N° Ordre...../Institut/UMBB/2014

**People's Democratic Republic of Algeria**
**Ministry of Higher Education and Scientific Research**
**University M'Hamed BOUGARA – Boumerdes**


**Institute of Electrical and Electronic Engineering**
**Department of Electronics**

Project Report Presented in Partial Fulfilment of
the Requirements of the Degree of

# "MAGISTER"

In **Electrical and Electronic Engineering**
Option: **Electrical and Electronic Systems Engineering**

Title:

# Single Channel Speech Denoising

Presented By:

**ZEGAR Chouki**

**Before the jury composed of:**

| | |
|---|---|
| Pr. BELOUCHRANI Adel (Professeur ENP, Alger) | President |
| Dr. FERGANI Belkacem (MC/A, USTHB, Alger) | Examiner |
| Dr. HAMADOUCHE M'hamed (MC/A, UMBB) | Examiner |
| Dr. CHERIFI Dalila (MC/B, UMBB) | Examiner |
| Dr. DAHIMENE Abdelhakim (MC/A, UMBB) | Supervisor |

AcademicYear: 2013/2014

# Acknowledgments

I would like to express a sincere thanks to my supervisor Dr. A. Dahimene for all his help. He was always there to offer help and guidance whenever there was a problem in the project.

I would also like to thank all the students that took part in helping me test out my project.

Finally , I want to say thanks to all my teachers, family members and friends for all their support throughout this project.

# ملخص

تعزيز نوعية الإشارة الكلامية والحد من الضوضاء لها تطبيقات واسعة في مجال معالجة الكلام. وغالبا ما تستخدم كمرحلة تمهيدية او ما قبل المعالجة في مختلف التطبيقات.

العمل الذي سنقوم به من خلال هذا المشروع يتضمن الخطوات التالية:

1- تقليل نسبة الضوضاء الملونة الخلفية المتضمنة في إشارة الكلام من خلال قناة واحدة من أجل تحسين نوعية محسوسة و كلام واضح ببنية جيدة.

2- تحقيق ما يسمى بالفصل المكفوف لخليط كلامي مشكل من مصادر مختلفة (دون معلومات سابقة عن هاته المصادر) إنطلاقا من خليطين فقط و خالين من إشارات الصدى. وذلك في نطاق الفرضية التي يطلق عليها إسم الإتصال العمودي التقريبي في المعلم الزمني الترددي.

في عالمنا الذي نعيش فيه معظم الضوضاء الموجودة تأتي على شكل ما يسمى بالضوضاء الملونة التي لها تأثير غير منتظم على إشارات الكلام المستعملة على طول مدى الطيف. هذا المشروع سيدرس ستة خوارزميات في إشارة الكلام من خلال قناة واحدة الآتي ذكرها :

- الطرح الطيفي.
- الطرح الطيفي متعدد الطبقات.
- مصفاة وينر.
- أدنى معدل خطأ تربيعي لسعات طيفية في مدات زمنية قصيرة (باستعمال وبدون استعمال المعدِّل الحسَّاس لوجود قطع كلامية).
- أدنى معدل خطأ تربيعي للوغارتم السعات طيفية (باستعمال وبدون استعمال المعدِّل الحسَّاس لوجود قطع كلامية).
- الخوارزمية الفائقة المعدلة للوغارتم السعات طيفية .

الطرق المقترحة تميزَّت بدرجة معتبرة من المرونة قي معالجة و مراقبة مستويات الضوضاء المحذوفة. من خلال النتائج المتحصل عليها تبين أنَّ طريقة المعاجة التمهيدية المقترحة الخوارزمية الفائقة المعدلة للوغارتم السعات طيفية تقدم نتائج أحسن من الطرق الأخرى ذلك إعتمادا على جملة من الإختبارات الموضوعية و الذاتية.

التقنية المستعملة في ما يسمى بالفصل الأعمى لخليط كلامي مكون من مصادر مختلفة هي تقنية تبني وسائطها النسبية لكل مصدر إشارة في الخليط عن طريق حساب النسب الزمنية الترددية المتضمنة في المزيجين اللذان سنعتمدهما في فصل الإشارات الكلامية. إذا إعتبرنا أن كل مصدر إشارة في الخلائط الكلامية متصل متعامد في إيطار نافذة المعاجة، يعني أنه يوجد إشارة مصدر واحد فقط ناشطة في المعلم الزمني الترددي في كل إحداثية تردد-زمن أو لا يوجد على الإطلاق. بعدما أخضعنا هاته التقنية إلى التجربة على خلائط كلامية إصطناعية ذات أعداد مختلفة من المصادر، وجدنا أنه تم الفصل الجيد لهاته المصادر الكلامية وذلك تبعا لإختبارات التقييم المنجزة

**الكلمات المفتاحية:** تعزيز نوعية الإشارة الكلامية، قناة واحدة، الضوضاء الملونة، الضوضاء الموسيقية، الفصل المكفوف لخليط كلامي، إتصال متعامد، إختبارات التقييم الموضوعية و الذاتية.

# Résumé

L’amélioration de la parole et la réduction du bruit ont de larges applications dans le domaine du traitement de la parole. Ils sont souvent employés en tant que prétraitement dans diverses applications.

Le travail à faire dans ce projet est:

1. L’élimination de bruit très non-stationnaire pour les signaux de parole à canal unique afin d'améliorer la qualité perceptible et l'intelligibilité de la parole.
2. La séparation aveugle d'un nombre quelconque de sources vocales multi-locuteurs (sans connaissances sur les sources réelles) étant donné seulement deux mélanges anéchoïques et étant donné l'hypothèse que nous appelons W-orthogonalité disjoint approximative.

Deux points doivent souvent être pris en compte dans les applications de la suppression de bruit pour les signaux: l’élimination du bruit indésirable pour améliorer le rapport signal sur bruit (SNR) et la préservation de la forme et les caractéristiques du signal original.

Le bruit du monde réel est le plus souvent non-stationnaire et n'affecte pas le signal de parole de manière uniforme sur le spectre. Ce projet explore un groupe d'algorithmes à base de DFT comme des techniques de prétraitement pour un débruitage de seul canal de parole qui sont comme suit:

- Soustraction spectrale en utilisant plus de soustraction et le plancher spectral.
- Soustraction spectrale multi bande.
- Filtre de Wiener.
- Minimum de l'Estimation du Carré de la Moyenne de l'amplitude spectrale à court terme (MMSE-STSA) avec, et sans l'utilisation de modificateur SPU.
- Minimum de l'Estimation du Carré de la Moyenne de l'amplitude Log-spectrale (MMSE-LSA) avec, et sans l'utilisation de modificateur SPU.
- l'Estimation optimale modifié de l'amplitude Log-spectrale.

Les algorithmes mis en œuvre fournissent divers degrés élevé de flexibilité et de contrôle sur les niveaux d'élimination de bruit qui réduit les artefacts au niveau de la parole améliorée. Les résultats de l'étude de comparaison sur la base de critères subjectifs et objectifs ont montré que la méthode de l'Estimation optimale modifié de l'amplitude Log-spectrale surpasse tous les algorithmes à base de DFT pour l’amélioration de la parole à canal unique.

La technique utilisée pour la séparation aveugle de sources de parole est basée sur la technique de démixage dégénéré et d'estimation (DUET) qui construit des estimations des

paramètres de mélange relatifs associés à chaque signal en prenant le rapport des représentations temps fréquence des deux mélanges. Si les sources de chaque mélange sont W-disjoints orthogonales, ce qui signifie que seul un signal est actif dans le plan temps-fréquence à un temps-fréquence donnée ou pas de signal du tout. Nous avons mis en œuvre et testé le comportement de la technique de DUET sur des mélanges artificiels instantanées de parole d'un nombre différent de sources, et ils peuvent être séparés parfaitement selon les tests d'évaluation des performances effectués (subjectifs et objectifs).

**Mots clés** : L'élimination de bruit de parole, Canal unique, Bruit non-stationnaire, Bruit musical, Les Techniques à base de DFT, Séparation aveugle de Sources, W-orthogonalité disjoint, les tests Subjectifs et Objectifs.

# Abstract

Speech enhancement and noise reduction have wide applications in speech processing. They are often employed as pre-processing stage in various applications.

The work to be done in this project is:

1. Denoising a single-channel speech signal in the presence of a highly non-stationary background noise in order to improve the perceptible quality and intelligibility of the speech.
2. Blind Multi-speaker speech separation of an arbitrary number of speakers (without knowledge about the actual speakers as speech sources) given just two anechoic mixtures provided the assumption which we call Approximate W-disjoint Orthogonality.

Two points are often required to be considered in signal denoising applications: eliminating the undesired noise to improve the signal to noise ratio (SNR) and preserving the shape and characteristics of the original signal.

Real world noise is mostly highly non-stationary and does not affect the speech signal uniformly over the spectrum. This project explores a set of DFT-based algorithms as single-channel pre-processing techniques which are as follows:

- Spectral Subtraction using over-subtraction and spectral floor.
- Multi-Band Spectral Subtraction (MBSS).
- Wiener Filter.
- MMSE of Short-Time Spectral Amplitude (MMSE-STSA) estimator with, and without using SPU modifier.
- MMSE Log-Spectral Amplitude Estimator with, and without using SPU modifier.
- Optimally-Modified Log-Spectral Amplitude estimator (OM-LSA).

All the implemented algorithms provide considerable, different degrees of flexibility and control on noise elimination levels that reduces artifacts in the enhanced speech, resulting in the improved quality, and intelligibility. The comparison study results based on subjective and objective tests showed that the Optimally Modified Log-Spectral Amplitude Estimator (OM-LSA) method outperforms all the implemented DFT-based single-channel speech enhancement algorithms.

The technique used for the Blind Multi-speaker speech separation is based on the Degenerate Unmixing and Estimation Technique (DUET) which constructs estimates of the relative mixing

parameters associated with each signal by taking the ratio of time-frequency representations of two mixtures. If the sources in each mixture are W-disjoint orthogonal, that means only one signal is active in the time-frequency plane at a given time-frequency or no signal at all. We have implemented and tested the behavior of the DUET technique on artificial instantaneous speech mixtures of different number of speakers, and they could be separated perfectly according to the performed evaluation tests (objective and subjective).

**Key Words:** Speech Denoising, Single Channel, non-stationary noise, Musical noise, DFT-based Techniques, Blind source separation, W-disjoint orthogonality, Subjective and Objective Tests.

# Table of Contents

# Table of Contents

# Table of Contents

# List of Illustrations

# List of Illustrations

# List of Illustrations

# List of Tables

# Abbreviations

| | |
|---|---|
| ASR | Automatic Speech Recognition |
| BSS | Blind Source Separation |
| DD | Decision Directed |
| DFT | Discrete Fourier Transform |
| DSP | Digital Signal Processing |
| DUET | Degenerate Unmixing Estimations Technique |
| FFT | Fast Fourier Transform |
| IDFT | Inverse Discrete Fourier Transform |
| IFFT | Inverse Fast Fourier Transform |
| LSA | Log-Spectral Amplitude LSA |
| MBSS | Multi-Band Spectral Subtraction |
| ML | Maximum-likelihood |
| MMSE | Minimum Mean-Squared Error |
| OLA | Overlap and add |
| OM-LSA | Optimally-Modified Log-Spectral Amplitude |
| PC | Personal Computer |
| PDF | Probability Density Function |
| PSD | Power Spectrum Density |
| RASR | Robust Automatic Speech Recognition |
| SAP | Speech Absence Probability |
| SAR | Source to Artifacts Ratio |
| SDR | Source to Distortion Ratio |
| SIR | Source to Interference Ration |
| SNR | Signal to Noise Ratio |
| SPL | Sound Pressure Level |
| SPU | Speech Presence Uncertainty |
| STFT | Short Time Fourier Transform |
| STSA | Short-Time Spectral Amplitude |
| TF | Time Frequency |
| DO | Disjoint Orthogonality |

# INTRODUCTION

A major part of the interaction between humans takes place via speech communication which is continuously gaining popularity as technology evolves. The field of speech processing is essentially an application of signal processing techniques to acoustic signals using the knowledge offered by researchers in the field of hearing sciences. Pre-processing of speech signals is considered as crucial step in the development of a various robust and efficient applications such as speech recognition systems that have to perform reliably in noisy environments. Recent advances in computer technology and the simultaneous advances in the digital signal processing theory, phonetics, acoustics, and artificial intelligence had a very profound effect on speech research.

Development and widespread deployment of digital communication systems during the last twenty years have brought increased attention to the role of speech enhancement in speech processing problems.

The degradation of the quality and intelligibility of speech signals, due to the presence of background noise severely affects the ability of speech related systems to perform well. Speech enhancement algorithms are used to improve the performance of communication systems when their input or output signals are corrupted by noise. The main objective of speech enhancement or noise reduction is to improve the perceptual aspects of speech, such as the speech quality and intelligibility. However, the problem of cleaning noisy speech still poses a challenge to the area of signal processing. Noise reduction techniques have some problems and questions. One of these problems is to reach a compromise between noise reduction, signal distortion, and the residual musical noise. Complexity and ease of implementation of the speech enhancement algorithms is also of concern in applications especially those related to portable devices such as mobile communications and digital hearing aids.

The DFT-based speech enhancement methods have been one of the most well-known techniques for noise reduction. The spectral subtraction estimates the power spectrum of clean speech by explicitly subtracting the noise power spectrum from the noisy speech power spectrum. Due to its minimal complexity and relative ease in implementation, it has enjoyed a great deal of attention over the past years. This approach generally produces a residual noise commonly called musical noise. In this project, we propose six DFT-based single-channel speech enhancement algorithms as speech signal pre-processing approaches that will be discussed in detail in chapter two.

# INTRODUCTION

Noise spectrum estimation is so important aspect for effective single-channel speech enhancement algorithms, especially when the background noise is non-stationary, in this case the spectrum will be varying rapidly over time. In our project, we use a noise estimation algorithm with rapid adaptation for highly non-stationary environments proposed in [1] during the implementation of the six DFT-based single-channel speech enhancement algorithms.
In order to make automatic speech recognition systems more effective in real world events it is necessary to deal with difficult environments with multiple speech sources. One classical example is when a number of people are talking simultaneously in one place and the ASR task is to recognize the speech content of one or more target speakers among other interfering sources.

Blind Multi-speaker speech separation refers to the problem of recovering two or more speaker's speech from a number of unknown mixtures. When the number of speakers is greater than the number of mixtures, the problem is degenerate where the traditional matrix inversion demixing cannot be applied. In this project we investigate the applicability of a version of BSS algorithms which is the degenerate unmixing estimations technique (DUET) to Multi-speaker speech signals separation relied on the assumption that the sources were approximately W-disjoint orthogonal. Motivated by the maximum likelihood mixing parameter estimators, we define a power weighted two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures that is shown to have one peak for each source with peak location corresponding to the relative attenuation and delay mixing parameters. The histogram is used to create time-frequency masks that partition one of the mixtures into the original sources.

# CHAPTER 1
# Literature Review

Improving the performance of speech communication systems in noisy environments has been a challenging task for researchers for more than three decades till now. In real world situations, it is very difficult to reliably predict the characteristics of the interfering noise signal or the exact characteristics of the speech waveform. Hence, in effect, the speech enhancement methods are sub-optimal and can only reduce the amount of noise in the signal to some extent. There is a trade-off between distortions in the processed speech and the amount of noise reduced. The effectiveness of the speech enhancement system can therefore be measured based on how well it performs in light of this trade-off.

### 1.1 Speech processing

The term speech processing basically refers to the scientific discipline concerning the analysis and processing of speech signals in order to achieve the best benefit in various practical scenarios. At present, the field of speech processing is undergoing a rapid growth in terms of both performance and applications. This is stimulated by the advances being made in the field of microelectronics, computation and algorithm design. Nevertheless, speech processing still covers an extremely broad area, which relates to the following three engineering applications [2]:

- Speech Coding and transmission that is mainly concerned with man-to man voice Communications.
- Speech Synthesis which deals with machine-to-man communications.
- Speech Recognition relating to man-to machine communications.

### 1.2 Speech signal characteristics

The speech signal is a time varying signal whose signal characteristics represent the different speech sounds produced. There are three ways of labeling events in speech. First is the silence state in which no speech is produced. Second state is the unvoiced state in which the vocal cords are not vibrating, thus the output speech waveform is aperiodic and random in nature. The last state is the voiced state in which the vocal cords are vibrating periodically when air is expelled from the lungs. This results in the output speech being quasi-periodic shows a speech waveform with unvoiced and voiced state [2].

Speech is produced as a sequence of sounds. The type of sound produced depends on shape of the vocal tract. The vocal tract starts from the opening of the vocal cords to the end of the lips. Its cross sectional area depends on the position of the tongue, lips, jaw and velum. Therefore the tongue, lips, jaw and velum play an important part in the production of speech.

Physically, the sounds of speech can be described in terms of a pitch contour and formant frequencies. Formants are resonant frequencies of the vocal tract which appear in the speech spectrum as clear peaks [3]. As an example, three distinct formant peaks can be seen in the frequency domain plot of a short speech recording, in Figure 1.1.

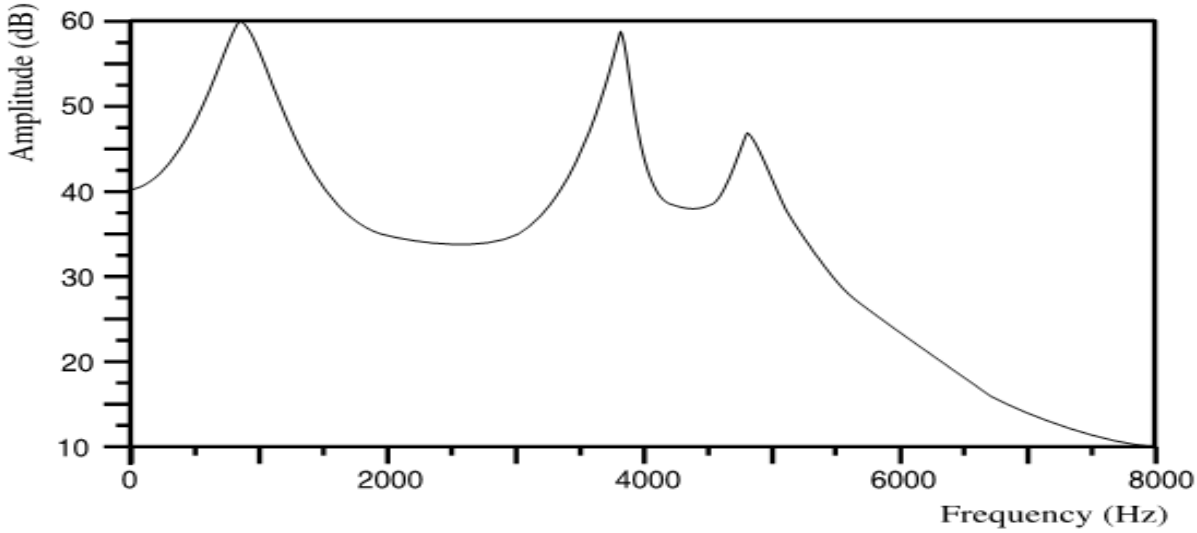


**Figure 1.1:** Spectrum plot of a 20ms recording of voiced speech, showing three distinct formant peaks [4].

Formants have been described by the famous researcher Klatt and others as being the single most important feature in speech communications [5]. Generally many formants will be present in a typical utterance, and the location of these will vary over time as the shape of the mouth changes. Formants are counted from the lowest frequency upwards, and usually only the first three (F1, F2 and F3) contribute significantly to the intelligibility of speech. Some fricative sounds like /ch/ can produce a lot of formants, but generally speaking F1 contains most of the speech energy while F2 and F3 between them contribute more to speech intelligibility [6].

The pitch contour (often called f0) is the parameter that describes the tone of the voice (the perceived frequency), and is in effect the fundamental vocal frequency. Again, pitch frequencies contain energy but contribute little to intelligibility for English and other European languages [7]

### 1.3 Hearing and perception

Audible sounds are transmitted to the human ears through the vibration of the particles in the air. Human ears consist of three parts, the outer ear, the middle ear and the inner ear. The function of

the outer ear is to direct speech pressure variations toward the eardrum where the middle ear converts the pressure variations into mechanical motion. The mechanical motion is then transmitted to the inner ear, which transforms these motion into electrical potentials that passes through the auditory nerve, cortex and then to the brain. Figure 1.2 shows the schematic diagram of the human ear [2].

