to enhance speech by McAulay and Malpass [28]. It is assumed that an N-point data set $y = \{y(0), y(1), , y(N1)\}$ is connected to an unknown parameter θ. In the speech enhancement y (observed data set) is specified as the noisy speech magnitude spectrum and in the interest parameter θ. It also refers to the PDF of y, denoted by $p(\mathbf{y}; \theta)$. The PDF of y is denoted by the unknown θ parameter and denoted by a semicolon. The θ parameter affects the probability of y, the values of θ are subtracted from the observed values of y; the following question can be asked against this statement: *Is θ the largest value produced y which is the observed data?* Mathematically, the maximizing θ of $p(\mathbf{y}; \theta)$ can be investigated as given in the following equation,

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(y; \theta) \tag{6.1}$$

The previous estimate $\hat{\theta}_{ML}$, is the expressed ML estimate of θ. The PDF $p(\mathbf{y}; \theta)$ is called the probability function, given that it can be seen as a function of an unknown parameter with $\mathbf{y}$ fixed. To find ,$p(\mathbf{y}; \theta)$ is distinguished by θ. The derivative zero is equally set and solved for θ. In some cases, it has been suggested that instead of this, it is more appropriate to find by derivation of the log-likelihood function $p(\mathbf{y}; \theta)$. The θ parameter is not known but must be considered deterministic. This assumption indicates Bayesian maximum likelihood estimation (MLE) when θ is assumed to be random.

The sampled noisy speech signal $y(n) = x(n) + d(n)$ consisting of the clean signal $x(n)$ and the noise signal $d(n)$ is specified. In the frequency domain, sampled noisy speech signal looks like the following:

$$Y(w_k) = X(w_k) + D(w_k) \tag{6.2}$$

for $w_k = 2\pi k/N$ and $k = 0, 1, 2, \cdots, N-1$, where N is the frame length in the samples.

The previous equation can also be expressed in polar form as

$$Y_k e^{j\theta_y(k)} = X_k e^{j\theta_x(k)} + D_k e^{j\theta_d(k)} \tag{6.3}$$

where

$\{Y_k, X_k, D_k\}$ defines the magnitudes,

$\{\theta_y(k), \theta_x(k), \theta_d(k)\}$ indicates the phases at noisy speech signal of frequency coefficient k, clean speech signal, and noise signal separately and respectively.

In the ML approach, recommended by McAulay and Malpass [28], the magnitude and phase spectra of the clean signal, i.e., X_k and $\theta_x(k)$ are unknown but are noted deterministic parameters. The PDF of the noise Fourier transform coefficients $D(w_k)$ is supposed to be zero-mean and complex Gaussian. It is expressed that the real and imaginary parts of $D(w_k)$ are assumed to have $\lambda_d(k/2)$ variances. Depending on these two assumptions, the probability density of investigated noisy speech DFT coefficients, $Y(w_k)$. The probability density of $Y(w_k)$ is also Gaussian, which denotes variance of $\lambda_d(k)$ and the average of $X_k e^{j\theta_x(k)}$ as follows:

$$\begin{aligned} p(Y(w_k); X_k, \theta_x(k)) &= \frac{1}{\pi\lambda_d(k)} exp\Big[- \frac{|Y(w_k) - X_k e^{j\theta_x(k)}|^2}{\lambda_d(k)} \Big] \\ &= \frac{1}{\pi\lambda_d(k)} exp\Big[- \frac{Y_k^2 - 2X_k Re\{e^{j\theta_x(k)} Y(w_k)\} + X_k^2}{\lambda_d(k)} \Big] \end{aligned} \tag{6.4}$$

To obtain the ML estimate of X_k, the maximum $p(Y(w_k); X_k, \theta_x(k))$ with about X_k. However, this is not clear, because $p(Y(w_k); X_k, \theta_x(k))$ is named function of two unknown parameters: the phase and the magnitude. The phase parameter is assumed as an annoyance parameter that can be easily removed by "integrating it out." More specifically, the phase parameter can be reduced by maximizing the following mean probability function:

$$p_L(Y(w_k); X_k) = \int_0^{2\pi} p(Y(w_k); X_k, \theta_x) p(\theta_x) d\theta_x \tag{6.5}$$

The index k is dropped from the phase to provide simplicity. Assuming a homogeneous distribution over $(0, 2\pi)$ for the phase θ_x, that is $p(\theta_x) = \frac{1}{2\pi}$ for $\theta_x \in [0, 2\pi]$ the probability function is as follows as:

$$p_L(Y(w_k); X_k) = \frac{1}{\pi\lambda_d(k)} exp\left[-\frac{Y_k^2 + X_k^2}{\lambda_d(k)}\right] \frac{1}{2\pi} \int_0^{2\pi} exp\left[-\frac{2X_k Re\{e^{j\theta_x(k)} Y(w_k)\}}{\lambda_d(k)}\right] d\theta_k \tag{6.6}$$

The integral of the previous equation is known as the first modified Bessel function is given as

$$I_o(|x|) = \frac{1}{2\pi} \int_0^{2\pi} = exp\big[Re(xe^{-j\theta_x})\big] d\theta_x \tag{6.7}$$

and the likelihood function in (6.6) simplifies as follows,

$$p_L(Y(w_k); X_k) = \frac{1}{\pi\lambda_d(k)} \frac{1}{\sqrt{2\pi \frac{2X_k Y_k}{\lambda_d(k)}}} exp\Big[-\frac{Y_k^2 + X_k^2 - 2X_k Y_k}{\lambda_d(k)}\Big] \tag{6.8}$$

After reproducing the log-likelihood function $p_L(Y(w_k); X_k)$, which is associated with the unknown parameter X_k, and after equalizing the derivative zero, the ML estimate of the size spectrum is obtained as follows:

$$\hat{X}_k = \frac{1}{2}\left[Y_k + \sqrt{Y_k^2 - \lambda_d(k)}\right] \tag{6.9}$$

When the noisy signal phase θ_y is used instead of θ_x, the estimate of the clean signal spectrum, the clean signal is used as an estimate of the signal frequency; it is expressed as:

$$\begin{aligned} \hat{X}(w_k) &= \hat{X}_k e^{j\theta_y} = \hat{X}_k \frac{Y(w_k)}{Y_k} \\ &= \left[\frac{1}{2}\sqrt{\frac{Y_k^2 - \lambda_k(k)}{Y_k^2}}\right] Y(w_k) \end{aligned} \tag{6.10}$$

$\gamma_k \triangleq Y_k^2/\lambda_d(k)$ has a posteriori or evaluated SNR value based on the described expressions, the previous equation can be written as:

$$\begin{aligned}\hat{X}(w_k) &= \left[\frac{1}{2} + \frac{1}{2}\sqrt{\frac{\gamma_k - 1}{\gamma_k}}\right] Y(w_k) \\ &= G_{ML}(\gamma_k) Y(w_k)\end{aligned} \tag{6.11}$$

where $G_{ML}(\gamma_k)$ represents the gain function of the ML estimator. The previous acquiring implies that the signal magnitude and phase (X_k and θ_k) are unknown, but it is deterministic. Defining that the signal and speech DFT coefficients are independent in Gaussian process with zero mean, but the signal variance is expressed as variance $\lambda_x(k)$ that is deterministic, and then we acquired a likelihood function. The signal and noise are denoted to be independent. The variance of $Y(w_k)$, indicated as $\lambda_y(k)$ is expressed by $\lambda_y(x) = \lambda_x(k) + \lambda_d(k)$. Hence, the probability density of $Y(w_k)$ is given by the following equation:

$$p_L(Y(w_k); \lambda_x(k)) = \frac{1}{\pi\lambda_x(k) + \lambda_d(k)} exp\left[- \frac{Y_k^2}{\lambda_x(k) + \lambda_x(k)}\right] \tag{6.12}$$