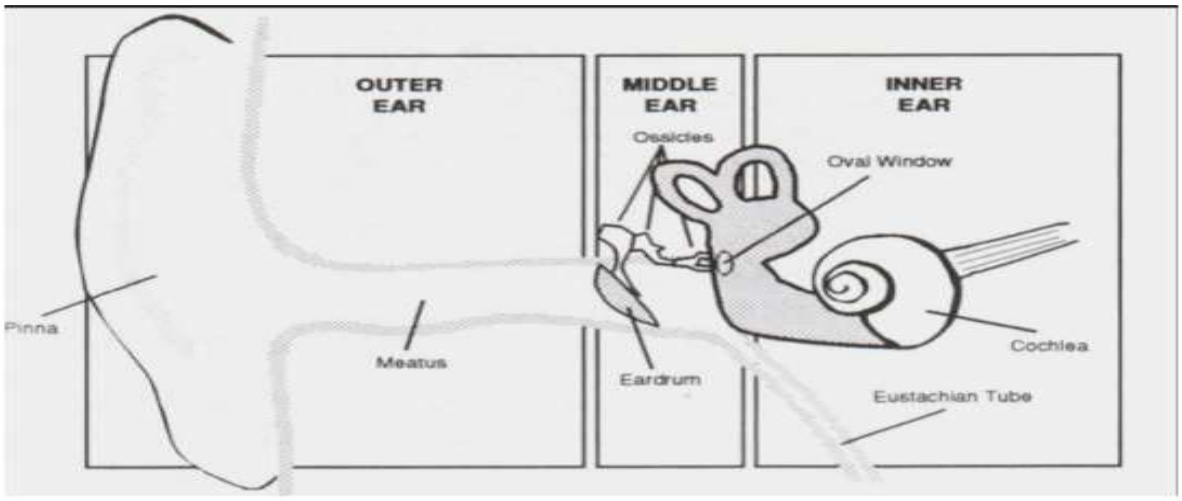


**Figure 1.2 :** The schematic diagram of the human ear [2]

## 1.4 Sound Basics

### 1.4.1 Concept of sound

A person perceives sound as any vibration of the eardrum in the audible frequency range that results from an incremental variation in air pressure at the ear. A variation in pressure above and below the atmospheric pressure is called sound pressure and is measured in units of Pascal (Pa) [8, 9].

The number of pressure variations per second is called the frequency of sound, which is measured in cycles per second, called Hertz (Hz). A young person with normal hearing can perceive sound in the frequency range of roughly 20 − 20,000 Hz, defined as the normal audible frequency range. A sound that has only one frequency is known as a pure tone [8, 9].

The speed of sound is the rate at which a sound wave propagates through a given medium, and is dependent on the elasticity and density of that medium. For all practical purposes, the speed of sound in air is dependent only on the absolute temperature, which directly affects its density. The equation for the speed of sound in air is $c = 20.05\sqrt{T}\ (m/s)$ where T is the absolute temperature

of air in degrees Kelvin. At room temperature and standard atmospheric pressure, the speed of sound in air is 343 m/s [8, 9].

### 1.4.2 The Sound levels and Decibel

Generally, Sound levels are described logarithmically because it compresses the large range of typical sound pressures into a smaller and more practical scale, which incidentally also more closely parallels the human ear's ability to judge the relative loudness of sounds according to the ratio of their pressure. The important thing to remember about the decibel is that it represents a relative measurement or ratio. 'Sound power level' and 'sound pressure level' are typically expressed in terms of decibels, as an indication that they are not absolute values, but rather, measurements relative to a reference quantity.

The quantity most often used to measure the "strength" of a sound wave is the sound pressure level (SPL) measured with respect to a standard reference pressure of $p_{ref} = 5.10^{-5}$ Pa [10]. The sound pressure level can be calculated using the following formula:

$$L_p = 20log_{10}(p/p_{ref}) \qquad (dB) \qquad (1.1)$$

The reference pressure represents the normal threshold of hearing for most individuals.

### 1.4.3 Discrete time representation of sound

When making a digital recording of a sound it is sampled with a certain frequency. Each sample stores information about the sound pressure level measured at that specific time. These values are taken at regular time periods, known as the sampling period, Ts.

For example, consider a continuous waveform given by $y(t)$. In order to process this waveform digitally we first must convert this into a discrete time vector. Each value in the vector represents the instantaneous value of this waveform at integer multiples of the sampling period. The values of the sequence, $y(t)$ corresponding to the value at n times the sampling period which is denoted as y[n].

$$y[n] = y(nT_s) \qquad (1.2)$$

To be able to reconstruct the sound correctly, the number of samples per second, called the sampling frequency ($f_s = {}^{1}/_{T_s}$), must be at least twice the highest frequency of the sound. This relation was discovered by Nyquist [11].

### 1.5 Noise Characteristics

Noise can be defined as an unwanted signal that interferes with the communication or measurement of another signal. A noise itself is a signal that conveys information regarding the source of the noise. For example, the noise from a car engine conveys information regarding the state of the engine. The sources of noise are many, and vary from audio frequency acoustic noise emanating from moving, vibrating or colliding sources such as revolving machines, moving vehicles, computer fans, keyboard clicks, wind, rain,…etc [12].

Noise and distortion are the main limiting factors in communication and measurement systems. Therefore the modeling and removal of the effects of noise and distortion have been at the core of the theory and practice of communications and signal processing.

The nature of the noise is an important factor in deciding on a speech enhancement method. Therefore, a good model of noise is important for the performance of speech enhancement system and vice-versa it is important to analyze how well a speech enhancement algorithm/model works with different types of noise.

Noise can be different based on various statistical, spectral or spatial properties. From [13] different noises can be summarized as in Table 1.1.

| **Property** | **Types** |
|---|---|
| Structure | Continuous/ Impulsive/ Periodic |
| Type of Interaction | Additive/ Multiplicative/ Convolutive |
| Temporal behavior | Stationary/ Non-stationary |
| Frequency range | Broadband/ Narrowband |
| Signal dependency | Correlated/ Uncorrelated |
| Statistical properties | Dependent/ Independent |
| Spatial properties | Coherent/ Incoherent |

**Table 1.1:** Classification of noise based on various properties

Based on the nature and properties of the noise sources, noise can be classified in the following ways [14]:

1. Background noise: additive noise, which is usually uncorrelated with the signal and present in various environment scenarios like trains, restaurants, streets, machines, climatic conditions, factory environment…etc.
2. Interfering speakers (speech like noise): additive noise composed of single or multiple "competing" speakers. This noise has characteristics and frequency range very similar to the speech signal of interest.
3. Impulse noise: slamming of doors, noise present in archived gramophone recordings.
4. Non-additive noise: noise due to non-linearities of microphones, speakers and channel distortion (speech on transmission lines).
5. Non-additive noise due to speaker stress: e.g. Lombard effect i.e. the effect induced in presence of noise, wherein the speaker has a tendency to increase his vocal effort [15]. This results in speech having different spectral properties as compared to clean speech.
6. Noise correlated with the signal: reverberations and echoes.
7. Convolutive noise: corresponds to convolution in time domain. For instance, changes in speech signal due to changes in room acoustics or changes in microphones etc. These are usually harder to deal with, as compared to additive noise.
8. Multiplicative noise: signal distortion due to fading in cellular channels.

In general, it is more difficult to deal with non-stationary noise, where there is no priori knowledge available about the characteristics of noise. Since non-stationary noise is time varying, the conventional method of estimating the noise from initial intervals assuming no speech signal is not suitable for estimation [13]. Noise types, which are similar in temporal, frequency or spatial characteristics to speech, are also difficult to remove or attenuate. Multitalker babble, for instance, retains some characteristics of speech and poses a particularly difficult problem for an algorithm intended to isolate speech signal from the background noise.

### 1.6 Classification of speech enhancement techniques

In many speech related systems, the desired signal is not available directly; rather it is mostly contaminated with some interference sources. These background noise signals degrade the quality and the intelligibility of the original speech resulting in a severe drop in the performance of the applications. The degradation of the speech signal due to the background noise is a severe

problem in speech related systems and therefore should be eliminated through speech enhancement algorithms.

Speech enhancement systems can be classified in a number of ways [13, 16] based on the criteria used or application of the enhancement system. (See Table 1.2).

| Domain | Possible Strategies |
|---|---|
| Number of input channels | One/ Two / Multiple |
| Domain of processing | Time / Frequency |
| Type of algorithm | Non-adaptive/ Adaptive |
| Additional constraints | Speech production / Perception |

**Table 1.2:** Speech enhancement processing strategies [13].

The speech signal can be acquired from single or multiple channel sensors. Additive noise can make speech enhancement particularly difficult. Non-stationarity of the noise process can further complicate the enhancement effort. One microphone input (single channel) could make speech enhancement difficult, as speech and noise are present in the same channel. Separation of the two signals would require relatively good knowledge of the speech and noise models or require that the interfering signal be present exclusively in a different frequency band than that of the speech signal. The noise source is assumed to be statistically independent and additive. This assumption is based on the fact that most environmental noise is typically additive in nature.

### 1.7 Applications of speech enhancement

In the current information technology, there are many areas that speech enhancement is used in order to improve the performance of the system [17];

- Robust Automatic Speech Recognition (RASR): the accuracy of automation speech recognition degrades in the presence of back ground noise or other interfering sources. Noise reduction of speech signals has therefore critical importance as pre-process as such type of systems, including human-computer interactions, robotics, and audio driven systems, etc.
- Telecommunications: Background noise is a common problem which degrades the quality of the communication for the human listener. Speech enhancement may be applied to such systems in order to remove the unwanted noise sources. Another problem in

telecommunication is the channel noise. By enhancing the speech signal before it goes into the channel, it is also possible to reduce the effect of the channel noise.

- Digital Hearing Aids: the digital hearing aid users often complain of difficulty in understanding speech in the presence of background noise. Therefore, speech enhancement is an important process to improve speech perception in a noisy environment for the human listener.

Figure 1.3 shows an illustration of the usage of speech enhancement. It can be observed that enhancement may also be applied directly to the clean speech signal in order to reduce the effect of the channel noise in the communication system [17].

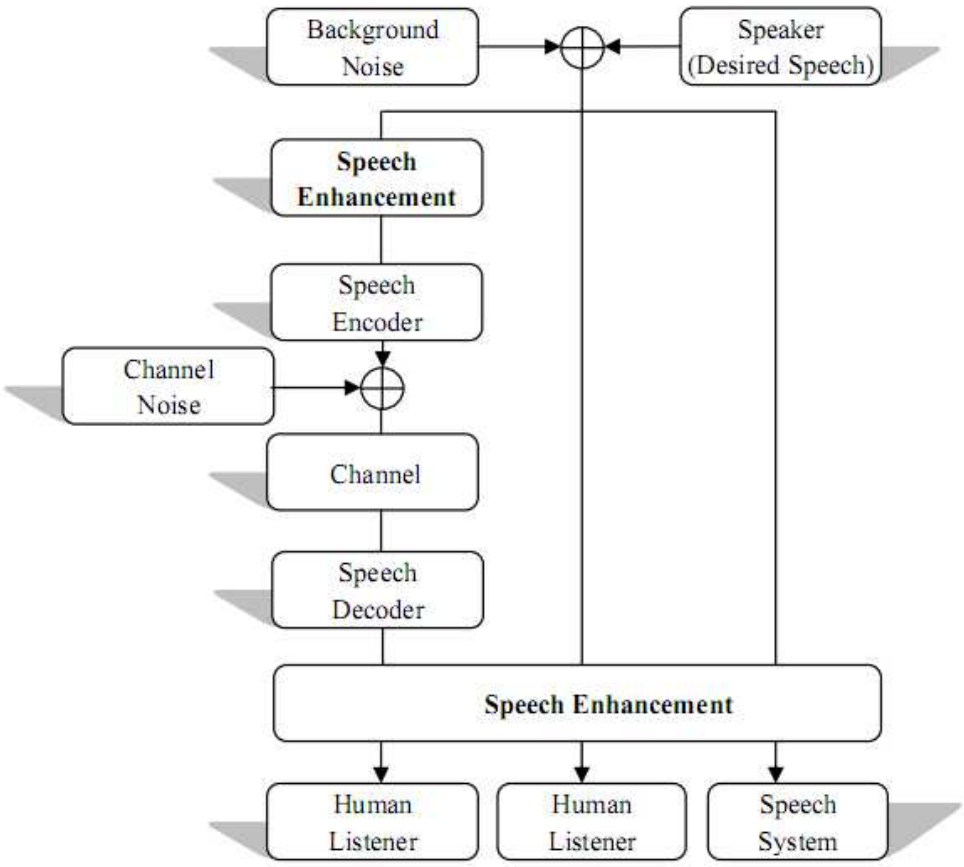


**Figure 1.3:** The application areas of speech enhancement [17].

# CHAPTER 2

# DFT-based techniques for single channel speech Enhancement

This chapter describes short time DFT-based single channel techniques for additive noise removal.

These methods are based on the analysis-modify-synthesis approach. They use fixed analysis window length (usually 20-32ms) and frame by frame based processing.

They are based on the fact that human speech perception is not sensitive to spectral phase but the clean spectral amplitude must be properly extracted from the noisy speech to have acceptable quality of speech at output and hence they are called short time spectral amplitude (STSA) based methods. Figure 2.1 shows the basic overview of a single-channel speech enhancement system.

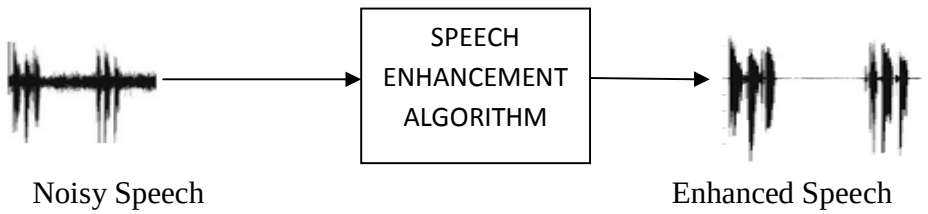


**Figure 2.1:** Basic overview of single channel speech enhancement system.

STSA based approaches assume that noise is additive and uncorrelated to the speech signal.

Most real world noise such as street noise, train station noise, restaurant noise, babble noise…etc are non-stationary in nature. Therefore complete noise cancellation is more complex as it is not possible to completely track such noises. However, using these assumptions, it is possible to achieve significant reduction in the background noise levels using simple techniques.

**2.1 Additive noise model**

Most of the speech enhancement techniques use the additive noise model to model the background noise. In the additive noise model the noisy speech is assumed to be the sum of the clean speech and the noise as defined in the following equation:

$$y(t) = x(t) + d(t) \quad (2.1)$$

Where y(t) is the noisy speech signal, x(t) is the clean speech signal, and d(t) is the background noise signal.

Let $y[n] = x[n] + d[n]$ be the sampled observed noisy speech signal consisting of the clean signal x[n] and the noise signal d[n] where, $0 \leq n \leq N - 1$, and N is the frame length. The additive noise model can be represented as shown in Figure 2.2.

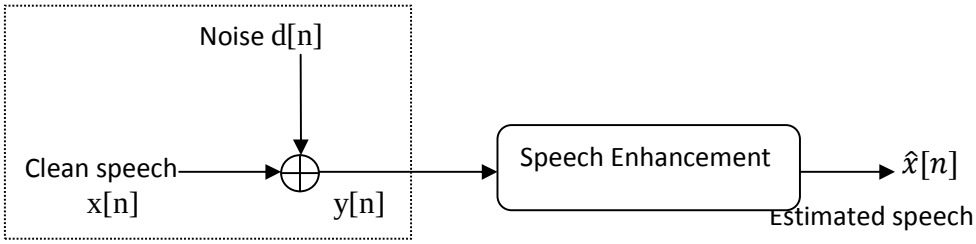


**Figure 2.2:** Additive noise model in single-channel speech enhancement [18].

In the DFT frequency domain, the additive relation can be expressed as follows:

$$Y_k = X_k + D_k \quad (2.2)$$

Where $X_k$, $Y_k$, $D_k$ denotes the $k$th spectral component of the original clean signal x[n], the noisy observation y[n], and the noise d[n] respectively, in the analysis frame interval [0, N − 1].

## 2.2 The general structure of the DFT-based speech enhancement

The overall structure of the DFT-based speech enhancement techniques is shown in Figure 2.3. The analog speech signal is sampled at the frequency Fs, and quantized to 16 bits. The digitized speech is then partitioned into overlapping frames. The commonly used amount of overlap is 50% or 75%. An appropriate window is then applied to each frame. The windowed frame then passes through the DFT transform stage. The output complex coefficients are then separated into magnitude and phase. The phase is kept unaltered and the magnitude is processed using one of the STSA speech enhancement techniques. After getting the estimated spectral magnitude component of the clean signal, then it is combined with the phase, and the inverse transform operation is carried out. Finally, the overlap and add (OLA) technique [19] is applied to reconstruct the output speech signal (the enhanced speech signal) [20].

## 2.3 Time to Frequency Domain Conversion

The statistical properties of a speech signal change over time, specifically, from one phoneme to the next. Within phonemes, which average about 80 ms in duration [21], the statistics of the signal are relatively constant. For this reason, the processing of speech signals is typically done in short time sections called frames. The size of frames is typically from 20 to 32 ms. In these short-time segments, speech can be considered stationary [22].

The frames of time domain data are windowed (the effects of the window employed are discussed in the next section 2.4) and then converted to frequency domain using the Discrete Fourier Transform (DFT).

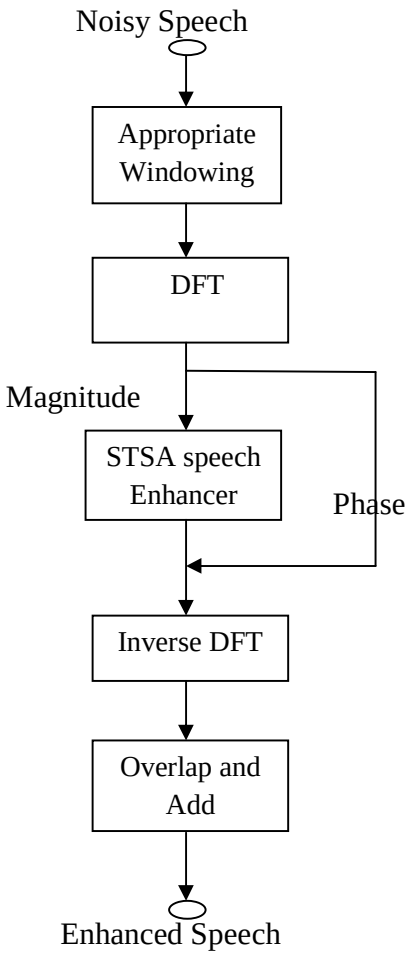


**Figure 2.3:** Block diagram of the DFT-based speech enhancement [20]

For a finite-duration, discrete-time signal $y(n)$ of length $N$ samples, the discrete Fourier transform (DFT) is defined as $N$ uniformly spaced spectral samples (see figure 2.4) given by :

$$Y_k = \sum_{n=0}^{N-1} y[n]\exp(-j\frac{2\pi}{N}kn) \qquad k = 0,1,2 \ldots N-1 \tag{2.3}$$

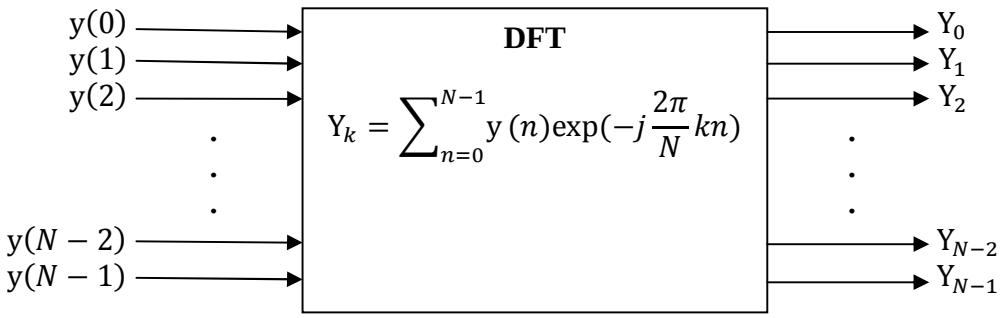


**Figure 2.4:** Illustration of the DFT as a parallel-input, parallel-output processor [12].

The inverse discrete Fourier transform (IDFT) is given by

$$y(n) = \frac{1}{N}\sum_{k=0}^{N-1} Y_k \exp(j\frac{2\pi}{N}kn) \qquad n = 0,1,2 \dots N-1 \tag{2.4}$$

The direct calculation of the Fourier transform requires $N(N-1)$ multiplications and a similar number of additions. Algorithms that reduce the computational complexity of the discrete Fourier transform are known as fast Fourier transforms (FFT) methods. FFT methods utilize the periodic and symmetric properties of $\exp\left(-j\frac{2\pi}{N}\right)$ to avoid redundant calculations [12].

**2.4 Windowing**

The speech signal needed to be split up into overlapping frames of size N. Overlapping of the frames is at 50% in this project. This is done to avoid discontinuities between frames. Figure 2.5 shows how the signal gets split up into frames [23].

As long as the length of frames is kept appropriate as mentioned in the previous section. It should be noted that the effective frequency resolution depends only on the frame size.

It is desirable that the window function satisfy two characteristics in order to reduce the spectral distortion caused by the windowing. One is a high-frequency resolution, principally, a narrow and sharp main lobe. The other is a small spectral leak from other spectral elements produced by the convolution, in other words, a large attenuation of the side lobe [24].

Since these two requirements are actually contrary to each other, and because it is impossible to satisfy both, several compromise window functions have been proposed. Among these, the Hamming window $W_H(n)$, defined as:

$$W_H(n) = 0.54 - 0.46\cos(\frac{2n\pi}{N-1}) \tag{2.5}$$

is usually used as the window function for speech analysis. The Hamming window is advantageous in that its resolution in the frequency domain is relatively high and its spectral leak is small since the attenuation of the side lobe is more than 43 dB [24].

On the other hand, a rectangular window which is defined as:

$$W_R(n) = 1 \qquad for\ (0 \le n \le N-1) \tag{2.6}$$

which corresponds to the simple extraction of N sample points of the speech wave, and it has the largest frequency resolution, whereas the attenuation of its first side lobe is only 13 dB. Thus the rectangular window is not suited to the analysis of a speech wave having a large dynamic range of spectral components.

Another window, called the Hanning window is given by:

$$W_N(n) = 0.5 - 0.5\cos(\frac{2n\pi}{N-1}) \quad (2.7)$$

is also employed. Although the advantage of this window is that its higher-order side lobes are lower than those of the Hamming window, the attenuation of the first side lobe is only roughly 30 dB. The shapes of these windows and the spectra for 10 periods of 1 $KHz$ sinusoidal waves extracted by using these windows are shown in Figure 2.6.

The relationship between the sampling period $T[s]$, number of samples for analysis $N$, and nominal frequency resolution of the calculated spectrum $\Delta f[Hz]$ is expressed as

$$\Delta f = \frac{1}{T.N} \quad (2.8)$$

In this project the analysis window of choice is the Hamming window. Based on the argument presented in Section 2.3, an appropriate window length is chosen [24].

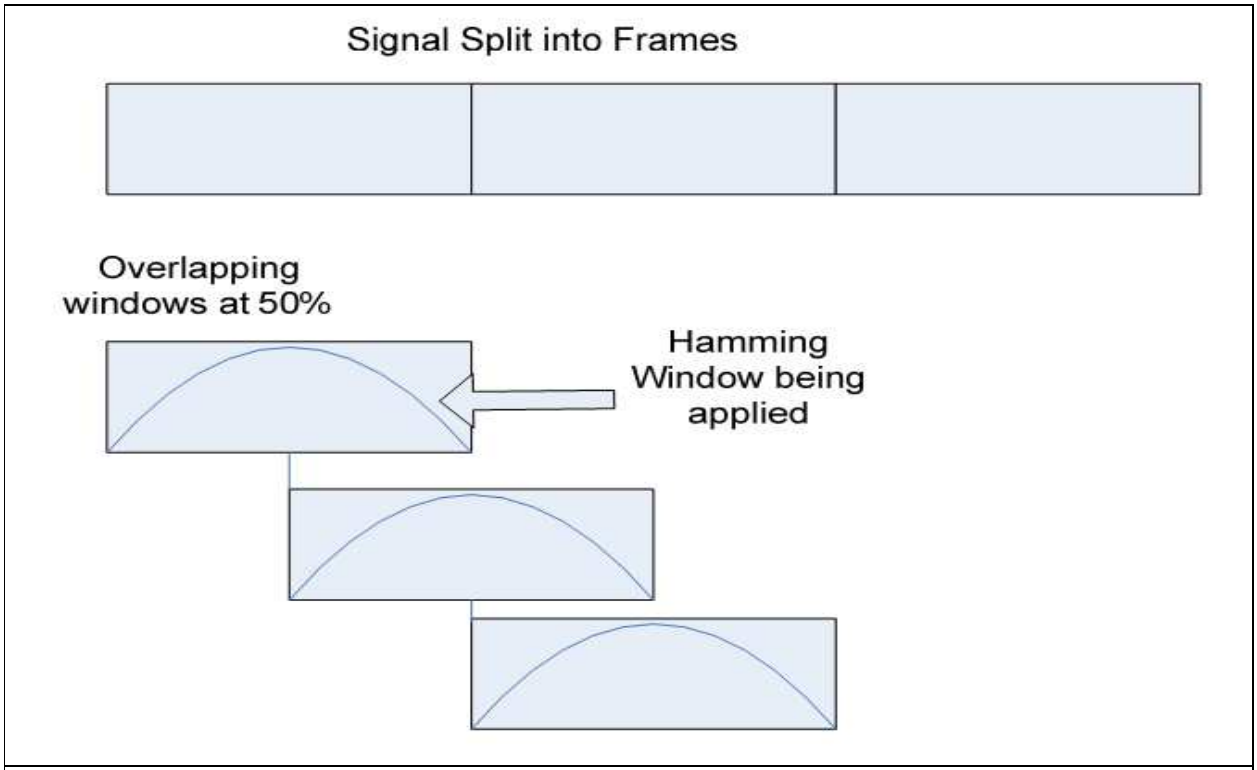


**Figure 2.5:** Frames getting split up and Hamming window applied with overlap [8]

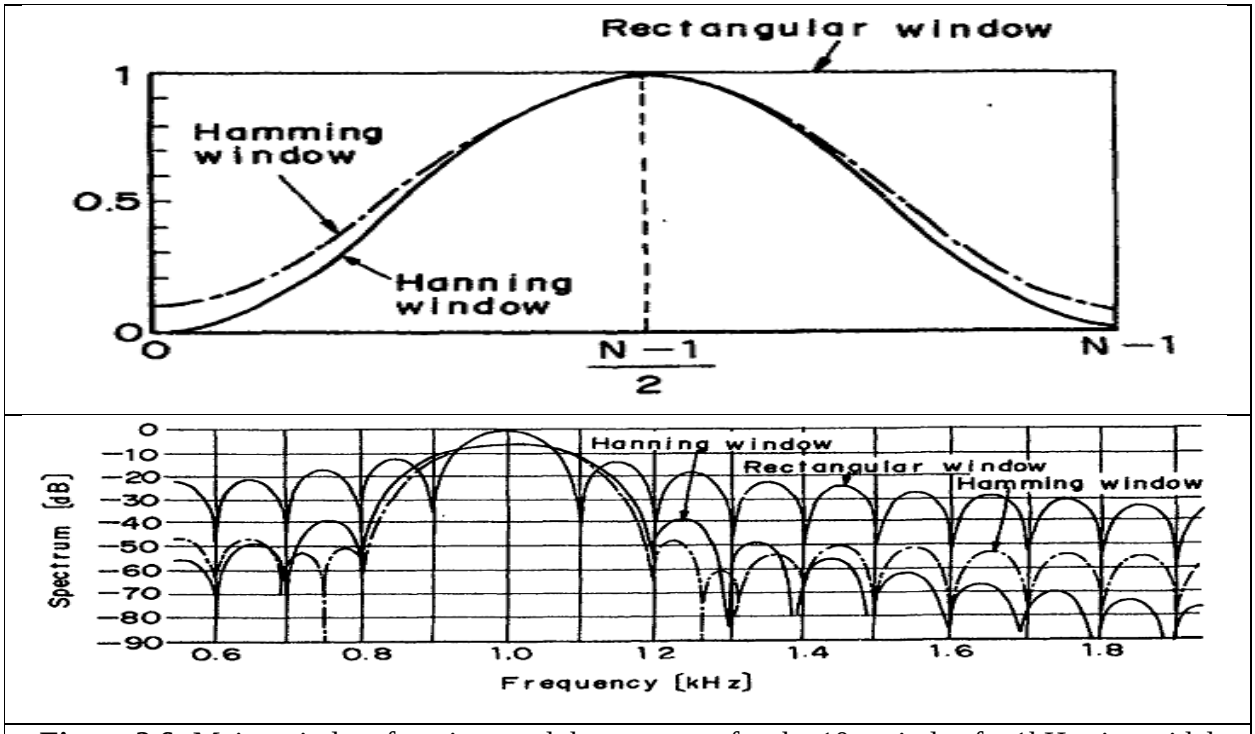


**Figure 2.6:** Major window functions and the spectrum for the 10 periods of a 1kHz sinusoidal wave extracted using each of the windows [7].

## 2.5 Noise Power Spectrum Estimation

Noise spectrum estimation is a challenging task for single-channel speech enhancement, where we have only the noisy speech available at the input. Non-stationary noise spectrum varies rapidly over time, hence it needs to be estimated and updated continuously.

In this project we use the algorithm proposed in [1] for estimating highly non-stationary noise environments.

Let y(n) = x(n) + d(n), where y(n) is the noisy speech signal, x(n) is the clean signal and d(n) is the additive noise. The smoothed power spectrum of the noisy speech signal can be estimated using a first-order recursive formula as follows:

$$p_k(l) = \eta . p_k(l-1) + (1-\eta)|Y_k(l)|^2 \quad (2.9)$$

Where $|Y_k(l)|^2$ is the estimate of the short-time power spectrum of y(n), $\eta$ is a smoothing constant, $l$ is the frame index, and $k$ is the frequency bin index.

We know that in the case of speech-absent frame, the noise power spectrum is the same as the power spectrum of the noisy speech. Hence, we can update the estimate of the noise power spectrum by tracking the noise-only frames (speech absent). In order to realize that, we calculate the ratio of the energy of the noisy speech power spectrum in three different frequency bands

(low: 0-1 KHz, middle: 1-3 KHz, high: 3 KHz and above) to the energy of the corresponding frequency band in the previous noise estimate. The three ratios are calculated as follows:

$$R_{low}(l) = \frac{\sum_{k=1}^{LF} p_k(l)}{\sum_{k=1}^{LF} N_k(l-1)}, \qquad R_{Middle}(l) = \frac{\sum_{k=LF+1}^{MF} p_k(l)}{\sum_{k=LF+1}^{MF} N_k(l-1)},$$

$$R_{High}(l) = \frac{\sum_{k=MF+1}^{Fs/2} p_k(l)}{\sum_{k=MF+1}^{Fs/2} N_k(l-1)} \tag{2.10}$$

Where $N_k(l)$ is the estimate of the noise power spectrum at frame $l$, and $LF, MF, Fs$ correspond to the frequency bins of 1 kHz, 3 kHz and the sampling frequency respectively. If the above three ratios $(R_{low}(l), R_{Middle}(l), R_{High}(l))$ are all smaller than a threshold $\sigma$, then it is concluded that it is a speech-absent frame and the noise estimate is updated according to:

$$N_k(l) = \lambda . N_k(l-1) + (1-\lambda)\,|Y_k(l)|^2 \tag{2.11}$$

Where λ is a smoothing constant. If any or all of the above three ratios are larger than the threshold $\sigma$, then another algorithm is going to be used for updating and estimating the noise spectrum.

The proposed algorithm used for speech-present segments is based on first finding the minimum of the noisy speech spectrum, and using that minimum to determine signal presence probability in sub-bands. The signal presence probability is used to determine a frequency-dependent smoothing parameter which replaces the fixed smoothing constant λ given in eq (2.11). The local minimum of the noisy speech is computed by averaging the past spectral values with a look-ahead factor as defined in [25]:

If $P_{k_{min}}(l-1) < P_k(l),\ then$

$$P_{k_{min}}(l) = \gamma . P_{k_{min}}(l-1) + \frac{1-\gamma}{1-\beta}\left(P_k(l) - \beta P_k(l-1)\right) \tag{2.12}$$

Else $P_{k_{min}}(l) = P_k(l)$

Where $P_{k_{min}}(l)$ denotes the local minimum of the noisy speech power spectrum, β,and γ are constants determined experimentally.

The approach taken to determine signal presence probability in sub-bands is similar to that proposed in [26].

Let $S_k(l) \triangleq P_k(l)/P_{k_{min}}(l)$ denotes the ratio between the energy of the noisy speech to its local minimum. This ratio is compared against a frequency-dependent threshold and if it is found to be larger than that threshold, then the corresponding frequency is considered to contain speech.

Using the above ratio $S_k(l)$, the new frequency-dependent smoothing constant can be estimated as follows:

$$\alpha_k(l) = \begin{cases} \alpha_1 \; if \; S_k(l) < \delta_k \\ \alpha_2 \;\; otherwise \end{cases} \tag{2.13}$$

Where $\alpha_1$, $\alpha_2$ are smoothing constants ($\alpha_2$>$\alpha_1$) and $\delta_k$ is a frequency-dependent threshold given by:

$$\delta_k = \begin{cases} 1.3 & 1 \le k \le LF \\ 3 & LF < k \le MF \\ 5 & MF < k \le Fs/2 \end{cases} \tag{2.14}$$

Finally, after computing the frequency-depending smoothing factor $\alpha_k(l)$, the noise spectrum estimate is updated according to:

$$N_k(l) = \alpha_k(l).\, N_k(l-1) + (1 - \alpha_k(l))|\mathrm{Y}_k(l)|^2 \tag{2.15}$$

To summarize, if the ratios defined in Eq.(2.10) indicate that the current frame is a speech-absent frame, then Eq. (2.11) is used to update the noise spectrum. Otherwise, Eq.(2.15) is used to update the noise spectrum [1].

**2.6 Spectral Subtraction**

Spectral subtraction is a method for restoration of the power spectrum or the magnitude spectrum of a signal observed in additive noise, through subtraction of an estimate of the noise spectrum from the noisy signal spectrum [12]. Figure 2.7 shows the simplified structure of basic spectral subtraction systems.

The first detailed treatment of spectral subtraction was performed by Boll [27, 28]. After that, papers [29, 16] expanded and generalized Boll's method to power subtraction, Wiener filtering.

Both spectral subtraction and Wiener filtering algorithms are derived under Gaussian assumption for each spectral component.

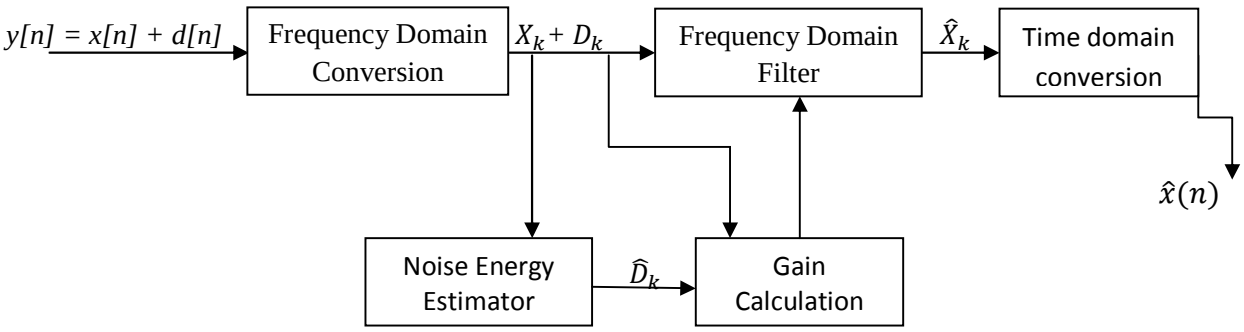


**Figure 2.7:** Basic structure of spectral subtraction systems [30]

### 2.6.1 Power Spectral Subtraction and its generalized form

The basic power spectral subtraction (PSS) principle involves the subtraction of the estimated noise variance, from the power spectrum of the observed noisy signal, to obtain an estimate of the modulus of speech power spectrum (taking into account that $\left|\hat{X}_k\right|^2 has\ to\ be\ positive$). Mathematically this is represented as:

$$\left|\hat{X}_k\right|^2 = \max\left(|Y_k|^2 - \left|\hat{D}_k\right|^2, 0\right) \tag{2.16}$$

However, there are limitations to this subtraction rule. The basic problem has been tackled by deriving several fundamentally and theoretically justified noise suppression rules.

Since the power spectrum of two uncorrelated signals is additive. By generalizing the exponent from 2 to *a*, Eq. (2.16) becomes

$$\left|\hat{X}_k\right|^a = \max\left(|Y_k|^a - \left|\hat{D}_k\right|^a, 0\right) \tag{2.17}$$

The speech phase is estimated directly from the noisy signal phase. Thus a general form of the estimated speech in frequency domain can be written as:

$$\hat{X}_k = \left(\max\left(|Y_k|^a - \alpha\left|\hat{D}_k\right|^a, 0\right)\right)^{\frac{1}{a}}.\exp\left(j\theta_k\right), \tag{2.18}$$

Where $\alpha > 1$ is used to overestimate the noise to account for the variance in the noise estimate. The inner term $|Y_k|^a - \alpha\left|\hat{D}_k\right|^a$ is limitied to positive values, since it is possible for the overestimated noise to be greater than the current signal [17].

### 2.6.2 Spectral Subtraction using over-subtraction and spectral floor

For more residual musical noise reduction, a modification of the spectral subtraction was proposed by Berouti et al [31]. The technique could be expressed as:

$$\left|\hat{X}_k\right|^2 = \max\left(|Y_k|^2 - \alpha.\left|\hat{D}_k\right|^2, \beta.\left|\hat{D}_k\right|^2\right) \tag{2.19}$$

Where: $\alpha$ is the over-subtraction factor, and it is given in terms of the frame noisy signal to noise ratio as follows:

$$\alpha = \alpha_0 - \frac{3}{20}.SNR \qquad -5dB \le SNR \le +20dB \tag{2.20}$$

$\alpha_0$ is the desired value of $\alpha$ at 0 dB SNR.

$\alpha$ plays the role of a time-varying factor, which provides a degree of control over the noise removal process between periods of noise update.

The parameter $\beta$ is the spectral floor which prevents the spectral components of the enhanced spectrum from being below the smallest value $\beta.\left|\widehat{D}_k\right|^2$. In this case $\beta$ plays the role of controller (the amount of remaining residual noise and the amount of perceived musical noise).

### 2.6.3 Multi-Band Spectral Subtraction (MBSS)

The idea of MBSS method proposed by [32] starts from the fact that the colored noise has different effects at the various frequencies of the speech spectrum. The MBSS technique performs spectral subtraction with different over subtraction factor in different non-overlapped frequency bands. The spectral subtraction rule in $i^{th}$ frequency band is given by:

$$\left|\hat{X}_{k,i}\right|^2 = \begin{cases} \left|\overline{Y_{k,i}}\right|^2 - \delta_i \alpha_i . \left|\widehat{D}_{k,i}\right|^2, & if \ \left|\overline{Y_{k,i}}\right|^2 > \delta_i \alpha_i . \left|\widehat{D}_{k,i}\right|^2 \\ \beta . \left|\overline{Y_{k,i}}\right|^2 & else \end{cases} \qquad for\ b_i \le k \le e_i \tag{2.21}$$

where the spectral floor parameter was set to $\beta = 0.002$, and $b_i$ and $e_i$ are the beginning and ending frequency bins of the $i^{th}$ frequency band.

$\overline{Y_{k,i}}$ is the $i^{th}$ frequency band of smoothed and averaged version of the noisy speech spectrum. A weighted spectral average is taken over preceding and succeeding frames of speech as follows:

$$\overline{Y_{k,j}} = \sum_{l=-M}^{M} W_l \, Y_{k,j-l} \tag{2.22}$$

Where $j$ is the frame index, and $0 < W_l < 1$. The averaging is done over M preceding and succeeding frames of speech.

The number of frames $M$ is limited to 2 to prevent smearing of the speech spectral content. The weights $W_l$ were empirically determined and set to $W_l = [0.09, 0.25, 0.32, 0.25, 0.09]$ for $-2 \le l \le +2$ [32].