Maximizing the likelihood function $p_L(Y(w_k); \lambda_x(k))$ with respect to $\lambda_x(k)$, we get as the following equation,

$$\hat{\lambda}_x(k) = Y_k^2 - \lambda_d(k) \tag{6.13}$$

Given that $Y_k^2 \approx \lambda_x(k)$ and $D_k^2 \approx \lambda_d(k)$ and $(Y_k^2 - \lambda_d(k) > 0)$ get an estimate of the signal size spectrum as obtained in the following equation:

$$\hat{X}_k = \sqrt{Y_k^2 - \lambda_x(k)} \tag{6.14}$$

It should be noted that this estimator of X_k is nothing more than a power spectrum subtraction estimator. For this reason, assuming that the original power spectrum subtraction approach is modeled as independent Gaussian random processes of signal and noise Fourier transform coefficients. ML can be derived using principles and signal variance $\lambda_x(k)$ can be calculated but is deterministic.

The estimation of the clean signal spectrum acquired by power spectrum subtraction is calculated as follows in (6.11):

$$\begin{aligned}\hat{X}(w_k) &= \hat{X}_k e^{j\theta_y} = \hat{X}_k \frac{Y(w_k)}{Y_k} \\ &= \sqrt{\frac{Y_k^2 - \lambda_d(k)}{Y_k^2}} Y(w_k)\end{aligned} \tag{6.15}$$

The previous equation for $\gamma(k)$ can be written as:

$$\begin{aligned}\hat{X}(w_k) &= \sqrt{\frac{\gamma_k - 1}{\gamma_k}} Y(w_k) \\ &= G_{PS}(\gamma_k) Y(w_k)\end{aligned} \tag{6.16}$$

The gain function of the power spectral subtraction method is expressed as $G_{PS}(\gamma_k)$. Finally, it should be noted that when $\lambda_x(k)$ in (6.6) replaces the ML estimate, the following equation is obtained,

$$\hat{X}(w_k) = \frac{\lambda_x(k)}{\lambda_x(k) + \lambda_d(k)} Y(w_k) \tag{6.17}$$

we get

$$\begin{aligned}\hat{X}(w_k) &= \frac{Y_k^2 - \lambda_d(k)}{Y_k^2} Y(w_k) \\ &= \frac{\gamma_k - 1}{\gamma_k} Y(w_k) \\ &= G_{PS}^2(\gamma_k) Y(w_k)\end{aligned} \tag{6.18}$$

6.3 Bayesian Estimators

In the previous section, the ML approach has been shown for the parameter estimation and it has been stated here that the interest parameter assumed to be the θ was deterministic but unknown. In this section, it is assumed that θ is a random variable, and thus, a random variable is predicted to occur. This approach is called Bayesian approach because this application is based on Bayes theorem. The primary motivation behind the Bayesian approach is the fact that

when knowledge of θ is identified in advance, i.e., when $p(\theta)$ is known, it is necessary to include the estimator in order to improve the estimation accuracy. Bayes estimators perform better than MLE predictors, because they typically use information in advance. In the other part, the Bayesian concept explains the methods of minimizing the mean square error between the true and estimated magnitude spectra.

6.4 MMSE Estimator

In the previous part, it was stated that the error between the linear model of the clean spectrum and the actual spectrum can be derived by minimizing the error. Many researchers who acknowledge the importance of short-time spectral amplitude (STSA) on speech intelligibility and quality of speech have suggested the most appropriate methods for obtaining spectral amplitudes from noisy observations. Specifically, the most suitable estimators were sought to reduce the mean square error between predicted and actual quantities to a minimum:

$$e = E\{(\hat{X}_k - X_k)^2\} \tag{6.19}$$

where

$\hat{X}_k$ is expressed as an estimate of the spectral magnitude in the frequency domain

The actual magnitude of the clean signal is specified as X_k

Depending on how the expectation is applied, the minimization of (6.19) can be achieved in two ways. $\mathbf{Y}$ is the investigated noisy speech spectrum, $p(\mathbf{Y}; X_k)$, which indicates $[Y(w_0)Y(w_1)\cdots Y(w_{N-1})]$ in the classical MSE approach. In Bayesian MSE approach, expectation is made about common joint PDF $p(\mathbf{Y}; X_k)$ and Bayes MSE is given by

$$\hat{X}_k = \iint (\hat{X}_k - X_k)^2 p(\mathbf{Y}; X_k) d\mathbf{Y} dX_k \tag{6.20}$$

The optimal MMSE estimator is generated to the minimization of the Bayes MSE according to $\hat{X}_k$

$$\begin{aligned}\hat{X}_k &= \int X_k p(X_k|\mathbf{Y}) dX_k = E[(X_k\mathbf{Y})] \\ &= E\big[(X_k\big|[Y(w_0)Y(w_1)\dots Y(w_{N-1})]\big]\end{aligned} \tag{6.21}$$

This expresses the mean of a posteriori PDF of X_k. The posterior PDF $p(\mathbf{Y}; X_k)$ is the clean spectrum amplitudes with the amplitudes acquired after investigating all the data in the posterior PDF. In contrast, the PDF of the clean amplitudes before the data is observed indicates a priori PDF of the X_k, that is $p(X_k)$. The MMSE estimate given in (6.21) is not assumed to be a linear relationship between the observed data and the estimator, but includes knowledge of the probability distributions of speech and noise DFT coefficients.

Assuming we already have knowledge about the distributions of speech and speech DFT coefficients, we can evaluate if we assume that we have averages of the posterior PDF of X_k, i.e., the mean of $p(\mathbf{Y}; X_k)$, and the distributions of speech and sound DFT coefficients.

However, it is difficult to measure the true probability distributions of the speech Fourier transform coefficients because the speech and sometimes the noise signal is not a stationary or ergodic operation. They have often tried to measure the probability distributions [46][47] by examining the long-term behavior of the method. Histograms of the Fourier coefficients have been questioned to measure the relative frequency of the Fourier transform coefficients rather than the actual probability density of the Fourier transform coefficients obtained using large quantities of data [29]. Ephraim and Malah [29] proposed a statistical model using the asymptotic statistical properties of the Fourier transform coefficients to remove these problems [48].

This model makes two assumptions:

1. The real and imaginary parts with Fourier transform coefficients have a Gaussian probability distribution. The averages of the coefficients are zero and the variances of the coefficients replace over time, depending on the instability of speaking.

2. The Fourier transform coefficients are statistically uncorrelated and thus independent.

The Gaussian hypothesis is motivated by the central limit theorem [52] because the Fourier transform coefficients are computed as a sum of N random variables. For example, the calculation of the Fourier transform coefficients of noisy speech, $Y(w_k)$, is expressed as:

$$Y(w_k) = \sum_{n=0}^{N-1} y(n)e^{-jw_k n} = y(0) + a_1 y(1) + a_2 y(2) + \cdots + a_{N-1} y(N-1) \quad (6.22)$$

where

$a_m = exp(-jw_k m)$ are constants

$y(n)$ are the time-domain samples of the noisy speech signals

When the statistically independent random variables $\{y(n)\}_{n=0}^{N-1}$ are independent, $Y(w_k)$ is Gaussian according to the resultant central limit theorem [52]. The central limit theorem applies in the case of speech signals as well as weakly dependent samples of sufficiently separated samples. The unrelated assumption is motivated by the approach of the correlation between different Fourier coefficients as the frame length of the analysis approaches N infinity. The assumption of independence is a direct result of the irrelevant assumption, since the Fourier coefficients are both uncorrelated and Gaussian, then they are also independent. However, in speech applications, due to the stability of the speech signal. This can lead to a degree of correlation of the Fourier transform coefficients. However, overlapping analysis windows are often used in practice. Although such "window overlap" clearly violates the unrelated assumption, the resulting models have proven to be practical and useful in practice.

6.4.1 MMSE Magnitude Estimator

To determine the MMSE estimator, first the posterior PDF of X_k, $p(X_k|Y(w_k))$, is calculated. It was used to determine Bayes' rule as follows:

$$\begin{aligned} p(X_k|Y(w_k)) &= \frac{p(Y(w_k)\big|X_k)p(X_k)}{p(Y(w_k))}) \\ &= \frac{p(Y(w_k)\big|X_k)p(X_k)}{\int_0^\infty p(Y(w_k)\big|x_k)p(x_k)dx_k} \end{aligned} \tag{6.23}$$

where x_k is a realization of the random variable X_k. It should be noted that $p(Y(w_k))$ is a normalization factor required to ensure that $p(X_k\big|Y(w_k))$ integrates with 1. If we assume statistical independence between Fourier transform coefficients, that is,

$$E\big[X_k|Y(w_0)Y(w_1)Y(w_2)\dots Y(w_{N-1})\big] = E[X_k|Y(w_k)] \tag{6.24}$$

and $p(x_k\big|Y(w_k))$, the estimator in (6.21) is simplified as shown in the following equation,

$$\begin{aligned} \hat{X}_k &= E[X_k|Y(w_k)] \\ &= \int_0^\infty x_k p(x_k|Y(w_k))d_k \\ &= \frac{\int_0^\infty x_k p(x_k|Y(w_k))dx_k}{\int_0^\infty p(Y_{w_k})|x_k)p(x_k)dx_k} \end{aligned} \tag{6.25}$$

After that,

$$p(Y(w_k)|X_k)p(X_k) = \int_0^{2\pi} p(Y(w_k)|x_k,\theta_k)p(x_k,\theta_k)d\theta_k \tag{6.26}$$

θ_x is the actualization of the phase random variable of $X(w_k)$ (for clarity, the index k is reduced at θ_x), we get

$$\hat{X}_k = \frac{\int_0^\infty \int_0^{2\pi} x_k p(Y(w_k)|x_k, \theta_x) dx_k}{\int_0^\infty \int_0^{2\pi} p(Y(w_k)|x_k, \theta_x) p(x_k, \theta_x) d\theta_x dx_k} \tag{6.27}$$

Next, $p(Y(w_k)|x_k, \theta_x)$ and $p(x_k, \theta_x)$ need to be estimated. It is known from the statistical model that $Y(w_k)$ is the sum of two zero-complex Gaussian random variables. For this reason, the conditional PDF $p(Y(w_k)|x_k, \theta_x)$ must also be Gaussian:

$$p(Y(w_k)|x_k, \theta_x) = p_D(Y(w_k) - X(w_k)) \tag{6.28}$$

where $p_D(\cdot)$ is expressed as the PDF of noisy Fourier transformation coefficients $D(w_k)$. The previous equation is then obtained by

$$p(Y(w_k)|x_k, \theta_x) = \frac{1}{\pi \lambda_d(k)} exp\Big\{ - \frac{1}{\lambda_d(k)} |Y(w_k) - X(w_k)|^2 \Big\} \tag{6.29}$$

Here, the variance of the kth noise spectral component k is given as $\lambda_d(k) = E\big\{|D(w_k)|^2\big\}$. For complex Gaussian random variables, it is known that the random variables (X_k) and phase $(\theta_x(k))$ of $X(w_k)$ are independent of the random variables [52], and for this reason the joint PDF $p(x_k, \theta_x)$ such as the product of the exclusive PDFs, that is, $p(x_k, \theta_x) = p(x_k)(\theta_x)$.

The PDF of X_k is Rayleigh, because $X_k = \sqrt{r(k)^2 + i(k)^2}$, over here $r(k) = Re[X(w_k)]$ and $i(k) = Im[X(w_k)]$ are Gaussian random variables [52]. $\theta_x(k)$ equals PDF uniformly in interval $(-\pi, \pi)$ and hence the common probability $p(x_k, \theta_x)$ is expressed in an equation as follow as

$$p(x_k, \theta_x) = \frac{x_k}{\pi \lambda_x(k)} exp\Big\{ - \frac{x_k^2}{\lambda_x(k)} \Big\} \tag{6.30}$$

where $\lambda_x(k) = E\big\{|X(w_k)|^2\big\}$ denotes the variance of the kth spectral parameter of the clean signal.

By replacing (6.29) and (6.30) with (6.25), the MMSE magnitude estimator was obtained:

$$\hat{X}_k = \sqrt{\lambda_k}\Gamma(1.5)\Phi(-0.5, 1; -v_k) \tag{6.31}$$

where $\Gamma(\cdot)$ is the gamma function, $\phi(a, b; c)$ is called the confluent hypergeometric function, and λ_k is given by

$$\lambda_k = \frac{\lambda_k}{\lambda_x(k) + \lambda_d(k)} = \frac{\lambda_x(k)}{1 + \xi_k} \tag{6.32}$$

and v_k is described by

$$v_k = \frac{\xi_k}{1 + \xi_k}\gamma_k \tag{6.33}$$

where γ_k and ξ_k are described by

$$\gamma_k = \frac{Y_k^2}{\lambda_d(k)} \tag{6.34}$$

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \tag{6.35}$$

The parameters of ξ and γ_k are called a priori and posteriori SNR, respectively. A priori SNR ξ can be viewed as the true SNR of the kth spectrum component, as a posteriori SNR γ_k is assumed to be the noticed or evaluated SNR after the noisy addition of the spectral component. Equation 6.31 can also be written as

$$\hat{X}_k = \frac{\sqrt{v_k}}{\gamma_k}\Gamma(1.5)\Phi(-0.5, 1; v_k) \tag{6.36}$$

Because $\sqrt{\lambda_k}$ is simplified by using (6.33) to (6.36)

$$\begin{aligned}\sqrt{\lambda_k} &= \sqrt{\frac{\lambda_x(k)}{1+\xi_k}} \\ &= \sqrt{\frac{\xi_k \lambda_d(k)}{1+\xi_k}} \\ &= \sqrt{\frac{\xi_k Y_k^2}{(1+\xi_k)\gamma_k}\frac{\gamma_k}{\gamma_k}} \\ &= \sqrt{\frac{\xi_k Y_k^2}{(1+\xi_k)}\frac{1}{(\gamma_k)^2}}Y_k \\ &= \sqrt{\frac{v_k}{\gamma_k}}Y_k\end{aligned} \tag{6.37}$$

Finally, the confluent hypergeometric function in (6.36) is written in terms of Bessel functions and the MMSE estimator can be expressed as:

$$\hat{X}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k} exp(-\frac{v_k}{2})\Big[(1+v_k)I_0\Big(\frac{v_k}{2}\Big) + v_k I_0\Big(\frac{v_k}{2}\Big)\Big]Y_k \tag{6.38}$$

where $I_1(\cdot)$ and $I_0(\cdot)$ are expressed as replaced Bessel functions of first and zero order, respectively. Equation (6.38) or (6.36) is preferred over the original (6.30), since the estimated size is expressed as a gain function, such as $\hat{X}_k = G(\xi_k, \gamma_k)Y_k$.