The band-specific over-subtraction factor $\alpha_i$ is a function of the segmental $SNR_i$ of the $i^{th}$ frequency band, which is calculated as:

$$SNR_i(dB) = \left[\frac{\sum_{k=b_i}^{e_i} \left|Y_{k,i}\right|^2}{\sum_{k=b_i}^{e_i} \left|\widehat{D}_{k,i}\right|^2}\right] \tag{2.23}$$

$\alpha_i$ can be expressed in terms of $SNR_i$ (defined previously) as follows:

$$\alpha_i = \begin{cases} 4.75 & SNR_i < -5 \\ 4 - \frac{3}{20} SNR_i & -5 \le SNR_i \le 20 \\ 1 & SNR_i > 20 \end{cases} \tag{2.24}$$

The additional over subtraction factor $\delta_i$ called tweaking factor provides additional degree of control in each frequency band. The values of this factor are empirically determined and set according to following equation (Usually 4-8 linearly spaced frequency bands are used).

$$\delta_i = \begin{cases} 1 & f_i < 1\ KHz \\ 2.5 & 1\ KHz \le f_i \le \dfrac{F_s}{2} - 2\ KHz \\ 1.5 & f_i > \dfrac{F_s}{2} - 2\ KHz \end{cases} \tag{2.25}$$

Where $f_i$ is the upper frequency of the the $i^{th}$ band, and $F_s$ is the sampling frequency [32].

**2.7 Wiener Filter**

In terms of our speech enhancement problem the Wiener filter proposed in [33] is given by:

$$|\hat{X}_k| = \frac{\xi_k}{\xi_k + 1} |Y_k| \tag{2.26}$$

Where $\xi_k\ is$ defined as the a priori $SNR\ found\ by\ Decision\ Directed\ Method.$

**2.8 MMSE of Short-Time Spectral Amplitude**

Ephraim and Malah [34] formulated an optimal spectral amplitude estimator, which, specifically, estimates the modulus (magnitude) of each complex Fourier coefficient of the speech signal in a given analysis frame from the noisy speech in that frame.

In order to derive the MMSE STSA estimator, the a priori probability distribution of the speech and noise Fourier expansion coefficients should be assumed since these are unknown in reality. Ephraim and Malah [34] assumed that the Fourier expansion coefficients of each process can be modeled as statistically independent Gaussian random variables, real and imaginary parts of each component is independent to each other, and the mean of each coefficient is assumed to be zero and the variance time-varying.

**2.8.1 The Gaussian based MMSE-STSA Estimator**

Since the spectral components are assumed to be statistically independent, the MMSE amplitude estimator $\hat{A}_k$ can be derived from $Y_k$ only, Fig 2.8. That is,

$$\hat{A}_k = E\{ A_k \backslash Y_k \}$$
$$= \frac{\int_0^{\infty} \int_0^{2\pi} a_k P(Y_k \backslash a_k, \alpha_k) P(a_k, \alpha_k) d\alpha_k da_k}{\int_0^{\infty} \int_0^{2\pi} P(Y_k \backslash a_k, \alpha_k) P(a_k, \alpha_k) d\alpha_k da_k} \tag{2.27}$$

Under the assumed Gaussian model, $P(Y_k \backslash a_k, \alpha_k)$ and $P(a_k, \alpha_k)$ are given by [27]

$$P(Y_k \backslash a_k, \alpha_k) = \frac{1}{\pi\lambda_d(k)} exp\left\{-\frac{1}{\lambda_d(k)}\left|Y_k - a_k e^{j\alpha_k}\right|^2\right\} \quad (2.28)$$

$$P(a_k, \alpha_k) = \frac{a_k}{\pi\lambda_x(k)} exp\left\{-\frac{{a_k}^2}{\lambda_x(k)}\right\} \quad (2.29)$$

Where $\lambda_x(k) = E\{|X_k|^2\}$, and $\lambda_d(k) = E\{|D_k|^2\}$ }, are variances of the $k^{th}$ spectral component of the speech and noise, respectively. Substituting Eq. (2.28) and Eq. (2.29) into Eq. (2.27) gives the desired gain function for the MMSE-STSA estimator, [34]:

$$G_{MMSE}(\nu_k) = \Gamma(1.5)\frac{\sqrt{\nu_k}}{\gamma_k}\exp\left(-\frac{\nu_k}{2}\right).\left[(1+\nu_k)I_0\left(\frac{\nu_k}{2}\right) + \nu_k I_1\left(\frac{\nu_k}{2}\right)\right] \quad (2.30)$$

Where $\Gamma(\cdot)$ is the Gamma function (with $\Gamma(1.5) = \sqrt{\pi/2}$) and $I_0(.)$ and $I_1(.)$ are the zeroth and first order modified Bessel functions, respectively, defined as:

$$I_n(z) = \frac{1}{2\pi}\int_0^{2\pi} \cos(\beta n)\exp(z\cos\beta)\, d\beta \quad (2.31)$$

In Eq. (2.30), $\nu_k$ is defined as:

$$\nu_k = \frac{\xi_k}{\xi_k + 1}\gamma_k \quad (2.32)$$

Where $\xi_k$ and $\gamma_k$ are defined by:

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \quad (2.33)$$

$$\gamma_k = \frac{{R_k}^2}{\lambda_d(k)} \quad (2.34)$$

$\xi_k$ and $\gamma_k$ are interpreted as the a priori and a posteriori signal-to-noise ratios (SNR), respectively. . $R_k$ denotes the spectral magnitude of the noisy signal.

Essentially, a priori SNR is the Signal-to-Noise Ratio of the $k^{th}$ spectral component of the "clean" speech signal, x[n], while a posteriori SNR is the $k^{th}$ spectral component of the corrupted signal, y[n]. Computation of $\gamma_k$ is straightforward—ratio of the variance of the noisy speech signal to the estimated noise variance. However, computation of a priori SNR is more involved, especially since the knowledge of "clean" signal is seldom available in real systems.

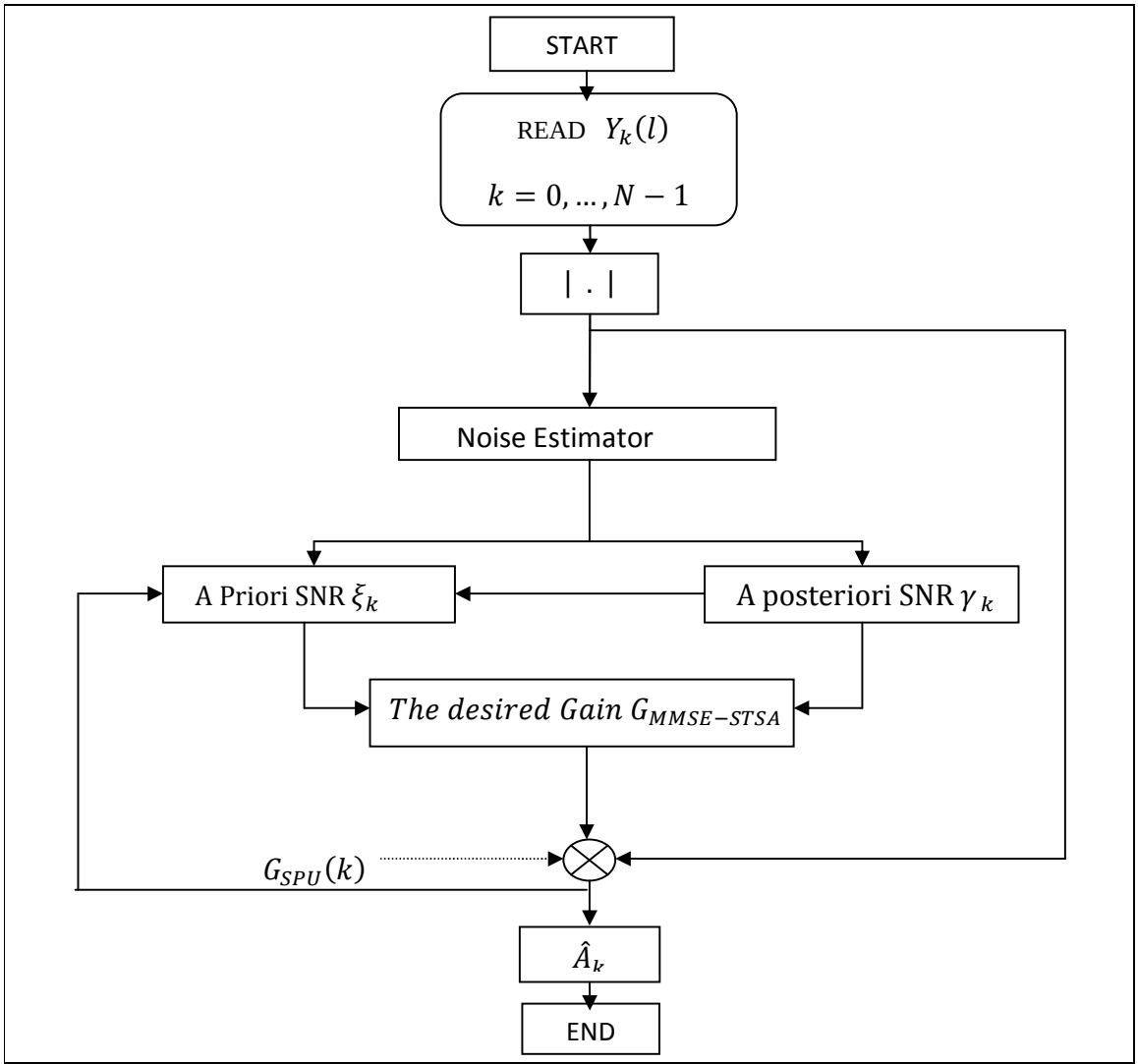


**Figure 2.8**: Algorithm for Spectral Amplitude estimation $\hat{A}_k$ using MMSE-STSA.

In this project "Decision-Directed" estimation has been exploited to compute a priori SNR which is addressed next.

### 2.8.2 Decision-Directed Estimation Approach

In the proposed estimator, the a priori SNR $\xi_k$ is unknown and we have to estimate it in order to implement the estimator. The reason $\xi_k$ is unknown is because the clean signal is unavailable. The "decision-directed" approach [34] was used to estimate $\xi_k$.

From Eq. (2.1) and the Fourier expansion definitions, it is seen that the expected value of the variance of noisy data can be expressed as:

$$E\{R_k^2\} = E\{A_k^2\} + 2E\{A_k\}E\{D_k\} + E\{D_k^2\} \quad (2.35)$$

Since, the clean speech is assumed uncorrelated to noise, the cross-term is set to zero. This simplifies the analysis and results in the instantaneous SNR:

$$\xi_k(n) = E\{\gamma_k(n) - 1\} \tag{2.36}$$

And also, based on the definition of the a priori SNR $\xi_k$

$$\xi_k(n) = \frac{E\{A_k^2(n)\}}{\lambda_d(k,n)} \tag{2.37}$$

*where* $\xi_k(n), A_k, \lambda_d(k,n), and\ \gamma_k(n)$ denote the a priori SNR, the speech magnitude, the noise variance, and the a posteriori SNR, respectively, of the $k^{th}$ corresponding spectral component in the $n^{th}$ analysis frame.

Combining Eq. (2.36) and Eq. (2.37) we have:

$$\xi_k(n) = E\left\{\frac{A_k^2(n)}{2.\lambda_d(k,n)} + \frac{1}{2}[\gamma_k(n) - 1]\right\} \tag{2.38}$$

The proposed estimator $\widehat{\xi_k}$ of $\xi_k$ is deduced [30] from Eq. (2.38), and is given by:

$$\widehat{\xi_k}(n) = \alpha\frac{\widehat{A_k}^2(n-1)}{\lambda_d(k,n-1)} + (1-\alpha)P\{\gamma_k(n) - 1\}, \qquad 0 \le \alpha < 1, \tag{2.39}$$

Where $\widehat{A_k}(n-1)$ is the amplitude estimator of the $k^{th}$ signal spectral component in the $(n-1)^{th}$ analysis frame and $\alpha$ is a weighting constant that is deduced from experimental data. The operator $P\{.\}$ is defined by:

$$P\{x\} = f(x) = \begin{cases} x, & x \ge 0 \\ 0, & otherwise \end{cases} \tag{2.40}$$

Where this positive operator is used to ensure the positiveness of the proposed estimator $\widehat{\xi_k}(n)$ in case $(\gamma_k(n) - 1)$ is negative.

The above estimator for $\xi_k(n)$ is a "decision-directed" type estimator, since $\widehat{\xi_k}(n)$ is updated on the basis of a previous amplitude estimate.

The initial conditions need to be determined and $\widehat{\xi_k}(0) = \alpha + (1-\alpha)P\{\gamma_k(0) - 1\}$ is found appropriate based on simulations since it minimizes initial transition effects in the enhanced speech [34].

### 2.8.3 Amplitude Estimator under Speech Presence Uncertainty (SPU)

Signal absence in noisy observations {y[n], 0 ≤ n ≤ N} is frequent, as speech signals generally contain large portions of silence, [34]. Nevertheless, it does not mean that speech is never present in noisy sections.

The idea of utilizing the uncertainty of signal presence in the noisy spectral components for improving speech enhancement results was first proposed by McAulay and Malpass [29, 34]. The MMSE estimator which accounts for uncertainty of speech presence in noisy observation was first developed by Middleton and Esposito, [34, 35] and it is based on the model of statistically independent random appearance of signal in noisy spectral components.

The derivation of the "likelihood ratio computation" and "a priori probability of speech absence" blocks are addressed below.

We consider a two-state model for speech events, that is, either speech is present at a particular frequency bin (hypothesis $H_1^k$) or that is not (hypothesis $H_0^k$). This is expressed mathematically using the following binary hypothesis model [36]:

- Null Hypothesis $H_0^k : speech\ absent: Y_k = D_k$.
- Alternate Hypothesis, $H_1^k : speech\ present: Y_k = X_k + D_k$.

In view of these hypotheses, Eq. (2.27), can be re-written more explicitly as:

$$\hat{A}_k = E\{ A_k \backslash Y_k, H_1^k \} P\{ H_1^k \backslash Y_k \} + E\{ A_k \backslash Y_k, H_0^k \} P\{ H_0^k \backslash Y_k \}, \tag{2.41}$$

Where $P\{ H_i^k \backslash Y_k \}$, $for\ (i = 0,1)$, is the probability that the speech in state $H_i^k$, for the $k^{th}$ spectral component, given that $Y_k$ is measured.

Obviously, $E\{ A_k \backslash Y_k, H_0^k \}$ is zero as it represents the average value of $A_k$ given $Y_k$ when speech is absent ($H_0^k$). Hence, Eq. (2.41), can be reduced to:

$$\hat{A}_k = E\{ A_k \backslash Y_k, H_1^k \} P\{ H_1^k \backslash Y_k \}. \tag{2.42}$$

We may remark that $E\{ A_k \backslash Y_k, H_1^k \}$ replaces $E\{ A_k \backslash Y_k \}$, identically, in Eq. (2.27) with an added assertion of the alternate hypothesis. Therefore, $P\{ H_1^k \backslash Y_k \}$ defines the multiplicative modifier on the optimal estimator under the signal presence hypothesis.

Bayes' rule [36] can be used to compute $P\{ H_1^k \backslash Y_k \}$:

$$P\{ H_1^k \backslash Y_k \} = \frac{P\{Y_k \backslash H_1^k\} P\{H_1^k\}}{P\{Y_k \backslash H_1^k\} P\{H_1^k\} + P\{Y_k \backslash H_0^k\} P\{H_0^k\}} = \frac{\Lambda(Y_k, q_k)}{1 + \Lambda(Y_k, q_k)} = G_{SPU}(k), \quad (2.43)$$

Where

$$\Lambda(Y_k, q_k) = \mu_k \frac{P\{ Y_k \backslash H_1^k \}}{P\{ Y_k \backslash H_0^k \}} \qquad \mu_k = \frac{P\{H_1^k\}}{P\{H_0^k\}} = \frac{1 - q_k}{q_k}. \quad (2.44)$$

$\Lambda(Y_k, q_k)$ is the generalized likelihood ratio while $q_k$ denotes the *a priori* probability of speech absence in the $k^{th}$ spectral component.

In order to derive the new amplitude estimator we need to calculate $\Lambda(Y_k, q_k)$. By using the Gaussian statistical model assumed for the spectral components we obtain:

$$P\{ Y_k \backslash H_0^k \} = \frac{1}{\pi \lambda_d(k)} \exp\left(-\frac{{Y_k}^2}{\lambda_d(k)}\right) \quad (2.45)$$

$$P\{ Y_k \backslash H_1^k \} = \frac{1}{\pi[ \lambda_d(k) + \lambda_x(k)]} \exp\left(-\frac{{Y_k}^2}{[ \lambda_d(k) + \lambda_x(k)]}\right) \quad (2.46)$$

Using equations (2.45) and (2.46) we get:

$$\Lambda(Y_k, q_k, {\xi'}_k) = \frac{1 - q_k}{q_k} \frac{\exp\left(\frac{{\xi'}_k}{1 + {\xi'}} \gamma_k\right)}{1 + {\xi'}_k} \quad (2.47)$$

Where ${\xi'}_k$ is the conditional a priori $SNR$:

$${\xi'}_k \triangleq E\{ A_k \backslash H_1^k \} \quad (2.48)$$

$${\xi'}_k = \frac{1}{1 - q_k} \xi_k \quad (2.49)$$

**2.9 Speech Enhancement using a MMSE Log-Spectral Amplitude Estimator**

Based on [37] Malah and Ephraim proposed a new short time spectral amplitude (STSA) estimator for speech signals which minimizes the mean squared error of the log spectra.

This section will briefly discuss the derivation of the minimum mean squared error log spectral amplitude (MMSE- LSA).

In order to derive MMSE- LSA, Malah and Ephraim used the same formulation of the estimation problem and the same statistical model as in [34] (modeling speech and noise spectral components as statistically independent Gaussian random variables).

We are looking for the estimator $\widehat{A_k}$, which minimizes the following distortion measure [37]:

$$E\{(logA_k - log\widehat{A_k})^2\} \quad (2.50)$$

Hence, by giving the noisy signal {y(n), $0 \leq n \leq N-1$}, the estimator $\widehat{A_k}$ can be expressed as:

$$\widehat{A_k} = exp\{E[lnA_k \backslash \mathrm{y(n)}, \quad 0 \leq \mathrm{n} \leq \mathrm{N} - 1]\} \tag{2.51}$$

And it is independent on the selected basis for the $log$ in (2.50). Under the assumed statistical model, the mean value of $A_k$ given {y(n), 0 ≤ n ≤ N −1} equals to the mean value of $A_k$ given only $Y_k$, the estimator in (2.51) can be expressed as:

$$\widehat{A_k} = exp\{E[lnA_k \backslash Y_k, \quad 0 \leq k \leq \mathrm{N} - 1]\} \tag{2.52}$$

In order to evaluate (2.52) for the Gaussian model assumed previously, it is conveniently done by using the moment generating function of $lnA_k$ given $Y_k$. Let $Z_k = lnA_k$. Hence the moment generating function of $\Phi_{Z_k \backslash Y_k}(\mu)$ can be expressed as:

$$\Phi_{Z_k \backslash Y_k}(\mu) = E\{\exp(\mu Z_k) \backslash Y_k\} = E\{A_k^{\mu} \backslash Y_k\}. \tag{2.53}$$

Therefore, $E\{lnA_k \backslash Y_k\}$ is found from $\Phi_{Z_k \backslash Y_k}(\mu)$ by:

$$E\{lnA_k \backslash Y_k\} = \frac{d}{d\mu} \Phi_{Z_k \backslash Y_k}(\mu) \Big| \mu = 0. \tag{2.54}$$

With the same definitions for a priori and a posteriori SNR (discussed during the MMSE- STSA derivation), it is not a difficult task to obtain the desired MMSE- LSA gain function (for more details refer to [37]):

$$G_{MMSE-LSA}(\xi_k, \gamma_k) = \frac{\xi_k}{1 + \xi_k} \left\{ \frac{1}{2} \int_{\nu_k}^{\infty} \frac{e^{-t}}{t} dt \right\}, \tag{2.55}$$

Where $\nu_k = \frac{\xi_k}{\xi_k + 1} \gamma_k$ as shown previously during MMSE- STSA estimator derivation.

The MMSE-LSA estimator may be also modified using the multiplicative gain $G_{SPU}(k)$ defined previously for the MMSE-STSA estimator.

### 2.10 Speech Enhancement using the Optimally-Modified Log-Spectral Amplitude estimator (OM-LSA)

The purpose of this section is to study the Optimally-Modified Log-Spectral Amplitude estimator (OM-LSA) proposed by I. Cohn [38]. As the name suggests, it estimates $\hat{A}_k$ by minimizing mean-squared error of the log-spectra for speech signals under signal presence uncertainty where

the spectral gain function is obtained as a weighted geometric mean of the hypothetical gains associated with signal presence and absence.