The spectral gain function $G(\xi_k, \gamma_k)$

$$G(\xi_k, \gamma_k) = \frac{\hat{X}_k}{Y_k} = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k} exp(-\frac{v_k}{2})\Big[(1+v_k)I_0\Big(\frac{v_k}{2}\Big) + v_k I_0\Big(\frac{v_k}{2}\Big)\Big] \tag{6.39}$$

is signified as a function of two parameters: a posteriori SNR and a value γ_k a priori SNR ξ_k. For large values of the instantaneous SNR, the MMSE acquisition function is given by the following equation:

$$G_w(\xi_k) = \frac{\xi_k}{1+\xi_k} \tag{6.40}$$

6.4.2 Estimating the a Priori SNR

The MMSE amplitude estimate in (6.36) has been regenerated assuming that there is an SNR and a noise variance ($\lambda_d(k)$) in advance. Nevertheless, in practice, only the noisy speech signal can be accessed. However, the estimation of ξ_k is more difficult to make.

First, Ephraim and Malah [29] investigated the sensitivity of the amplitude estimator for the errors of a priori SNR ξ_k. They found that the MMSE estimator is insensitive to small fluctuations of ξ_k. Furthermore, from another perspective, it should be assumed that the MMSE estimator is sensitive to low estimates and the a priori SNR ξ_k is greater. Several different methods have been proposed for estimating the a priori SNR ξ_k [29][51][53]. Many of these methods have been described as extensions and improvements of the methods recommended in [30].

6.4.3 Maximum-Likelihood Method

This method [29], which is expressed as an ML approach, is used to estimate the SNR ξ_k of $\lambda_x(k)$, which is firstly determined as unknown and accepted as a deterministic parameter. When $\lambda_x(k)$ is given and $\lambda_d(k)$ is assumed to be estimated during non-speech activity, (6.35) is used to obtain ξ_k.

This estimate is based on L, which is the past consecutive observations of the noisy speech magnitudes obtained in the mth analysis and the frequency coefficient k of $\mathbf{Y}_k(m) \triangleq \big\{Y_k(m), Y_k(m-1), \ldots, Y_k(m-L+1)\big\}$. Assuming statistical independence between L observations and using the Gaussian statistical model, the following probability function is constructed:

$$p(\mathbf{Y}_k(m); \lambda_x(k), \lambda_d(k)) = \prod_{j=0}^{L=1} \frac{1}{\pi(\lambda_x(k) + \lambda_d(k))} exp\bigg(- \frac{Y_k^2(m-j)}{\lambda_x(k) + \lambda_d(k)} \bigg) \quad (6.41)$$

The previous likelihood function is maximized with respect to $\lambda_x(k)$ and $\lambda_d(k)$ is estimated for frame m:

$$\hat{\lambda}_x(k,m) = \begin{cases} \frac{1}{L}\sum_{j=0}^{L-1} Y_k^2(m-j) - \lambda_d(k,m) & if \quad nonnegative \\ 0 & else \end{cases} \tag{6.42}$$

$$= max\left(\frac{1}{L}\sum_{j=0}^{L-1} Y_k^2(m-j) - \lambda_d(k,m), 0\right)$$

After dividing both sides by $\lambda_d(k,m)$, the following equation is obtained

$$\hat{\xi_k}(m) = max\left(\frac{1}{L}\sum_{j=0}^{L-1} \gamma_k^2(m-j) - 1, 0\right) \tag{6.43}$$

where represents $\gamma_k(m) \overset{\Delta}{Y_k^2}(m)/\lambda_d(k,m)$ the posteriori SNR of frame m, and max $(\cdot)$ operator is applied so that ξ_k is always nonnegative. In practice, the moving average is replaced by a recursive averaging operation of (6.43)

$$\bar{\gamma}_k(m) \overset{\Delta}{=} a\bar{\gamma}_k(m-1) + (1-a)\frac{\bar{\gamma}_k(m)}{\beta} \tag{6.44}$$

In the range given here $0 \leq a < 1$ is the correction constant expressed and $\beta \geq 1$ means a correction factor. The final ML estimator of ξ has the form

$$\hat{\xi_k}(m) = max(\hat{\gamma}(m) - 1, 0) \tag{6.45}$$

Interestingly, when $L = 1$, the previous ML estimator $\xi_k(m) = max(\gamma(m) - 1, 0)$ only produces a gain function depending on γ_k.

6.5 Implementation and Evaluation of the MMSE Estimator

The minimum mean square error algorithm has been reviewed and applied in the following four basic steps. For each speech frame separated by windows, the following steps are applied:

Step 1: The DFT of the noisy speech signal is calculated: $(w_k) = Y_k exp(j\theta_y(k))$

Step 2: λ_k is predicted as a posteriori term γ_k as $\gamma_k = Y_k^2/\lambda_d(k)$ over the power spectrum of the noise signal calculated during the non-speech activity. Thereafter, ξ_k is calculated by applying (6.45).

Step 3: Applying (6.38), the enhanced signal amplitude $\hat{X}_k$ is estimated.

Step 4: The enhanced signal spectrum must be constructed as $\hat{X}_k = \hat{X}_k exp(j\theta_y(k))$ and the inverse DFT of $\hat{X}_k(w_k)$ must be computed to obtain the enhanced time-domain signal $\hat{x}(n)$ corresponding to a given input speech frame.

The Hamming-window was opened prior to the DFT analysis of the speech signal [6]. In each frame, the overlap and add method is used to synthesize the enhanced signal $\hat{x}(n)$. In each frame, the overlap and add method is used to synthesize the enhanced signal . By explaining the performance of the MMSE algorithm [29], the effects on the development of spectral subtraction and ML algorithms have been explored and compared [29]. Enhanced speech had residual noise; that is, the value produced in the spectral extraction of the speech signal, was not the residue "musical." On the contrary, when a priori SNR was estimated using the ML approach in (6.45), "musical noise" was obtained in the enhanced speech. Furthermore, the MMSE estimator gave a similar speech quality with a priori SNR ($a = 0.725$ and $\beta = 2$ $a = 2$ in (6.44)) estimated when using the ML approach.

Chapter 7

Experimental Work

7.1 Subjective Listening Tests

The subjective listening test [55] helps people to compare and evaluate the quality of their speech by providing listening to recorded speech. It is known that listeners of recorded voices can evaluate the same speech differently. However, when a sufficiently large sample size is applied, the corresponding subjectivity average can be obtained. Enhanced speech listening is intended for human ears, and perhaps this is the best evaluation method. The major disadvantage of these tests is that they are very tedious and time-consuming to perform subjective listening tests. The two most commonly used subjective tests are the MOS and the paired listening test.