In this algorithm, Cohen [38] proposed two important estimators:

1) An estimator for the a priori signal-to-noise ratio.
2) An efficient estimator for the a priori speech absence probability (SAP) which is based on the time-frequency distribution of the a priori SNR.

### 2.10.1 The Optimal Gain Modification

Let $H_0^k$ and $H_1^k$ designate respectively hypothetical speech absence and presence in the $k^{th}$ frequency bin, and assuming a complex Gaussian distribution of the STFT coefficients for both speech and noise [34]:

- Null Hypothesis $H_0^k : speech\ absent: Y_k = D_k.$
- Alternate Hypothesis, $H_1^k : speech\ present: Y_k = X_k + D_k.$

The LSA estimator for the clean speech spectral amplitude (Assuming statistically independent spectral components [37]), which minimizes the mean-squared error of the log spectra, is given by:

$$\widehat{A_k} = exp\{E[lnA_k \backslash Y_k,\ 0 \leq k \leq \mathrm{N} - 1]\} \triangleq G_{k\ OM-LSA}.|Y_k| \tag{2.56}$$

Based on the speech presence uncertainty [38]

$$E\{lnA_k \backslash Y_k\} = E\{lnA_k \backslash Y_k, H_1^k\}.p_k + E\{lnA_k \backslash Y_k, H_0^k\}.(1 - p_k) \tag{2.57}$$

Where: $p_k = P\left(H_1^k \backslash Y_k\right),\ 0 \leq k \leq \mathrm{N} - 1.$

When speech is absent, the spectral gain is constrained to be larger than a threshold $G_{min}$, which is determined by subjective criteria for the noise naturalness [38]. Hence,

$$exp\{E[lnA_k \backslash Y_k, H_0^k]\} = G_{min}.|Y_k| \tag{2.58}$$

When speech is present, we use Ephraim and Malah's MMSE-LSA estimator [37]:

$$exp\{E[lnA_k \backslash Y_k, H_1^k]\} = G_{k\ H_1}.|Y_k| \tag{2.59}$$

Where, $G_{k\ H_1}$ is defined (as defined previously in eq (2.55) by:

$$G_{H1}(\xi'_k, \gamma_k) = \frac{\xi'_k}{1+\xi'_k}\left\{\frac{1}{2}\int_{\nu_k}^{\infty}\frac{e^{-t}}{t}dt\right\}, \quad (2.60)$$

$where\ \xi'_k: a\ priori\ SNR\ ,\ \gamma_k: a\ posteriori\ SNR, and\ \nu_k = \frac{\xi'_k}{\xi'_k+1}\gamma_k.$

By substituting (2.58), and (2.59) into (2.56), The Optimally Modified LSA estimator gain is given by:

$$G_{k\ OM-LSA} = \{G_{H1}(\xi'_k, \gamma_k)\}^{p_k}.{G_{min}}^{1-p_k}, \qquad where\ 0 \le k \le \mathrm{N}-1 \quad (2.61)$$

### 2.10.2 *A Priori* SNR Estimation

According to the decision-directed approach, proposed by Ephraim and Malah [34], it provides a useful estimation method for the non-conditional a priori SNR $\xi_k$ which was given previously by eq (2.39): $\widehat{\xi_k}(l) = \alpha\frac{\widehat{A_k}^2(l-1)}{\lambda_d(k,l-1)} + (1-\alpha)max\{\gamma_k(l)-1, 0\}$, $where\ 0 < \alpha < 1$, and $l$ is the frame number.

Therefore the estimate for the a priori SNR should be given by: $\xi'_k = \frac{1}{1-q_k}\xi_k$. According to this expression, there is an interaction between the estimated $q_k$ and the a priori SNR which may deteriorate the performance of the speech enhancement system [39], [40], [41].

Hence, Cohen in [38] proposed a new estimator of the *a Priori SNR* which is given as follows:

$$\widehat{\xi'_k}(l) = \alpha G_{H1}(l-1)^2 + (1-\alpha)max\{\gamma_k(l)-1, 0\} \quad (2.62)$$

To explain more this equation, if $H_1^k$ is true then the spectral gain should degenerate to $G_{H1}$, and the a priori SNR $\xi'_k$ estimate should coincide with $\xi_k$. In opposite, if $H_0^k$ is true, then the spectral gain should decrease to $G_{min}$, or equivalently the a priori SNR estimate should be as small as possible[38].

### 2.10.3 *A Priori* Speech Absence Probability (SAP) Estimation

In [38] , Cohen proposed a new estimator for the speech absence probability $\widehat{q_k}$. The estimator utilizes a soft-decision approach in order to find three parameters $(P_{k\ local}(l), P_{k\ global}(l), P_{frame}(l))$ based on the time-frequency distribution of the estimated a priori SNR, $\widehat{\xi'_k}(l)$. These parameters exploit the strong correlation of speech presence in neighboring frequency bins of consecutive frames [38].

Let $\xi'_k(l)$ be a recursive average of the *a priori* SNR:

$$\xi'_k(l) = \beta\xi'_k(l-1) + (1-\beta)\,\widehat{\xi'_k}(l-1) \tag{2.63}$$

Where $\beta$ is a time constant.

Local and global averaging windows are applied in the frequency domain to obtain local and global averages of the *a priori* SNR:

$$\xi'_{k_\lambda}(l) = \sum_{i=-w_\lambda}^{w_\lambda} h_\lambda(i)\,\xi'_{k-i}(l) \tag{2.64}$$

Where the subscript $\lambda$ designates either "local" or "global", and $h_\lambda$ is a normalized window of size $2w_\lambda$+1.

We define two parameters, $P_{k\,local}(l)$ and $P_{k\,global}(l)$, which represent the relation between the above averages and the likelihood of speech in the $k^{th}$ frequency bin of the $l^{th}$ frame[38]. The local and global parameters are given by the following expression:

$$P_{k_\lambda}(l) = \begin{cases} 0, & if\ \xi'_{k_\lambda}(l) \le \xi'_{min} \\ 1, & if\ \xi'_{k_\lambda}(l) \ge \xi'_{max} \\ \dfrac{\ln\left(\xi'_{k_\lambda}(l)/\xi'_{min}\right)}{\ln\left(\xi'_{max}/\xi'_{min}\right)}, & otherwise \end{cases} \tag{2.65}$$

Where $\xi'_{min}$ and $\xi'_{max}$ are empirical constants.

For more noise attenuation in noise-only frames, a third parameter named, $P_{frame}(l)$ is defined. This parameter is based on the speech energy in neighboring frames. If we average $\xi'_k(l)$ in the frequency domain we obtain:

$$\xi'_{frame}(l) = \underset{1\le k\le \frac{N}{2}+1}{\text{mean}} \left\{\xi'_k(l)\right\} \tag{2.66}$$

Figure 2.9 shows the block diagram for $P_{frame}(l)$ computation.

Where $I(l)$ is given by:

$$I(l) \triangleq \begin{cases} 0, & if\ \xi'_{frame}(l) \le \xi'_{peak}(l).\xi'_{min} \\ 1, & if\ \xi'_{frame}(l) \ge \xi'_{peak}(l).\ \xi'_{max} \\ \dfrac{\ln\left(\xi'_{frame}(l)/\xi'_{peak}(l)/\xi'_{min}\right)}{\ln\left(\xi'_{max}/\xi'_{min}\right)}, & otherwise \end{cases} \tag{2.67}$$

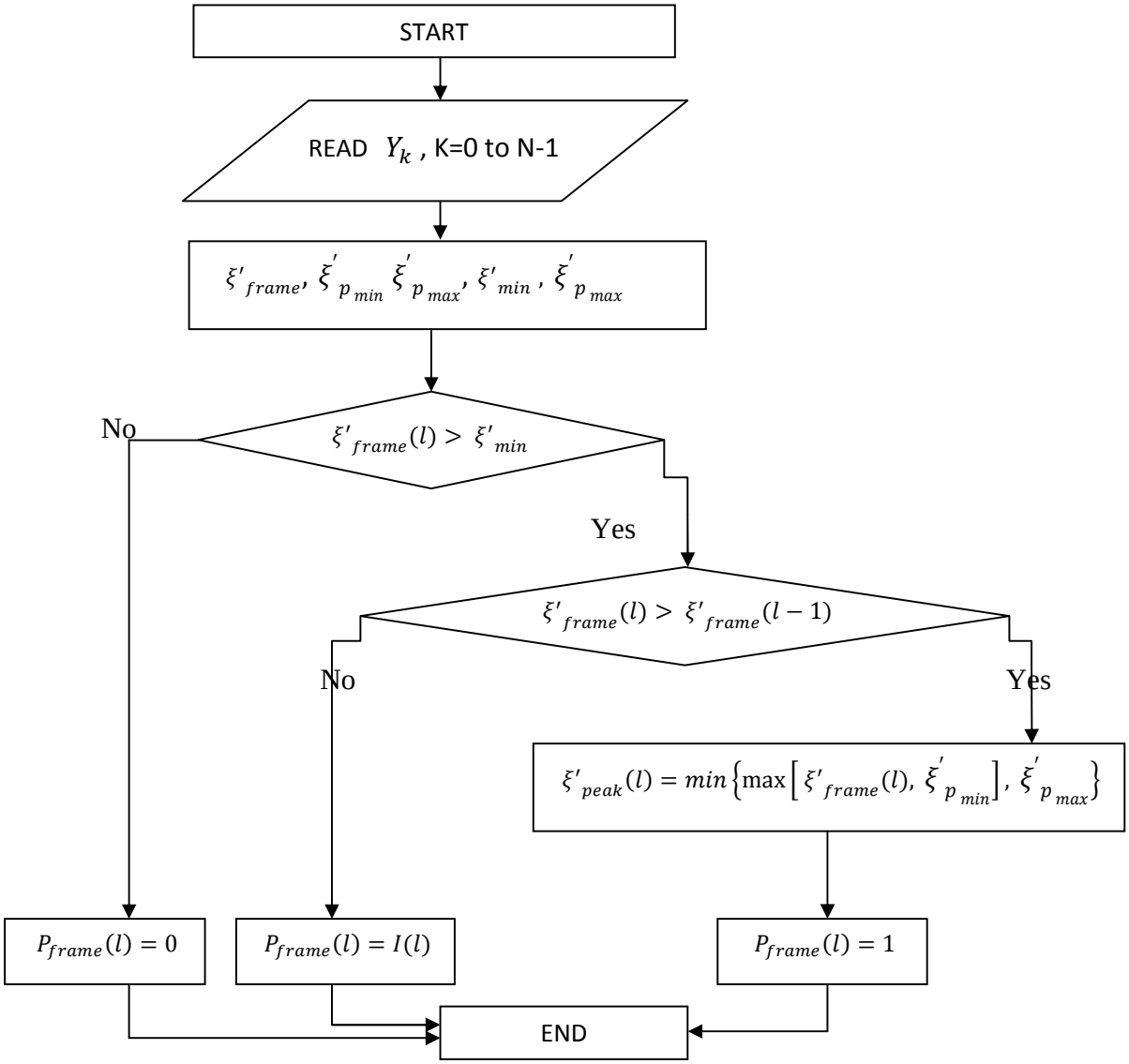


**Figure 2.9**: The Block diagram for $P_{frame}(l)$ computation.

And it represents a soft transition from "speech" to "noise", $\xi'_{peak}$ is a confined peak value of $\xi'_{frame}$, and $\xi'_{p_{min}}$, $\xi'_{p_{max}}$ are empirical constants that determine the delay of the transition [38].

Hence, the proposed estimate for the a priori probability for speech absence is obtained by:

$$\hat{q}_k(l) = 1 - P_{k_{local}}(l).P_{k_{global}}(l).\ P_{frame}(l) \qquad (2.67)$$

$\hat{q}_k(l)$ is larger if either previous frames, or recent neighboring frequency bins, do not contain speech. In order to reduce the possibility of speech distortion we restrict $\hat{q}_k(l)$ to be smaller than a threshold $q_{max}$ (<1).

# CHAPTER 3

# Blind Multi-speaker speech signal separation using DUET Algorithm

Blind Source Separation is a newly formed field of fundamental research with wide areas of applications. This later is motivated by practical problems that involve several source signals and several microphones. Each microphone receives a linear mixture of the source signals. The problem of the blind source separation consists, then, of recovering a set of the original unobserved speech signals from a set of a given mixtures [42]. Typically the observations are obtained at the output of a set of sensors (microphones), where each sensor receives different mixtures of source signals. The adjective “blind” emphasizes two facts [43]:

1) The source signals are not observed.
2) No information is available about the mixing system.

The mixture is often a convolutive mixture. However, in this project, our main concern is the blind separation of an instantaneous linear mixture. There exist ambiguities. To cite ref. [42]: "This problem of blind source separation has two inherent ambiguities. First, it is not possible to know the original labeling of the sources; hence, any permutation of the estimated sources is also a satisfactory solution. The second ambiguity is that it is inherently impossible to uniquely identify the source signals. This is because the exchange of a fixed scalar factor between a source signal and the corresponding column of the mixture matrix does not affect the observations".

When the number of mixtures is less than the number of sources, the blind source separation problem is undetermined [44]. However, the property of disjoint orthogonality allows us to overcome this problem. The ratio of the Short time Fourier Transform (STFT) coefficients of signals received at two sensors can factor out the role of the power spectrum of emitting sources, under an assumption called Disjoint Orthogonality. Thus, it can reveal parameters specific to the mixing scenario and serve as a basis for channel estimation techniques. In this project we investigate the applicability of the Blind source separation method (with unknown number of sources) based on the use STFT ratios of two sensor inputs only, called the Degenerate Unmixing and Estimation Technique (DUET) [45](see figure 3.1).

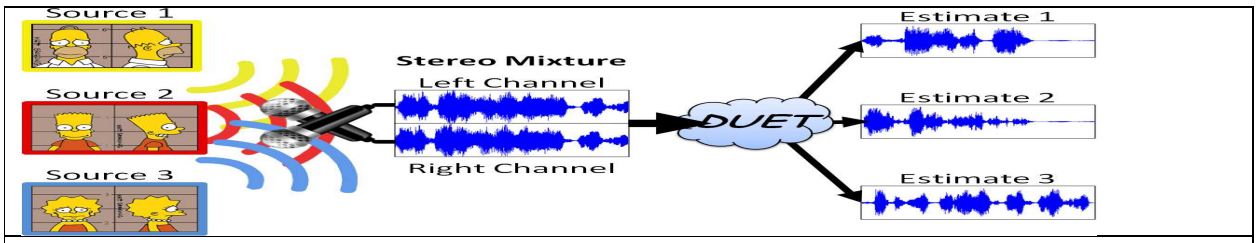


**Figure 3.1:** Blind source separation using DUET algorithm [46]

### 3.1 Introduction to Degenerate Unmixing Estimation Technique (DUET)

DUET [45] is a Blind source separation technique capable of separating a number of sources from 2 Mixtures. The principle behind DUET can be summarized in this sentence:

"It is possible to blindly separate an arbitrary number of sources given just two anechoic mixtures provided the time–frequency representations of the sources do not overlap too much, which is true for speech"[47].

### 3.2 Sources assumptions

#### 3.2.1 Anechoic Mixing

Let's consider the mixtures of N speakers' speech signals, $s_j(t), j = 1, \dots, N$ being detected at a pair of sensors (microphones) where only the direct path is present. The two anechoic mixtures can thus be given as,

$$x_1(t) = \sum_{j=1}^{N} s_j(t), \tag{3.1}$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t - \delta_j), \tag{3.2}$$

Where N is the number of speakers, $\delta_j$ is the arrival delay between the sensors, and $a_j$ is a relative attenuation factor corresponding to the ratio of the attenuations of the paths between speakers and sensors [48].

#### 3.2.2 W-Disjoint Orthogonality (WDO)

In mathematics, disjoint means if two or more sets are disjoint they have no element in common, or say their intersection is the empty set.

The two functions $s_j(t)$ *and* $s_k(t)$ are called **W-disjoint orthogonal** if, for a given windowing function $W(t)$, the supports of the windowed Fourier transforms of $s_j(t)$ *and* $s_k(t)$ *are* disjoint.

The windowed Fourier transform of $s_j(t)$ is defined,

$$\hat{s}_j(\tau, \omega) := F^W[s_j](\tau, \omega) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(t - \tau)\, s_j(t) e^{-i\omega t}\, dt \tag{3.3}$$

The WDO assumption can be expressed shortly and clearly as,

$$\hat{s}_j(\tau,\omega)\,\hat{s}_k(\tau,\omega) = 0,\ \forall\tau,\omega,\ \forall j \neq k \tag{3.4}$$

This assumption is the mathematical idealization of the condition that it is likely that every T-F point in the mixture with significant energy is dominated by the contribution of one source. Note that, if $W(t) \equiv 1$, $\hat{s}_j(\tau,\omega)$ becomes the Fourier transform of $s_j(t)$ which we will denote $\hat{s}_j(\omega)$. In this case, W-disjoint orthogonality can be expressed as,

$$\hat{s}_j(\omega)\,\hat{s}_k(\omega) = 0,\ \forall j \neq k, \forall\omega, \tag{3.5}$$

which we call **disjoint orthogonality**[49].

Since the W-disjoint orthogonality assumption is not exactly satisfied for many categories of signals, the concept of approximate W-disjoint orthogonality introduced in [37] provides a practical version for the basic assumption. Approximate W-disjoint orthogonality assumes that at each point of the time-frequency representation of a mixture, the power of, at most, one source signal will be dominant.

### 3.2.3 Local Stationarity and Microphones separation [47]

The Fourier transform pair is:

$$s_j(t-\delta) \leftrightarrow e^{-i\omega\delta}\,\hat{s}_j(\omega). \tag{3.7}$$

The local stationarity assumption can be officially given as,

$$F^W\left[s_j(.-\delta)\right](\tau,\omega) = e^{-i\omega\delta}F^W\left[s_j(.)\right](\tau,\omega), \quad \forall\delta, |\delta| \leq \Delta \tag{3.8}$$

where $\Delta$ is the maximum time difference possible in the mixing model (the sensors separation divided by the speed of signal propagation).

Of course, the multiplicative factor $e^{-i\omega\delta}$ only uniquely specifies δ if $|\omega\delta| < \pi$ as otherwise we have an ambiguity due to phase-wrap. So we require,

$$|\omega\delta_j| < \pi, \forall\omega, \forall j, \tag{3.9}$$

to avoid phase ambiguity [48]. This is guaranteed when the separation of the sensors is less than $\pi c/\omega_m$, where $\omega_m$ is the maximum frequency present in the speech sources and c is the speed of sound.

## 3.3 DUET demixing model and parameters [47]

The assumptions of anechoic mixing and local stationarity allow the mixing equations (3.1) and (3.2) in the T-F domain to be written as,

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 \\ a_1 e^{-i\omega\delta_1} & \dots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(\tau,\omega) \\ . \\ . \\ . \\ \hat{s}_N(\tau,\omega) \end{bmatrix} \tag{3.10}$$

With the further assumption of W-disjoint orthogonality, at most one source is active at every $(\tau,\omega)$, and the mixing process can be expressed as,

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} \hat{s}_j(\tau,\omega), \quad for\ some\ j \tag{3.11}$$

Where $j$ is the index of the speech source active at $(\tau,\omega)$. The main DUET observation which is the ratio of the time-frequency representations of the mixtures does not depend on the speech source components but only on the mixing parameters associated with the active speech source components.

The mixing parameters associated with each T-F point can be computed as,

$$\tilde{a}(\tau,\omega) := |\hat{x}_2(\tau,\omega)/\hat{x}_1(\tau,\omega)| \tag{3.12}$$

$$\tilde{\delta}(\tau,\omega) := \left(-\frac{1}{\omega}\right) \angle \left(\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)}\right) \tag{3.13}$$

Under the assumption that if the two sensors are sufficiently close then the delay estimation can be ignored, the local attenuation estimator $\tilde{a}(\tau,\omega)$ and the local delay estimator $\tilde{\delta}(\tau,\omega)$ can only take on the values of the actual mixing parameters.

We can demix via binary masking by determining the indicator function of each source. So the indicator functions are found via,

$$M_j(\tau,\omega) := \begin{cases} 1 & \left(\tilde{a}(\tau,\omega),\tilde{\delta}(\tau,\omega)\right) = (a_j,\delta_j) \\ 0 & otherwise \end{cases} \tag{3.14}$$

And then we demix using the masks. The union of the $\left(\tilde{a}(\tau,\omega),\tilde{\delta}(\tau,\omega)\right)$ pairs taken over the entire time-frequency plane $(\tau,\omega)$ is the set of mixing parameters $(a_j,\delta_j), J = 1 \dots N$.