7.2 Mean Opinion Score Test

A subjective test based on MOS is also carried out on some selected utterances. The results of the subjective test were also compared with those of the objective test to determine the most appropriate objective measure for the evaluation of speech enhancement algorithms. The strengths and weaknesses of the various algorithms are analyzed and compared. The first objective is by far the most common focus of most researchers in the area. The quality of speech, i.e., how pleasant it sounds to a human listener, is a high subjective measurement. On the other hand, intelligibility, or how much information can be removed from a speech, is an objective measure. Quality can be measured using the Mean Opinion Score (MOS) in which a listener rates the quality of a speech from 1 to 5. Intelligibility analyses are conducted differently as the emphasis is on the understanding of

speech. In such tests, listeners are asked to listen to several sentences or separate words, and they should write down the words they can recognize. Based on the percentage of correctly recognized words, an intelligibility score is obtained.

The mean opinion score (MOS) is the most commonly used global method for evaluating the user's recorded speech data. In this method, listeners can rate speech that is individually tested at all on a numeric scale of five based on their auditory impressions. The MOS scale is given in Figure 7.1 The measurements required for the test are regularly made with a group of auditors and then the scores are averaged to obtain the final MOS. Sample test speeches with known MOS values is provided for the good and unbiased evaluation of auditors where normalization may be performed to reduce the prejudice of the audiences concerning the MOS values being high or low. MOS rating of 3 and above is considered to be" toll quality." However, relative scores in comparing speech enhancement algorithms are more important than total ratings. A detailed explanation of how this particular subjective test is conducted in ITU-T Recommendation P.830 is given.

Rating	Speech Quality	Level of Distortion
5	Excellent	Imperceptible
4	Good	Just perceptible but not annoying
3	Fair	Perceptible and slightly annoying
2	Poor	Annoying but not objectionable
1	Unsatisfactory/Bad	Very annoying and objectionable

Figure 7.1: Mean Opinion Score five-point scale in table

7.3 Comparison of Algorithms using MOS

34 different sound sample that used in SYMPES code are treated with the geometric approach and Richard Hendriks algorithms. It is observed that it is not possible to evaluate outputs objectively. Then MOS test has been decided as a subjective evaluation method. In the test, 40 people listened to 34 sound samples with different sampling frequencies, 16 kbps, 32 kbps, 64 kbps and voted them from 1 to 5 as worst to best. The $Sympes16+GA$, $Sympes16+hen$, $Sympes16$, $Sympes32+GA$, $Sympes32+hen$, $Sympes32$, $Sympes64+GA$, $Sympes64+hen$, Sympes64 averages were calculated.

Average Ratings per Output									
Filename	Sympes16+GA	Sympes16+He	Sympes16	Sympes32+GA	Sympes32+Hen	Sympes32	Sympes64+GA	Sympes64+Hen	Sympes64
Overall Average	4.732	4.582	3.791	3.556	3.275	2.516	2.252	1.938	1.029
Speaker1_1_Original.wav	4.556	4.333	3.667	3.333	2.889	2.444	2.000	1.889	1.000
Speaker1_2_Original.wav	4.778	4.556	3.778	3.222	2.667	2.444	2.000	1.778	1.000
Speaker1_3_Original.wav	4.778	4.667	3.889	3.667	3.333	2.778	2.444	1.889	1.111
Speaker1_4_Original.wav	4.778	4.556	3.778	3.667	3.333	2.778	2.556	1.889	1.111
Speaker1_5_Original.wav	4.778	4.778	4.000	3.667	3.778	2.667	2.000	1.889	1.000
Speaker1_6_Original.wav	4.778	4.778	4.000	3.667	3.778	2.667	2.111	2.000	1.000
Speaker2_1_Original.wav	4.889	4.778	4.000	3.667	3.667	2.556	2.667	2.000	1.000
Speaker2_2_Original.wav	4.889	4.667	3.889	3.667	3.222	2.778	2.778	2.111	1.000
Speaker2_3_Original.wav	4.889	4.667	4.000	3.778	3.222	2.778	2.222	2.111	1.222
Speaker2_4_Original.wav	4.889	4.778	4.000	3.889	3.222	2.889	2.000	2.000	1.111
Speaker2_5_Original.wav	4.889	4.667	3.889	3.778	3.778	2.667	2.556	2.000	1.000
Speaker3_1_Original.wav	4.778	4.556	3.889	3.667	3.667	2.667	2.667	1.889	1.000
Speaker3_2_Original.wav	4.778	4.667	3.778	3.556	3.444	2.556	2.000	1.889	1.000
Speaker3_3_Original.wav	4.889	4.667	3.778	3.667	3.667	2.667	2.111	1.889	1.000
Speaker3_4_Original.wav	4.889	4.667	3.778	3.778	3.000	2.444	2.000	1.889	1.000
Speaker3_5_Original.wav	4.778	4.556	3.556	3.667	3.222	2.556	2.111	1.889	1.000
Speaker4_1_Original.wav	4.889	4.667	3.556	3.667	3.111	2.556	2.111	2.000	1.000
Speaker4_2_Original.wav	4.778	4.667	3.889	3.778	3.111	2.667	2.111	2.111	1.111
Speaker4_3_Original.wav	4.889	4.667	3.667	3.556	3.000	2.667	2.000	2.000	1.111
Speaker4_4_Original.wav	4.778	4.667	3.889	3.778	3.111	2.667	2.222	1.889	1.000
Speaker4_5_Original.wav	4.000	3.889	3.444	3.444	3.778	2.333	1.889	1.444	1.000
Speaker5_1_Original.wav	4.778	4.111	3.667	3.111	2.778	2.222	2.444	1.556	1.000
Speaker5_2_Original.wav	4.111	3.889	3.556	3.556	3.556	2.556	1.889	1.889	1.000
Speaker5_3_Original.wav	4.222	4.111	3.222	2.889	3.778	2.111	2.111	2.111	1.111
Speaker5_4_Original.wav	4.778	4.667	3.778	3.111	3.000	2.000	2.556	2.000	1.000
Speaker5_5_Original.wav	4.778	4.778	3.778	3.111	2.889	2.000	2.444	2.444	1.000
Speaker5_6_Original.wav	4.667	5.000	4.000	3.222	3.333	2.333	2.111	2.111	1.111
Speaker5_7_Original.wav	4.778	4.333	4.000	3.778	3.333	2.556	2.222	2.000	1.000
Speaker6_1_Original.wav	4.889	4.111	3.667	3.333	2.778	2.111	2.111	1.667	1.000
Speaker6_2_Original.wav	5.000	4.889	4.111	3.778	3.111	2.667	2.444	2.111	1.000
Speaker6_3_Original.wav	4.667	4.667	3.889	3.667	3.444	2.111	1.889	1.667	1.000
Speaker6_4_Original.wav	4.333	4.556	3.778	3.556	3.111	2.556	2.556	1.889	1.000
Speaker6_5_Original.wav	4.667	4.889	3.889	3.889	3.111	2.667	2.556	2.000	1.000
Speaker6_6_Original.wav	4.889	4.889	3.444	3.333	3.111	2.444	2.667	2.000	1.000

Figure 7.2: Average Ratings per Output in Table

Each result of MOS test was given in Figure 7.2. As seen in the table, for 16 kbps the best results are obtained at $Sympes16 + GA$ with score 5.000. The worst results are obtained at Sympes16 with 3.225. As seen in the table, for 32 kbps the best results are achieved at $Sympes_32 + GA$ with score 3.889. The worst results are obtained at Sympes32 with 2.000.

Moreover, the spectrum analyzes of the best, the mid and the worst averages for 16 kbps were given in figures below. Red areas represent high frequencies signal with high powers. The graphs are colored from green to red representing the power of the signal from low to high. The high powers signals in low frequencies in the spectrum of the original signal (Figure 7.4) are observed strongly in Hendrikss spectrum above 0,2-0,3 frequency (Figure 7.6).

In other words, background noise remains still exist due to distortion in Hendrikss spectrum. On the other hand, the high power signals are rarely observed in GA and SYMPES spectrum. It is seen that sound is improved in mid and high frequencies, therefore, it is better than GA as also seen in MOS test result.

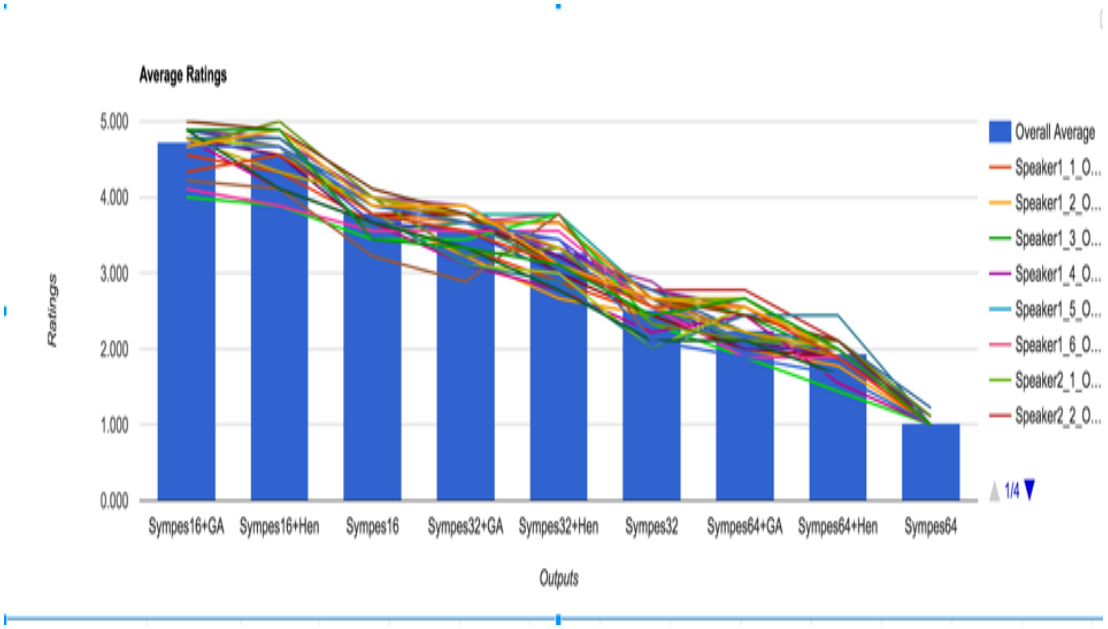

Figure 7.3: Average Ratings of MOS Score

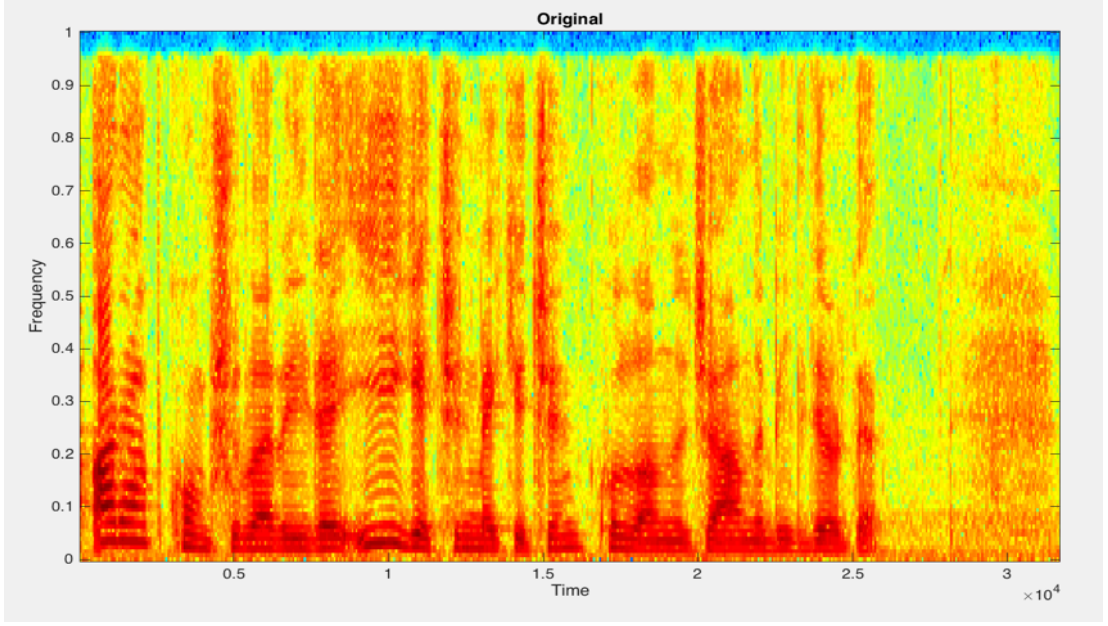

Figure 7.4: Spectrum Analyze of 16kbps original speech signal

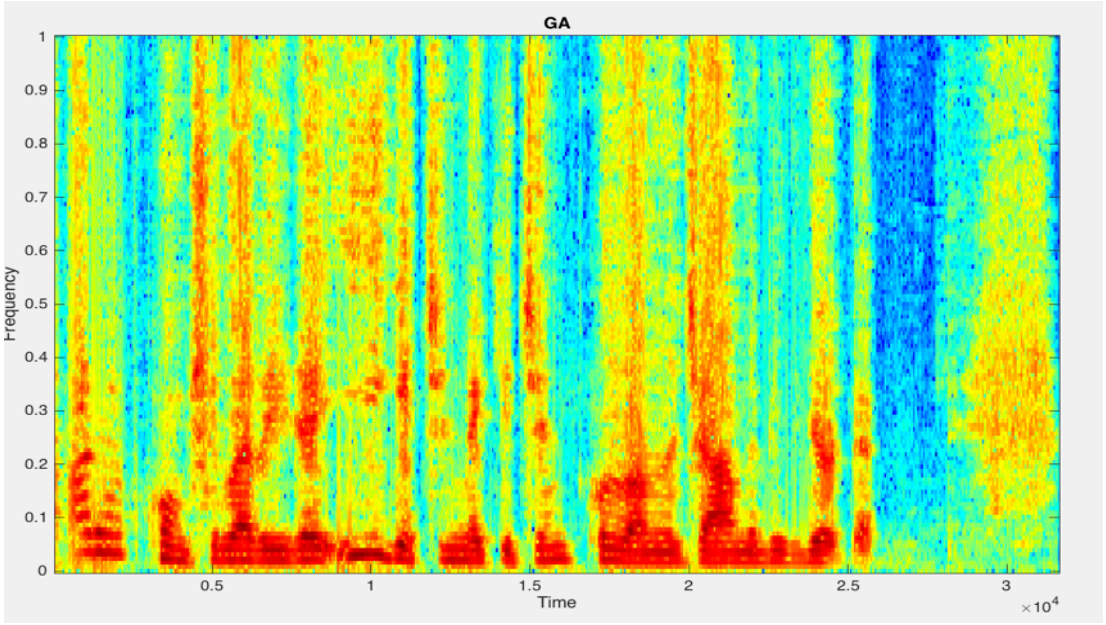

Figure 7.5: Spectrum Analyze of 16kbps SYMPES16+GA (Geometric Approach)