### 3.4 Construction of the 2D weighted histogram

Histogram is the key structure used for localization and separation. By using $\left(\tilde{a}(\tau,\omega),\tilde{\delta}(\tau,\omega)\right)$ pairs to indicate the indices into the histogram, clusters of weight will emerge centered on the actual mixing parameter pairs $(a_j,\delta_j)$ [48]. Figure 3.2 shows the two-dimensional weighted histogram.

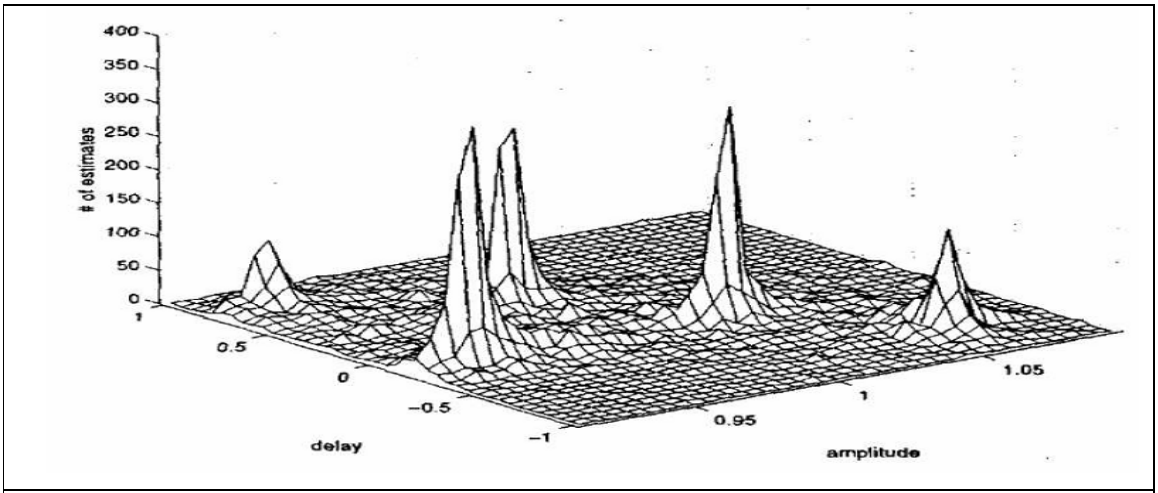


**Figure 3.2:** duet two-dimensional cross power weighted histogram of symmetric attenuation ( $a_j-\frac{1}{a_j}$) and delay estimate pairs from two mixtures of five sources. each peak corresponds to one source and the peak locations reveal the source mixing parameters [50].

We can formally define that the weighted histogram separates and clusters the parameter estimates of each source. The number of peaks corresponding to the number of sources, and the peak locations reveal the associated source's anechoic mixing parameters.

There are several different automatic peak identification methods including weighted $k-means$, model-based peak removal, and peak tracking [51]. Once the peaks have been identified, our purpose is to determine the time-frequency masks which will separate each source from the mixtures.

**3.5 Maximum-likelihood estimators** [47]

The assumptions made previously will not be satisfied in real-time (real signals with noise) cases, we need a mechanism for clustering the relative attenuation-delay estimates. Thus, we consider the ''maximum likelihood (ML) estimators'' for the $a_j$ attenuation factor and the $\delta_j$ delay factor in the following mixing model:

$$\begin{bmatrix} \hat{x}_1(\tau,\omega) \\ \hat{x}_2(\tau,\omega) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} \hat{s}_j(\tau,\omega) + \begin{bmatrix} \hat{n}_1(\tau,\omega) \\ \hat{n}_2(\tau,\omega) \end{bmatrix} \quad (3.15)$$

Where $\hat{n}_1$ and $\hat{n}_2$ are noise terms which represent the assumption inaccuracies. One thing we need to point out is: rather than estimating $a_j$, we estimate $\alpha_j := a_j - \frac{1}{a_j}$ which we call the "symmetric attenuation". That is, the attenuation is reflected symmetrically about a centre point ($\alpha = 0$) because it has the property that the two microphone (sensor) signals can be swapped [36]. We can define the local symmetric attenuation estimate,

$$\tilde{\alpha}(\tau,\omega) := |\hat{x}_2(\tau,\omega)/\hat{x}_1(\tau,\omega)| - |\hat{x}_1(\tau,\omega)/\hat{x}_2(\tau,\omega)| \quad (3.16)$$

Motivated by the form of the ML estimators [45], a pair of estimators appear:

$$\tilde{\alpha}_j = \frac{\iint_{(\tau,\omega)\in\Omega_j} |\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)|^p . \omega^q \tilde{\alpha}(\tau,\omega) d\tau d\omega}{\iint_{(\tau,\omega)\in\Omega_j} |\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)|^p . \omega^q d\tau d\omega} \quad (3.17)$$

And

$$\tilde{\delta}_j = \frac{\iint_{(\tau,\omega)\in\Omega_j} |\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)|^p . \omega^q \tilde{\delta}(\tau,\omega) d\tau d\omega}{\iint_{(\tau,\omega)\in\Omega_j} |\hat{x}_1(\tau,\omega)\,\hat{x}_2(\tau,\omega)|^p . \omega^q d\tau d\omega} \quad (3.18)$$

Where $\Omega_j := \{(\tau,\omega): |\hat{s}_j(\tau,\omega)| \gg |\hat{s}_k(\tau,\omega)|, \forall\, k \neq j\}$. The equations are parameterized by p and q.

$p = 1, and\ q = 0$ : motivated by the ML symmetric attenuation estimator [45]

**3.6 Separation of the sources** [47]

The estimators suggest the construction of a two-dimensional weighted histogram to determine the clusters and the estimated mixing parameters pairs $(a_j, \delta_j)$. Thus, the mixing parameters can be extracted by locating the peaks in the histogram.

Let's have the histogram peak centers $(\tilde{\alpha}_j, \tilde{\delta}_j), j = 1 \ldots N$, we convert the symmetric attenuation back to the attenuation via:

$$\tilde{a}_j = \frac{\tilde{\alpha}_j + \sqrt{\tilde{\alpha}_j^{\,2} + 4}}{2} \quad (3.19)$$

assign a peak to each time–frequency point via

$$J(\tau,\omega) := \arg_{\min_k} \frac{\left|\tilde{a}_k e^{-i\tilde{\delta}_k \omega} \hat{x}_1(\tau,\omega) - \hat{x}_2(\tau,\omega)\right|^2}{1+\tilde{a}_k^2} \tag{3.20}$$

and then assign each time–frequency point to a mixing parameter estimate via

$$\tilde{M}_j(\tau,\omega) := \begin{cases} 1 & if\ J(\tau,\omega) = j \\ 0 & otherwise \end{cases} \tag{3.21}$$

We demix via masking and ML combining [47],

$$\tilde{s}_j(\tau,\omega) = \tilde{M}_j(\tau,\omega).\left(\frac{\hat{x}_1(\tau,\omega) + \tilde{a}_j e^{-i\tilde{\delta}_j \omega} \hat{x}_2(\tau,\omega)}{1+\tilde{a}_j^2}\right) \tag{3.22}$$

**3.7 Summary of DUET Algorithm** [47]

The Degenerate Unmixing Estimate Technique algorithm can be summarized as follows:

1) Constructing time-frequency representations $\hat{x}_1(\tau,\omega)$ *and* $\hat{x}_2(\tau,\omega)$ from anechoic mixtures $x_1(t)$ and $x_2(t)$.
2) Calculating the mixing parameters

$$\left(\tilde{a}(\tau,\omega), \tilde{\delta}(\tau,\omega)\right) = \left(\left|\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)}\right| - \left|\frac{\hat{x}_1(\tau,\omega)}{\hat{x}_2(\tau,\omega)}\right|, \left(-\frac{1}{\omega}\right)\angle\left(\frac{\hat{x}_2(\tau,\omega)}{\hat{x}_1(\tau,\omega)}\right)\right).$$

3) Constructing a 2D smoothed weighted histogram for all weights associated with time-frequency plane.
4) Locating peaks and finding peak centers which determine the mixing parameter estimates.
5) Constructing the time-frequency binary masks for each peak centre as given in Eq(3.21).
6) Applying each mask to the appropriately aligned mixtures as given in Eq(3.22).
7) Finally, we convert each estimated source time-frequency representation back into the time domain.

# CHAPTER 4

# Implementation and performance Evaluation

This chapter describes the implementation details and performance evaluation of the proposed pre-processing algorithms to understand their functionality and behavior. Evaluation of speech enhancement algorithms is not simple. While objective quality assessment methods can indicate an improvement or degradation in speech quality based on mathematical measures, the human listener does not believe in a simple mathematical error criterion. Therefore, subjective measurements of intelligibility and quality are also required.

It is necessary to conduct off-line simulations to check the validity and feasibility of an algorithm before it can be implemented on a real-time system. The simulations were carried out on an Acer Aspire 5735Z PC.

**4.1 Implementation and performance evaluation of DFT-based single channel Algorithms**

This Section 4.1 describes the implementation and performance evaluation of six DFT-based single channel speech enhancement algorithms (explained in chapter two) which are as follows:

1. Spectral Subtraction using over-subtraction and spectral floor.
2. Multi-Band Spectral Subtraction (MBSS).
3. Wiener Filter (*a priori SNR* $\xi_k$ *is* calculated using the Decision-Directed method).
4. MMSE of Short-Time Spectral Amplitude (MMSE-STSA) estimator with, and without using SPU multiplicative modifier.
5. MMSE Log-Spectral Amplitude (MMSE-LSA) estimator with, and without using SPU multiplicative modifier.
6. Optimally-Modified Log-Spectral Amplitude estimator (OM-LSA)

The IEEE standard database NOIZEUS (noisy corpus) [52] is used to test algorithms. The database contains clean speech sample files as well as real world noisy speech files at different SNRs and noise conditions like street, car, restaurant, train, station, babble…etc.

The performance comparisons of various implemented algorithms are carried out which are based on visual examinations of signals in the time domain and the spectrograms (clean, noisy, and enhanced speech signals), and also the objective and subjective tests.

**4.1.1 Implementation Details**

The factors contributing in the efficient implementation of some of the functional blocks are discussed below.

- Frame size: 20 ms was chosen as the optimum frame size for our implementations.

- Window Type and Overlap: the most commonly used Hamming window [16][12][53]. After a few informal listening tests and comparing spectrograms, the Hamming window was chosen. The amount of overlap between consecutive frames is also associated with the frame-size, and is required to prevent discontinuities at frame boundaries. For this study we chose the overlap to be 50%, which is also usually the percentage overlap found commonly in the literature.
- The enhanced signal is obtained by taking the IFFT of the enhanced spectrum using the phase of the original noisy spectrum.
- The standard overlap-and-add method is used to obtain the enhanced signal.

For the Spectral Subtraction using over-subtraction and spectral floor, the spectral floor parameter is set to $\beta = 0.002$, and $\alpha = \begin{cases} 4.75 & SNR < -5 \\ 4 - \frac{3}{20}SNR & -5 \leq SNR \leq 20 \\ 1 & SNR > 20 \end{cases}$

For the Multi-Band Spectral Subtraction (MBSS) implementation, the spectral floor parameter is also set to $\beta = 0.002$, and all other parameters are taken as given in chapter two.

For the Wiener filter, MMSE-STSA, and MMSE-LSA algorithms implementations, the *a priori* $SNR\ \xi_k\ is$ calculated using the Decision-Directed estimation approach with $\alpha = 0.98\ in\ Eq.(2.38)$.

For Speech Presence Uncertainty (SPU) multiplicative modifier implementation in Eq. (2.42), the *a priori* probability of speech absence $q_k$, is set to $q_k = 0.3\ in\ Eqs.(2.46), and\ (2.48)$.

For the OM-LSA estimator implementation, the value $\alpha = 0.92\ in\ Eq.(2.61)$, and the values of parameters used for the estimation of the *a priori* SAP are given as follows:

$\beta = 0.7 \qquad \xi'_{min} = -10dB \qquad \xi'_{max} = -5dB \qquad \xi'_{p_{min}} = 0dB \qquad \xi'_{p_{max}} = 10dB$

$w_{local} = 1 \qquad w_{global} = 15 \qquad q_{max} = 0.95 \qquad h_\lambda: Hanning\ windows$

For the implementation of noise estimation algorithm discussed in section 2.5, the following parameters are used:

$\eta = 0.7\ in\ Eq.(2.9), \qquad the$ threshold $\sigma, is\ to\ \sigma = 1.3, \qquad \lambda = 0.8\ in\ Eq.(2.11),$

$\gamma = 0.998, and\ \beta = 0.8\ in\ Eq.(2.12), \qquad \alpha_1 = 0.8, and\ \alpha_2 = 1\ in\ Eq.(2.13).$

### 4.1.2 The noisy database

To test the implemented algorithms, sentences presented in table 4.1 from the noisy database (NOIZEUS) that contains 30 IEEE sentences [52] (produced by three male and three female speakers) corrupted by eight different real-world noises at different SNRs are used.

The noise was taken from the AURORA database [54] and includes suburban train noise, babble noise, car noise, restaurant noise, street noise, and train-station noise.

The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.

| Filename | Speaker | Gender | Sentence text |
|---|---|---|---|
| sp01.wav | CH | M | The birch canoe slid on the smooth planks. |
| sp02.wav | CH | M | He knew the skill of the great young actress. |
| sp03.wav | CH | M | Her purse was full of useless trash. |
| sp04.wav | CH | M | Read verse out loud for pleasure. |
| sp05.wav | CH | M | Wipe the grease off his dirty face. |
| sp06.wav | DE | M | Men strive but seldom get rich. |
| sp07.wav | DE | M | We find joy in the simplest things. |
| sp08.wav | DE | M | Hedge apples may stain your hands green. |
| sp09.wav | DE | M | Hurdle the pit with the aid of a long pole. |
| sp10.wav | DE | M | The sky that morning was clear and bright blue. |
| sp11.wav | JE | F | He wrote down a long list of items. |
| sp12.wav | JE | F | The drip of the rain made a pleasant sound. |
| sp13.wav | JE | F | Smoke poured out of every crack. |
| sp14.wav | JE | F | Hats are worn to tea and not to dinner. |
| sp15.wav | JE | F | The clothes dried on a thin wooden rack. |
| sp16.wav | KI | F | The stray cat gave birth to kittens. |
| sp17.wav | KI | F | The lazy cow lay in the cool grass. |
| sp18.wav | KI | F | The friendly gang left the drug store. |
| sp19.wav | KI | F | We talked of the sideshow in the circus. |
| sp20.wav | KI | F | The set of china hit the floor with a crash. |
| sp21.wav | SI | M | Clams are small, round, soft and tasty. |
| sp22.wav | SI | M | The line where the edges join was clean. |
| sp23.wav | SI | M | Stop whistling and watch the boys march. |
| sp24.wav | SI | M | A cruise in warm waters in a sleek yacht is fun. |
| sp25.wav | SI | M | A good book informs of what we ought to know. |
| sp26.wav | TI | F | She has a smart way of wearing clothes. |
| sp27.wav | TI | F | Bring your best compass to the third class. |
| sp28.wav | TI | F | The club rented the rink for the fifth night. |
| sp29.wav | TI | F | The flint sputtered and lit a pine torch. |
| sp30.wav | TI | F | Let's all join as we sing the last chorus. |

**Table 4.1** List of sentences used in NOIZEUS.

### 4.1.3 Visual Examinations for the implemented algorithms

Applying the implemented algorithms to the noisy speech signal sentence in "sp10.wav" corrupted with train noise at 0 dB SNR, and car noise at 5 dB SNR yields to the results presented along with the original noisy signal in following figures:

From Figure 4.1 to Figure 4.10 show the signals in the time domain of the original sentence in "sp10.wav" along with the same corrupted with speech-shaped train noise at 0 dB SNR, and the enhanced speech obtained by the implemented algorithms.

From Figure 4.11 to Figure 4.20 show the spectrograms of the original sentence in "sp10.wav" along with the same corrupted with speech-shaped car noise at 5 dB SNR, and the enhanced speech obtained from implemented algorithms.