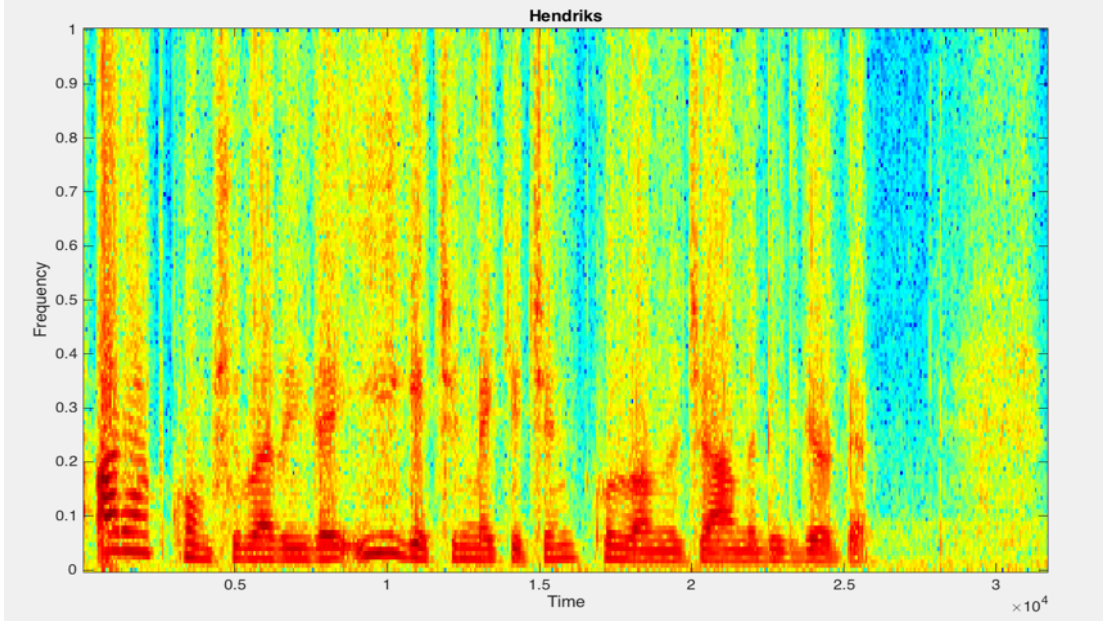

Figure 7.6: Spectrum Analyze of 16kbps SYMPES16+hen (Hendriks Approach)

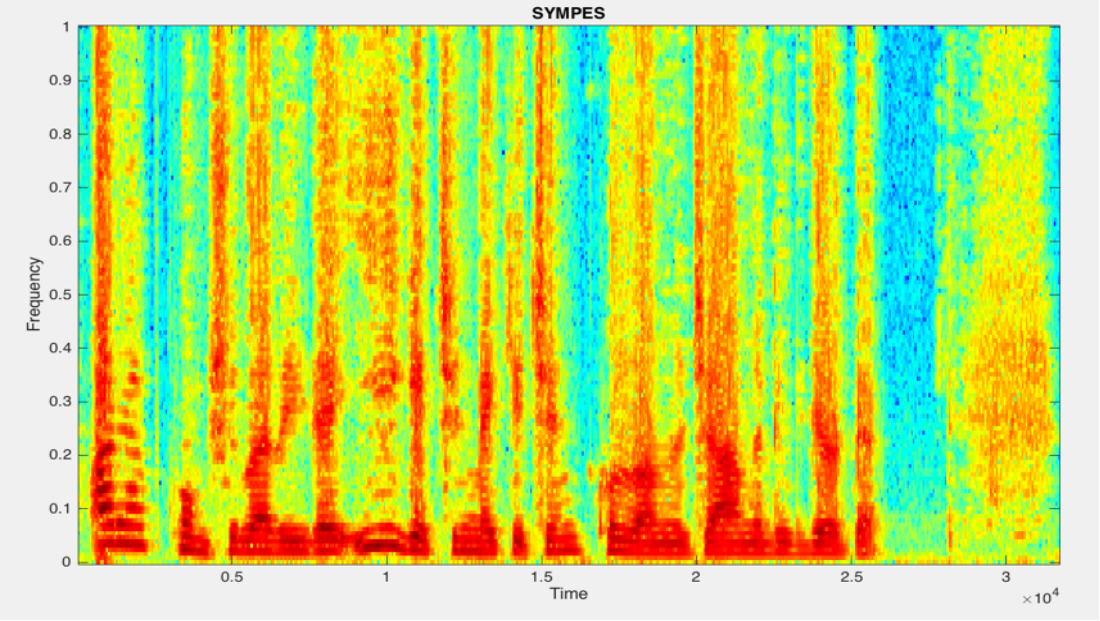

Figure 7.7: Spectrum Analyze of 16kbps SYMPES16 (Speech Coding)

Chapter 8

Conclusion

In this study, the coding method SYMPES, which is used to compress speech signals is used. The SYMPES method used is important to enhance speech signals in processing and storing the intended speech signals in modern communication systems. The SYMPES method used is important in estimating speech signals in processing and storing the expected speech signals in modern communication systems. In particular, as a result of modeling or reconstruction of speech signals, reduces the amount of information needed for compressed signals. The compression and transmission of digital speech signals is considerably improved by compressing the speech data. However, compression results in data loss or background noise.

This study has attempted to improve the distortion of the background that was made without being compressed. The speech enhancement algorithms used in this study have been done with speech enhancement algorithms; spectral subtraction algorithm and statistical-based model, which give very useful results to improve the distortion of the background. These algorithms are explained in detail, and the differences between them are examined and tested to determine which is better. However, a subjective test was used instead of an objective test. It is seen that the enhanced signals are distorted in the waveform. It has been found that the SNRs of the output of the signals that are improved after being compressed due to the distortion of this waveform will not be a good result. For this reason, the MOS test, which is a subjective test, is compared with the two speech enhancement algorithms mentioned above.

Thirty-four compressed speech records using SYMPES coding were enhanced using speech enhancement algorithms. These enhanced sounds were listened to by 40 people for 16 kbps, 32 kbps, and 64 kbps sampling frequency outputs for MOS

testing. In the MOS test [Section 7], the best was scored 5 and the worst as 1 by the listeners. According to the MOS test results shown in Table ... the best enhancement is made by the spectral subtraction algorithm. The best average result for 16 kbps is 5.00 for SYMPES16 + GA and the worst is 3.222 in the MOS test. The other enhancement algorithm is best at 2.889 and the worst at 1.444 for the study of Hendriks.