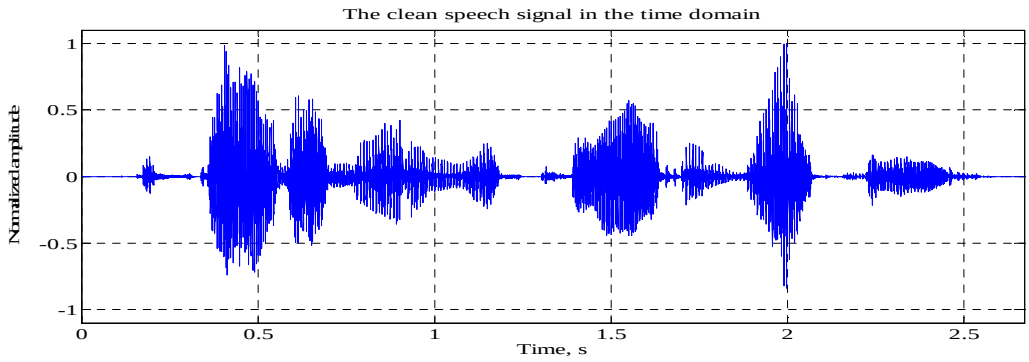


**Figure 4.1:** The clean speech signal in the time domain.

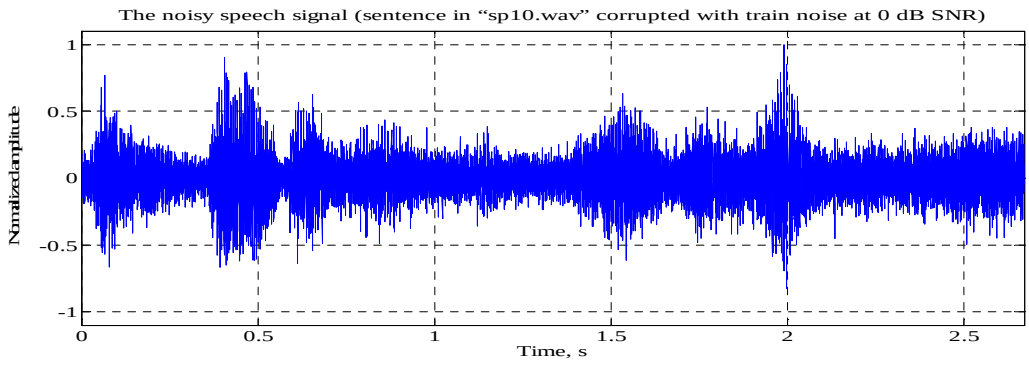


**Figure 4.2:** The noisy speech signal (sentence in "sp10.wav" corrupted with train noise at 0 dB SNR).

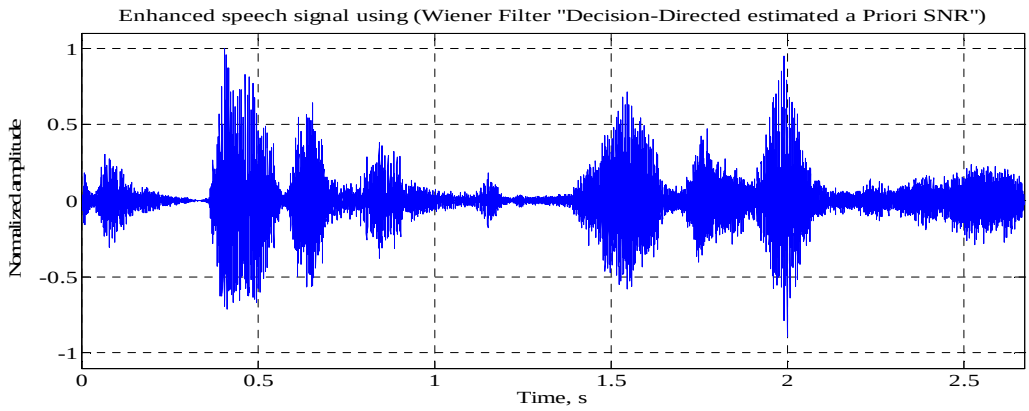


**Figure 4.3:** Enhanced speech signal (Wiener Filter "Decision-Directed estimated a priori SNR").

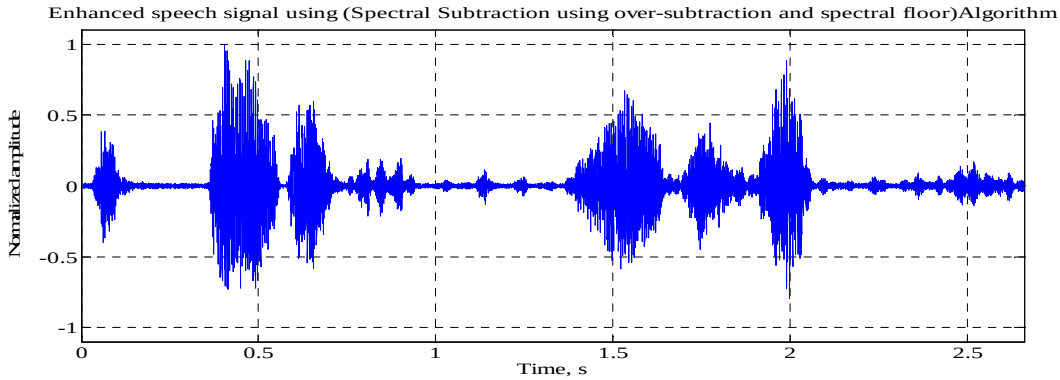


**Figure 4.4:** Enhanced speech signal (Spectral Subtraction using over-subtraction and spectral floor) Algorithm.

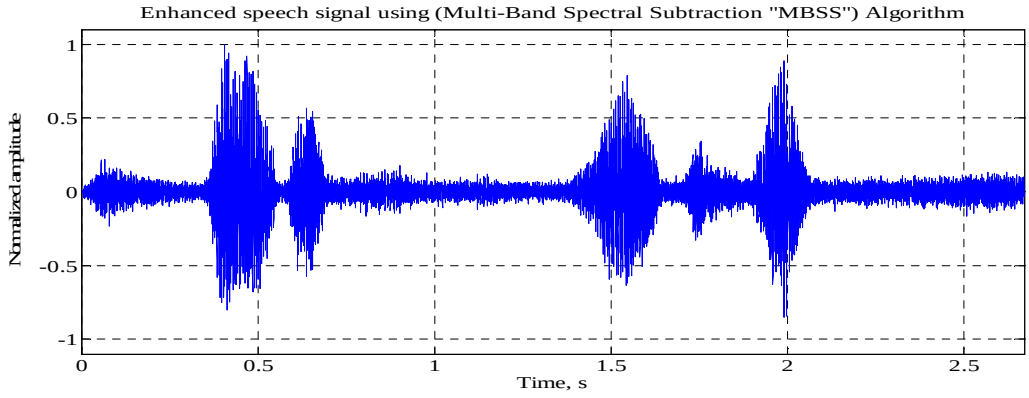


**Figure 4.5:** Enhanced speech signal (Multi-Band Spectral Subtraction "MBSS") Algorithm.

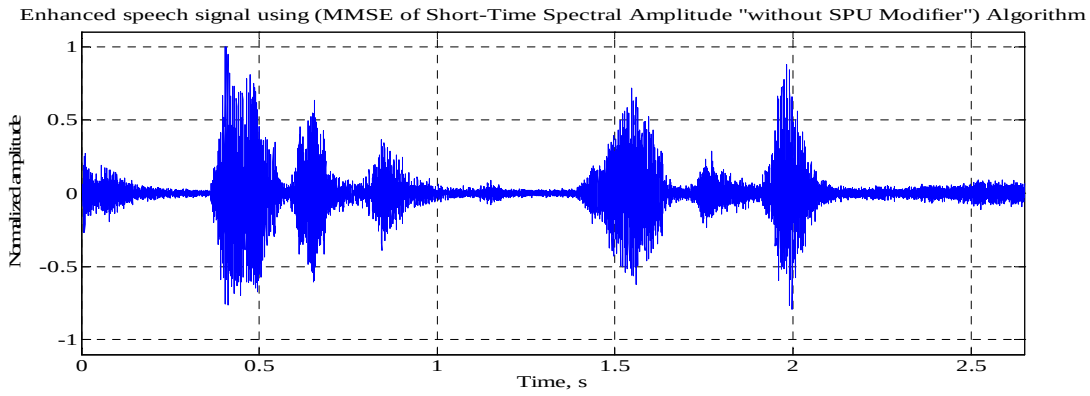


**Figure 4.6:** Enhanced speech signal (MMSE-STSA "without using SPU modifier") Algorithm.

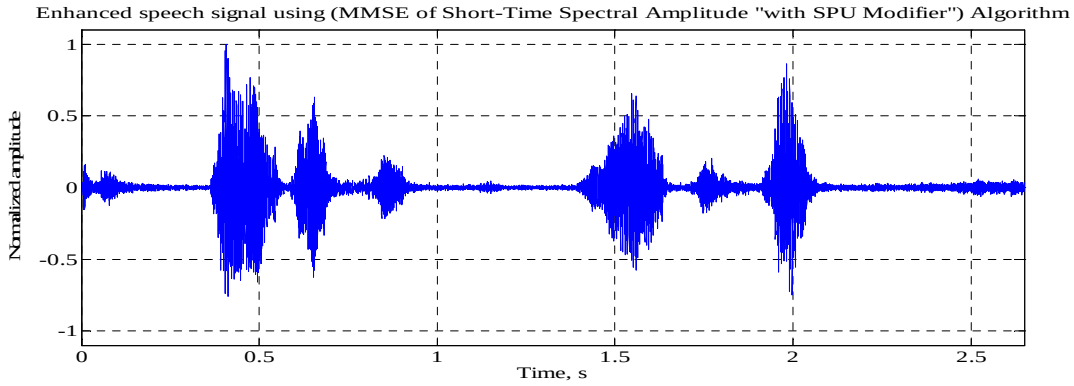


**Figure 4.7:** Enhanced speech signal (MMSE-STSA "using SPU modifier") Algorithm.

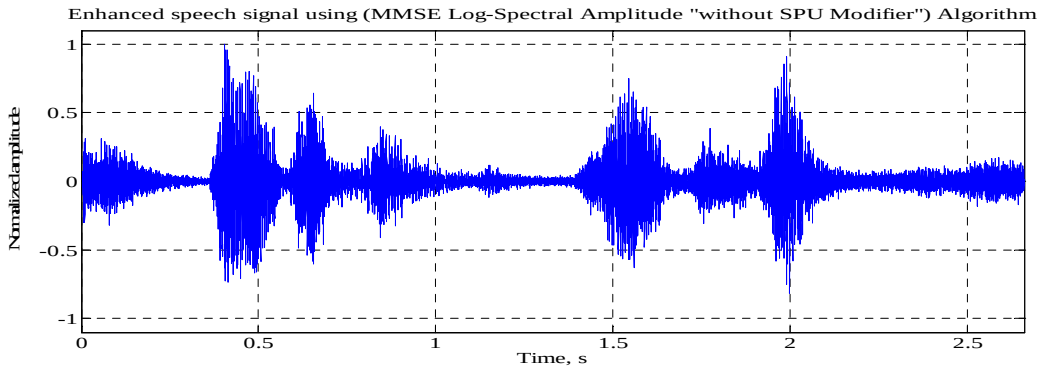


**Figure 4.8:** Enhanced speech signal (MMSE-LSA "without using SPU modifier") Algorithm.

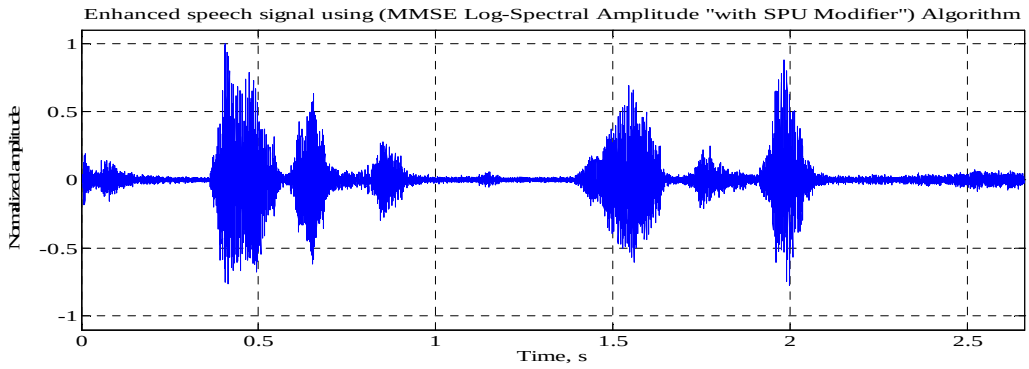


**Figure 4.9:** Enhanced speech signal (MMSE-LSA "using SPU modifier") Algorithm.

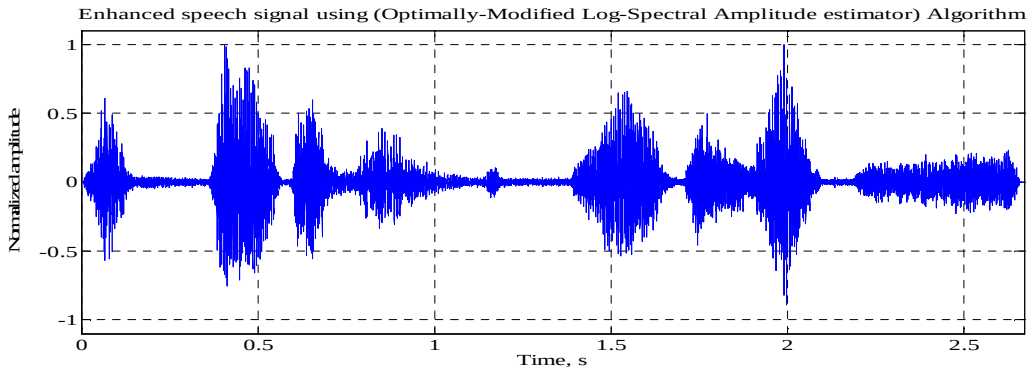


**Figure 4.10:** Enhanced speech signal (OM-LSA) Algorithm.

From visual examinations of figures presented above we can notice that:

- A significant amount of noise has been reduced from the noisy speech signal after applying each of the DFT-based speech enhancement algorithms.
- The enhanced speech panels using Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method show more distortions in the shape of the signals when compared to the original clean speech signal.
- The enhanced speech panels using MMSE-STSA, MMSE-LSA (without using SPU modifier) algorithms show small distortions in the shape of the signals when compared to the original clean speech signal.
- The enhanced speech panels using MMSE-STSA, MMSE-LSA (using the SPU modifier), and the OM-LSA algorithms show that the obtained processed signal shapes are more nearer to the original clean speech signal.