In addition, spectral analysis graphs show that the GA noise is better enhanced at mid and high frequencies than the study of Hendriks. Thus, the explanation for the above spectrum analysis graph proves the results of the MOS test.

References

[1] U. Guz, H. Gurkan, and B. S. Yarman, "A new speech modeling method: Sympes," pp. 4 pp.–, May 2006.

[2] U. Guz, H. Gurkan, and S. Yarman, "A novel noise robust and low bit rate speech coding algorithm," pp. 471–474, Sept 2009.

[3] Y. Lu and P. C. Loizou, "A geometric approach to spectral subtraction," *Speech Commun.*, vol. 50, no. 6, pp. 453–466, Jun. 2008.

[4] B. S. Yarman, Ü. Güz, and H. Gürkan, "On the comparative results of sympes: A new method of speech modeling," *AEU-International Journal of Electronics and Communications*, vol. 60, no. 6, pp. 421–427, 2006.

[5] U. Guz, H. Gurkan, and B. S. Yarman, "A novel noise robust and low bit rate speech coding algorithm," in *Computer and Information Sciences, 2009. ISCIS 2009. 24th International Symposium on.* IEEE, 2009, pp. 471–474.

[6] P. Loizou, "Speech enhancement: Theory and practice," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 113–120, Apr 2007.

[7] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoust*, pp. 359–366, 2001.

[8] M. Yektaeian and R. Amirfattahi, "Comparison of spectral subtraction methods used in noise suppression algorithms," in *2007 6th International Conference on Information, Communications Signal Processing*, Dec 2007, pp. 1–4.

[9] R. Martin, "Speech enhancement based on minimum mean-square error estimation and supergaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, Sept 2005.

[10] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, no. 4, pp. 113–116, April 2002.

[11] C. Breithaupt and R. Martin, "Mmse estimation of magnitude-squared dft coefficients with supergaussian priors," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 1, April 2003, pp. I–896–I–899 vol.1.

[12] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, vol. 2, Mar 1999, pp. 789–792 vol.2.

[13] X. Hou, S. Guo, H. Cui, K. Tang, and Y. Li, "Speech enhancement for non-stationary noise environments," in *2009 International Conference on Information Engineering and Computer Science*, Dec 2009, pp. 1–3.

[14] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, Jan 2002.

[15] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 8, pp. 799–807, Nov 2001.

[16] Ü. Güz, B. Yarman, and H. Gürkan, "A new method to represent speech signals via predefined functional bases," in *Proceedings of the IEEE European Conference on Circuit Theory and Design*, vol. 2, 2001, pp. 5–8.

[17] U. Guz, H. Gurkan, and B. S. Yarman, "A novel method to represent the speech signals by using language and speaker independent predefined functions sets," in *Circuits and Systems, 2004. ISCAS'04. Proceedings of the 2004 International Symposium on*, vol. 3. IEEE, 2004, pp. III–457.

[18] Ü. Güz, H. Gürkan, and B. S. Yarman, "A new coding method for speech and audio signals," in *Signal Processing Conference, 2005 13th European*. IEEE, 2005, pp. 1–4.

[19] H. R. Akansu AN, in *Multiresolution signal decomposition, transforms, subbands, wavelets.*, San Diego, USA, 1992.

[20] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.

[21] B. R. Pearsons, K. and S. Fidell, "Speech levels in various noise environments, technical report," *Philosophical Magazine*, 1977).

[22] J. Lim, A. Oppenheim, and L. Braida, "Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 4, pp. 354–358, 1978.

[23] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar 1999.

[24] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, Oct 1992.

[25] Q. Lin, E.-E. Jan, and J. Flanagan, "Microphone arrays and speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 622–629, 1994.

[26] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[27] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[28] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 2, pp. 137–145, Apr 1980.

[29] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[30] H. Van Trees, in *Detection, Estimation and Modulation Theory.* Wiley, 1968.

[31] K. Paliwal and A. Basu, "A speech enhancement method based on kalman filtering," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '87.*, vol. 12, Apr 1987, pp. 177–180.

[32] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec 1979.

[33] F.-M. Wang, P. Kabal, R. P. Ramachandran, and D. O'Shaughnessy, "Frequency domain adaptive postfiltering for enhancement of noisy speech," *Speech Communication*, vol. 12, no. 1, pp. 41 – 56, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939390017F

[34] B. T. Quackenbush, S. and M. Clements, in *Objective measures of speech quality.* Prentice-Hall, 1988.

[35] P. T. Weiss M, Aschkenasy E, "Nicolet scientific corporation; 1974. study and the development of the intel technique for improving speech intelligibility," 1974.

[36] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, Apr 1979.

[37] K. K. Paliwal and L. D. Alsteris, "On the usefulness of {STFT} phase spectrum in human listening tests," *Speech Communication*, vol. 45, no. 2, pp. 153 – 170, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639304000950

[38] L. P. Niederjohn, R. and F. Josse, "Factors related to spectral subtraction for speech in noise enhancement," *SPIE*, p. 985996., 1987.

[39] P. Vary and M. Eurasip, "Noise suppression by spectral magnitude estimation : Mechanism and theoretical limits," *Signal Processing*, vol. 8, no. 4, pp. 387 – 400, 1985. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0165168485900027

[40] 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.*, vol. 4, Apr 1979, pp. 208–211.

[41] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection, for robust speech recognition in cars," *Speech Communication*, vol. 11, no. 2, pp. 215 – 228, 1992. [Online]. Available: http://www.sciencedirect.com/science/article/pii/016763939290016Z

[42] B. L. Sim, Y. C. Tong, J. S. Chang, and C. T. Tan, "A parametric formulation of the generalized spectral subtraction method," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 328–337, Jul 1998.

[43] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, Apr 1985.

[44] J. G. Proakis and D. G. Manolakis, "Digital signal processing 3 rd edition," 1996.

[45] S. K. Sengijpta, "Fundamentals of statistical signal processing: Estimation theory," 1995.

[46] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma distributed speech priors," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–253.

[47] T. Lotter and P. Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.

[48] W. Pearlman and R. Gray, "Source coding of the discrete fourier transform," *IEEE Transactions on information theory*, vol. 24, no. 6, pp. 683–692, 1978.

[49] C. H. You, S. Koh, and S. Rahardja, "Adaptive β-order mmse estimation for speech enhancement." in *ICASSP (1)*, 2003, pp. 900–903.

[50] A. Papoulis and S. U. Pillai, *Stochastic processes*. McGraw-Hill New York, 1991, vol. 3.

[51] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, 2005.

[52] A. Papoulis and S. U. Pillai, *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education, 2002.

[53] M. K. Hasan, S. Salahuddin, and M. R. Khan, "A modified a priori snr for speech enhancement using spectral subtraction rules," *IEEE signal processing letters*, vol. 11, no. 4, pp. 450–453, 2004.

[54] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2008.

[55] A. W. Rix, M. P. Hollier, J. G. Beerends, and A. P. Hekstra, "Pesq-the new itu standard for end-to-end speech quality assessment," in *Audio Engineering Society Convention 109*. Audio Engineering Society, 2000.