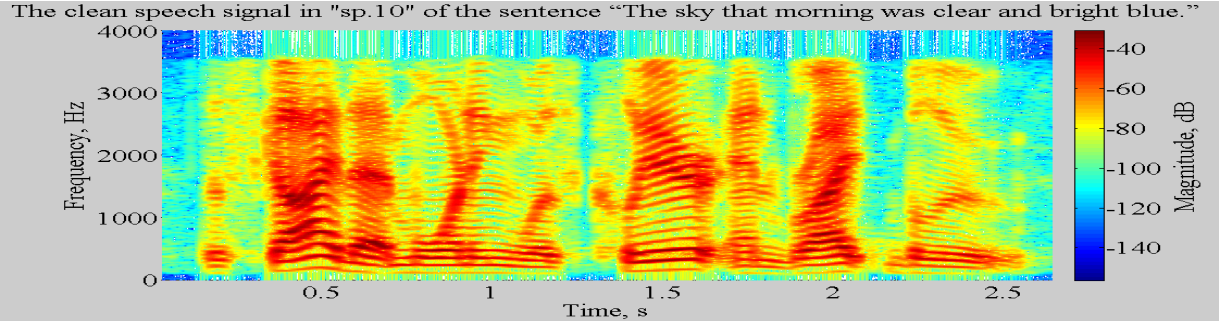


**Figure 4.11:** The spectrogram of the clean speech signal in "SP.10".

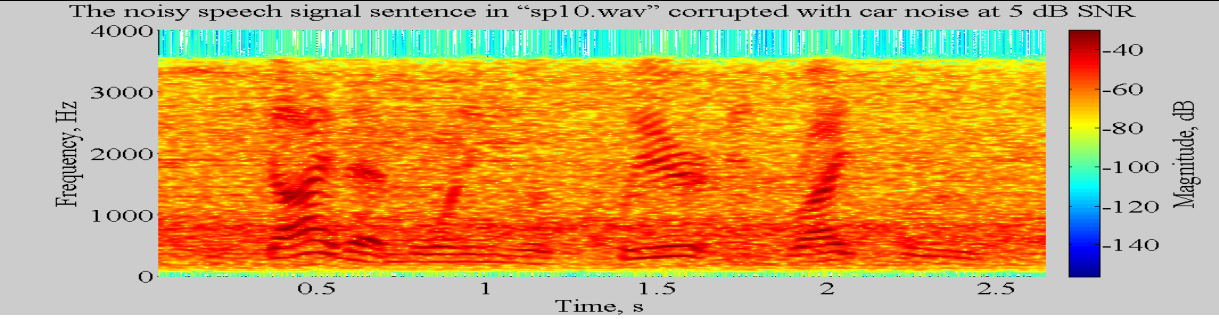


**Figure 4.12:** The spectrogram of the noisy signal in "SP.10" corrupted with car noise at 5 dB SNR.

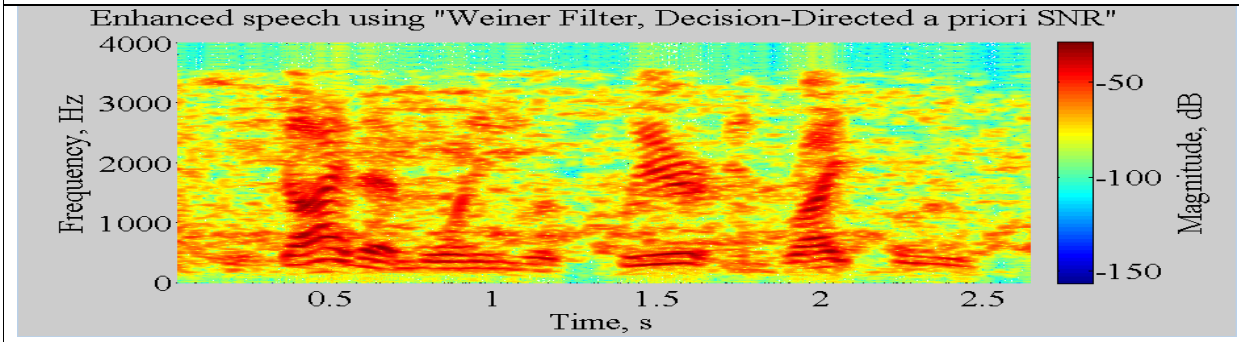


**Figure 4.13:** The spectrogram of the enhanced speech using "Weiner Filter, Decision-Directed a priori SNR".

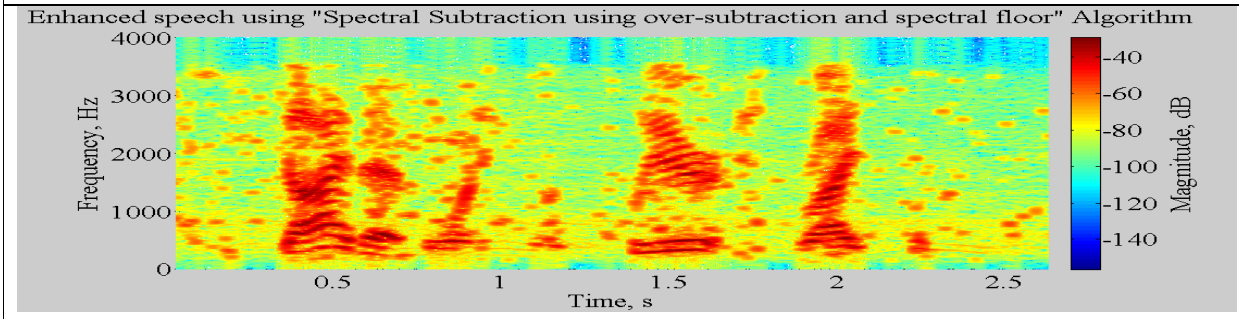


**Figure 4.14:** The spectrogram of the enhanced speech using Over-Subtraction and spectral floor" algorithm.

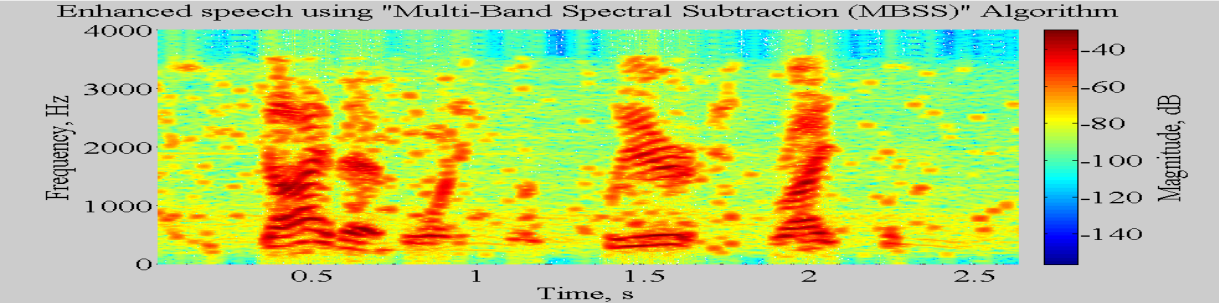


**Figure 4.15:** The spectrogram of the enhanced speech using "Multi-Band Spectral Subtraction (MBSS)" Algorithm.

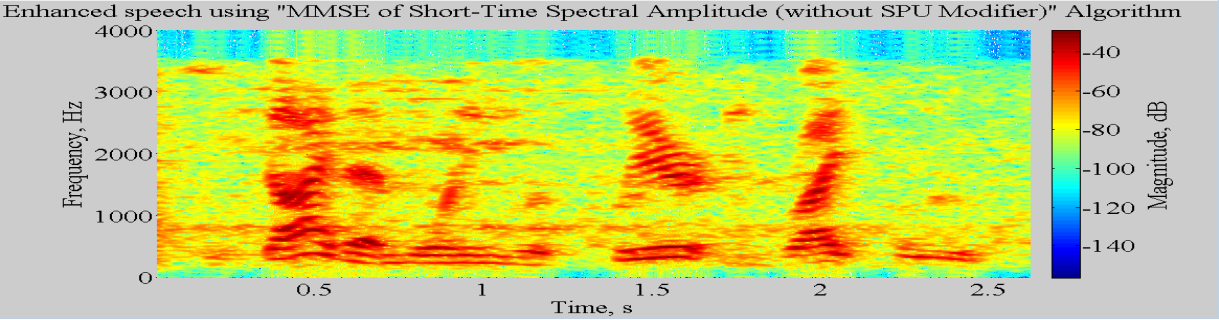


**Figure 4.16:** The spectrogram of the enhanced speech using "MMSE-STSA (without using SPU modifier)" Algorithm.

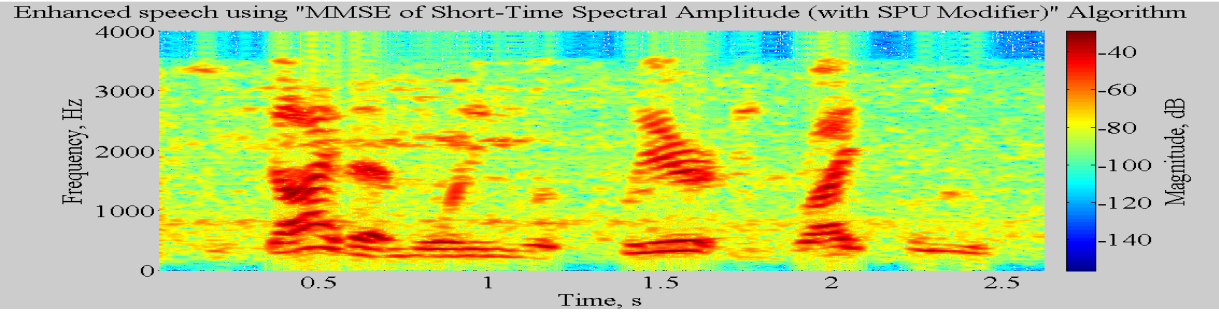


**Figure 4.17:** The spectrogram of the enhanced speech using "MMSE-STSA (using SPU modifier)" Algorithm.

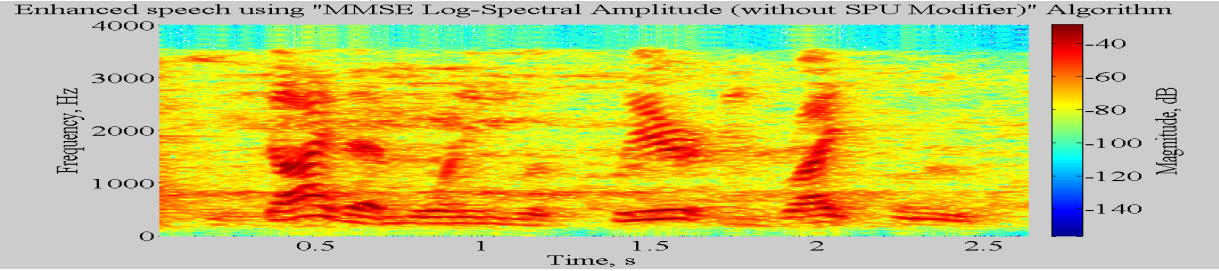


**Figure 4.18:** The spectrogram of the enhanced speech using "MMSE-LSA (without using SPU modifier)" Algorithm.

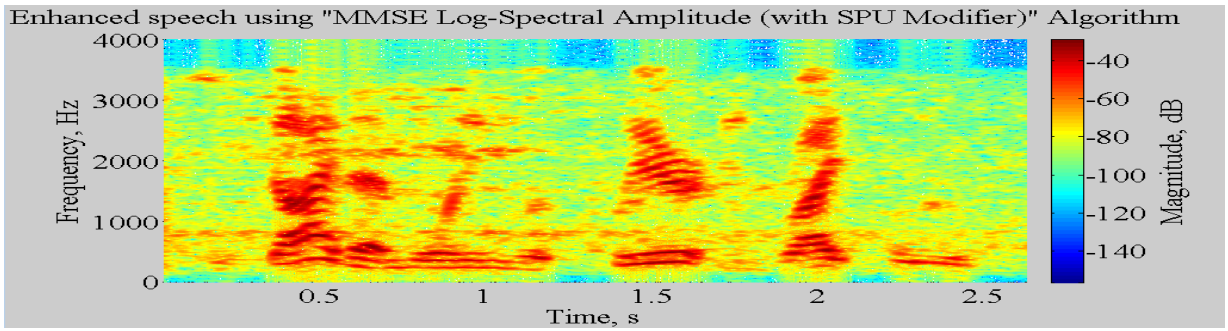


**Figure 4.19:** The spectrogram of the enhanced speech using "MMSE-LSA (using SPU modifier)" Algorithm.

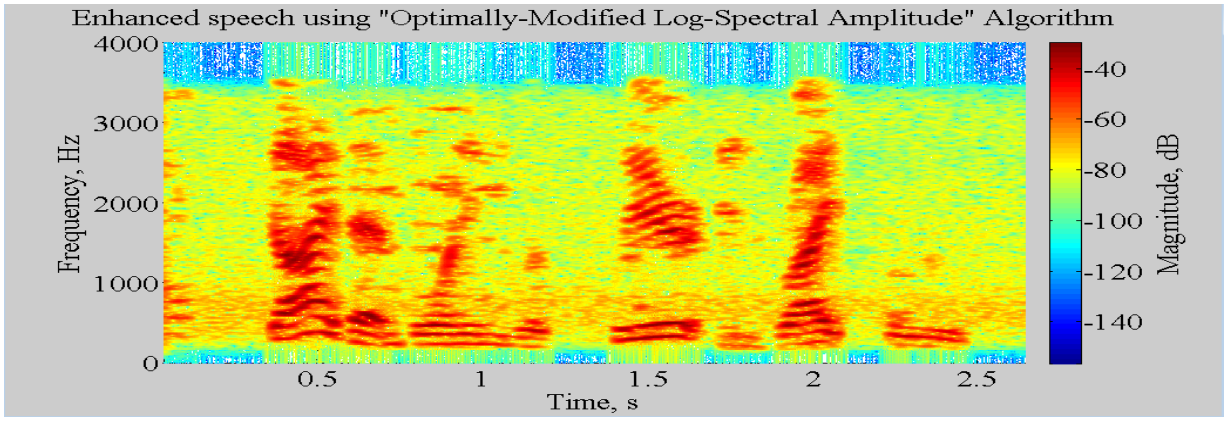


**Figure 4.120:** The spectrogram of the enhanced speech using "OM-LSA" Algorithm.

From the visual examinations of the spectrograms in figures presented above, we can remark that:

- In all the enhanced speech spectrograms, the formants are much clearer and visible than in the noisy speech spectrogram, which indicates that there is a considerable amount of noise has been reduced from the noisy speech.
- The enhanced speech spectrograms using Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method contain some random isolated dots which cause audible artifact known as musical noise.
- The enhanced speech spectrograms using MMSE-STSA, and MMSE-LSA algorithms show better results, and less amount of isolated dots compared with the spectrograms of Weiner filter, MBSS method, and Spectral Subtraction (using over-subtraction and spectral floor) method.

- Applying the multiplicative SPU modifier to the MMSE-STSA, and MMSE-LSA algorithms provides more noise reduction in spectrograms.
- The enhanced speech spectrogram using OM-LSA algorithm is the nearest to the to the original clean speech spectrogram.

### 4.1.4 Objective measures for implemented algorithms performance evaluation

Objective measures are based on a mathematical comparison of the original and enhanced speech signals.

**Signal-to-Noise Ratio (SNR)**

As the name suggests, SNR is the ratio of the signal energy to the noise energy:

$$SNR_{dB} = 10.log_{10}(\frac{\sum_n s^2[n]}{\sum_n (s[n]-\hat{s}[n])^2}) \quad (4.1)$$

Where $s(n)$ is the clean signal and $\hat{s}(n)$ is the processed signal. If the summation is performed over the whole signal length, the operation is called global SNR.

**Segmental Signal-to-Noise Ratio ( $\mathbf{SNR_{Seg}}$):**

The SNRseg in dB is the average SNR computed over short frames of the speech signal. The SNRseg over M frames of length N is computed as:

$$\text{SNR}_{Seg} = \frac{1}{M}\sum_{i=0}^{M-1} 10log_{10}\left[\frac{\sum_{n=iN}^{iN+N-1} s^2[n]}{\sum_{n=iN}^{iN+N-1}(s[n]-\hat{s}[n])^2}\right] dB, \quad (4.2)$$

In order to perform our objective tests, each algorithm is evaluated using all the sentences from NOIZEUS data base corrupted by 4 different SNR values (0, 5,10 and 15dB) in 6 colored noise environments which are as follows:

- Train
- Car
- Street
- Restaurant
- Train station
- Babble

In addition to that, a synthesized white noise added to clean speech sentences of NOIZEUS database at SNR range 0-15dB is also used to test the algorithms.

The results (all the obtained SNR and $\text{SNR}_{Seg}$ values are averages of 30 measures the number of sentences in the database and are given in dB**)** are shown in tables from 4.2 to 4.7.

| **Train noise** | SNR=0 dB $\mathbf{SNR_{Seg}}$ ≅-4.50 | | 5 dB $\mathbf{SNR_{Seg}}$ ≅-1.67 | | 10 dB $\mathbf{SNR_{Seg}}$ ≅1.50 | | 15 dB $\mathbf{SNR_{Seg}}$ ≅4.50 | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ |
| **Weiner DD** | 6.05 | ***-0.20*** | 8.77 | ***1.51*** | 13.01 | ***4.19*** | 16.16 | ***8.00*** |
| **SS using using over-subtraction and spectral floor** | 6.14 | ***-0.14*** | 8.07 | ***1.51*** | 12 | ***4.26*** | 17.44 | ***8.13*** |
| **Mband** | 6.40 | ***-0.10*** | 8.56 | ***2.32*** | 12.07 | ***5.12*** | 17.44 | ***8.54*** |
| **MMSE-STSA** | 6.48 | ***0.50*** | 8.32 | ***2.40*** | 12.18 | ***5.76*** | 15.10 | ***8.64*** |
| **MMSE-LSA** | 6.23 | ***0.55*** | 8.44 | ***2.44*** | 12.22 | ***5.77*** | 15.03 | ***8.68*** |
| **MMSE-STSA using SPU modifier** | 6.10 | ***1.20*** | 8.00 | ***3.11*** | 11.81 | ***5.80*** | 15.32 | ***8.70*** |
| **MMSE-LSA using SPU modofier** | 6.31 | ***1.25*** | 8.24 | ***3.10*** | 12.09 | ***5.81*** | 15.53 | ***8.73*** |
| **OM-LSA** | 6.40 | ***1.50*** | 8.76 | ***3.20*** | 13.66 | ***6.00*** | 17.20 | ***8.90*** |

Table 4.2: Objective quality evaluation with train noise

| **Car noise** | SNR=0dB $\mathbf{SNR_{Seg}} \cong -4.95$ | | 5 dB $\mathbf{SNR_{Seg}} \cong \mathbf{-2.00}$ | | 10 dB $\mathbf{SNR_{Seg}} \cong \mathbf{1.05}$ | | 15 dB $\mathbf{SNR_{Seg}} \cong \mathbf{4.05}$ | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ |
| **Weiner DD** | 6.08 | ***-0.3*** | 10.17 | ***2.32*** | 13.83 | ***5.19*** | 17.37 | ***8.50*** |
| **SS using over-subtraction and spectral floor** | 4.75 | ***-0.4*** | 9.46 | ***1.83*** | 13.47 | ***4.90*** | 18.75 | ***9.00*** |
| **Mband** | 4.90 | ***-0.2*** | 9.40 | ***2.33*** | 13.00 | ***5.20*** | 17.30 | ***9.05*** |
| **MMSE-STSA** | 5.16 | ***0.50*** | 9.53 | ***2.80*** | 13.46 | ***5.30*** | 16.51 | ***9.10*** |
| **MMSE-LSA** | 5.34 | ***0.55*** | 9.38 | ***2.80*** | 13.23 | ***5.33*** | 16.54 | ***9.10*** |
| **MMSE-STSA using SPU modifier** | 4.67 | ***0.80*** | 9.21 | ***3.24*** | 13.28 | ***5.41*** | 16.15 | ***9.12*** |
| **MMSE-LSA using SPU modofier** | 6.05 | ***0.90*** | 9.49 | ***3.25*** | 13.42 | ***5.43*** | 15.01 | ***9.15*** |
| **OM-LSA** | 6.10 | ***1.30*** | 10.48 | ***3.50*** | 14.43 | ***5.55*** | 18.38 | ***9.20*** |

Table 4.3: Objective quality evaluation with car noise

| **Street noise** | SNR=0 dB $\mathbf{SNR_{Seg}} \cong -4.25$ | | 5 dB $\mathbf{SNR_{Seg}} \cong \mathbf{-1.20}$ | | 10 dB $\mathbf{SNR_{Seg}} \cong \mathbf{1.75}$ | | 15 dB $\mathbf{SNR_{Seg}} \cong 4.70$ | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ |
| **Weiner DD** | 4.00 | **-0.15** | 7.74 | **1.85** | 11.81 | **3.94** | 16.10 | **8.13** |
| **SS using using over-subtraction and spectral floor** | 4.12 | **-0.2** | 8.35 | **1.86** | 12.21 | **4.70** | 16.01 | **8.94** |
| **Mband** | 4.50 | **0.05** | 7.83 | **2.00** | 11.80 | **4.90** | 16.01 | **9.00** |
| **MMSE-STSA** | 5.02 | **0.40** | 7.96 | **2.30** | 12.17 | **5.00** | 15.02 | **9.01** |
| **MMSE-LSA** | 5.05 | **0.44** | 7.67 | **2.49** | 12.00 | **5.12** | 15.01 | **9.04** |
| **MMSE-STSA using SPU modifier** | 5.10 | **0.70** | 7.83 | **2.90** | 12.18 | **5.30** | 15.06 | **9.05** |
| **MMSE-LSA using SPU modofier** | 5.20 | **0.90** | 8.00 | **3.05** | 12.18 | **5.37** | 15.17 | **9.10** |
| **OM-LSA** | 5.57 | **1.26** | 8.72 | **3.40** | 13.00 | **5.50** | 16.24 | **9.12** |

Table 4.4: Objective quality evaluation with Street noise

| **Restaurant noise** | SNR=0 dB $\mathbf{SNR_{Seg}} \cong \mathbf{-4.18}$ | | 5 dB $\mathbf{SNR_{Seg}} \cong \mathbf{-1.15}$ | | 10 dB $\mathbf{SNR_{Seg}} \cong \mathbf{1.80}$ | | 15 dB $\mathbf{SNR_{Seg}} \cong \mathbf{4.80}$ | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ | SNR | $\mathbf{SNR_{Seg}}$ |
| **Weiner DD** | 4.00 | **-0.02** | 7.71 | **1.62** | 11.51 | **4.21** | 16.10 | **8.49** |
| **SS using using over-subtraction and spectral floor** | 4.02 | **-0.10** | 7.68 | **1.60** | 11.52 | **4.20** | 17.31 | **8.50** |
| **Mband** | 4.45 | **0.04** | 7.70 | **1.80** | 12.00 | **4.39** | 17.42 | **8.66** |
| **MMSE-STSA** | 5.05 | **0.10** | 8.28 | **1.90** | 11.90 | **4.80** | 15.48 | **8.68** |
| **MMSE-LSA** | 5.60 | **0.40** | 8.22 | **1.95** | 12.05 | **5.02** | 15.30 | **8.71** |
| **MMSE-STSA using SPU modifier** | 6.00 | **0.52** | 8.20 | **2.40** | 12.05 | **5.02** | 15.44 | **8.72** |
| **MMSE-LSA using SPU modofier** | 6.05 | **0.80** | 8.34 | **2.60** | 12.10 | **5.10** | 15.60 | **8.90** |
| **OM-LSA** | 6.12 | **1.22** | 8.56 | **3.12** | 12.40 | **5.30** | 17.43 | **9.30** |

Table 4.5: Objective quality evaluation with restaurant noise

| **Train station noise** | SNR=0 dB **$SNR_{Seg}$**=-4.7 dB | | 5 dB **$SNR_{Seg}$**=-1.95 dB | | 10 dB **$SNR_{Seg}$**= dB | | 15dB **$SNR_{Seg}$**=4.60dB | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** |
| **Weiner DD** | 4.98 | **0.30** | 9.06 | **2.01** | 12.50 | **4.31** | 15.23 | **8.60** |
| **SS using using over-subtraction and spectral floor** | 5.88 | **0.44** | 8.89 | **2.04** | 13.00 | **4.20** | 16.38 | **8.94** |
| **Mband** | 5.80 | **0.50** | 8.73 | **2.11** | 12.89 | **4.50** | 16.50 | **8.80** |
| **MMSE-STSA** | 5.86 | **0.61** | 8.92 | **2.42** | 13.02 | **4.65** | 15.58 | **8.95** |
| **MMSE-LSA** | 5.80 | **0.66** | 8.77 | **2.70** | 13.05 | **4.80** | 15.70 | **9.05** |
| **MMSE-STSA using SPU modifier** | 6.05 | **0.80** | 8.56 | **2.95** | 13.00 | **4.91** | 15.77 | **9.10** |
| **MMSE-LSA using SPU modofier** | 6.14 | **1.10** | 8.84 | **3.02** | 13.06 | **5.10** | 15.51 | **9.22** |
| **OM-LSA** | 6.20 | **1.66** | 9.84 | **3.50** | 13.5 | **5.70** | 16.54 | **9.61** |

Table 4.6: Objective quality evaluation with train station noise

| **Babble noise** | SNR=0 dB **$SNR_{Seg} \cong -4.48$** | | 5 dB **$SNR_{Seg} \cong -1.50$** | | 10 dB **$SNR_{Seg} \cong 1.48$** | | 15 dB **$SNR_{Seg}$** ≅4.40 | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** |
| **Weiner DD** | 4.00 | **-0.10** | 8.26 | **1.55** | 12.00 | **4.22** | 15.26 | **8.40** |
| **SS using using over-subtraction and spectral floor** | 4.69 | **-0.20** | 9.15 | **1.60** | 13.33 | **4.60** | 17.15 | **8.61** |
| **Mband** | 4.90 | **-0.05** | 8.90 | **2.00** | 13.20 | **4.80** | 17.50 | **8.70** |
| **MMSE-STSA** | 4.92 | **0.10** | 8.26 | **2.50** | 12.00 | **5.01** | 15.26 | **8.95** |
| **MMSE-LSA** | 4.90 | **0.25** | 7.90 | **2.68** | 12.05 | **5.10** | 15.31 | **9.05** |
| **MMSE-STSA using SPU modifier** | 5.50 | **0.30** | 8.56 | **2.80** | 12.03 | **5.33** | 15.20 | **9.11** |
| **MMSE-LSA using SPU modofier** | 5.40 | **0.60** | 8.44 | **3.01** | 12.07 | **5.41** | 15.02 | **9.20** |
| **OM-LSA** | 5.60 | **1.15** | 9.30 | **3.19** | 13.00 | **5.60** | 16.96 | **9.55** |

Table 4.7: Objective quality evaluation with babble noise

| **white noise** | SNR=0 dB **$SNR_{Seg}$** ≅-4.50 | | 5 dB **$SNR_{Seg}$** ≅-1.3 | | 10 dB **$SNR_{Seg}$** ≅2.03 | | 15 dB **$SNR_{Seg}$** ≅4.2 | |
|---|---|---|---|---|---|---|---|---|
| **Objective test** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** | SNR | **$SNR_{Seg}$** |
| **Weiner DD** | 6.65 | **1.02** | 10.32 | **3.00** | 14.00 | **5.34** | 16.94 | **8.00** |
| **SS using using over-subtraction and spectral floor** | 5.99 | **1.05** | 9.71 | **3.01** | 13.23 | **5.31** | 18.32 | **9.23** |
| **Mband** | 6.60 | **1.10** | 10.20 | **3.20** | 13.55 | **5.34** | 18.00 | **9.25** |
| **MMSE-STSA** | 7.12 | **1.50** | 10.46 | **4.00** | 13.65 | **6.40** | 16.50 | **8.61** |
| **MMSE-LSA** | 6.86 | **1.71** | 10.20 | **4.15** | 13.54 | **6.55** | 16.35 | **8.72** |
| **MMSE-STSA using SPU modifier** | 6.68 | **1.80** | 10.10 | **4.22** | 13.24 | **6.61** | 16.14 | **8.88** |
| **MMSE-LSA using SPU modofier** | 7.01 | **1.88** | 10.34 | **4.30** | 13.47 | **6.89** | 16.22 | **9.01** |
| **OM-LSA** | 7.73 | **2.00** | 10.89 | **4.42** | 14.16 | **7.00** | 18.00 | **9.80** |

Table 4.8: Objective quality evaluation with white noise

**A**ccording to the objective test results presented above, we can observe the following:

- There are remarkable improvements in both global and segmental SNRs of noisy speech signals after being processed by the implemented DFT-based speech enhancement algorithms and the noise reduction is more when it is white.
- The speech enhancement using Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method provides less segmental SNR values when compared to the other implemented algorithms in most cases.
- The speech enhancement using MMSE-STSA, and MMSE-LSA algorithms provides more better segmental SNR values, and using the SPU modifier gives a remarkable improvement in segmental SNRs.
- The speech enhancement using Optimally Modified Log-Spectral Amplitude estimator (OM-LSA) provides the best results (global SNR, and Segmental SNR) in most cases.

### 4.1.5 Subjective tests for Algorithm performance evaluation

Subjective tests rely heavily on the opinion of a group of listeners to judge the quality or intelligibility of processed speech. These tests are often time consuming as they require proper

training of listeners. In addition to this, a constant listening environment (e.g., playback volume), identically tuned output device (e.g., headphones and/or speakers) are necessary. Nevertheless, subjective test results present the most accurate system of performance, insofar as intelligibility and speech quality are concerned, as they are determined perceptually by the human auditory system. These tests can be structured under two types of evaluation procedures: speech quality evaluation and also intelligibility testing. Quality refers to the clarity, freedom of distortion and ease for listening whereas Intelligibility refers to the number of words that can be identified correctly by a listener or to the likelihood of being correctly understood.

#### 4.1.5.1 Subjective test for speech quality evaluation

In this test, we asked 5 normal-hearing students who speak, and understand English very well to listen twice to the different samples of speech for each input SNR used in the previous objective tests. In the first time, we presented to them the signals in their noisy form, whereas in the second time we presented to them the processed ones. After that, we asked our listeners to grade each speech heard on a scale from 1 to 5, based on how pleasant their listening experience was, the highest grade corresponding to the most pleasant one. Then we averaged the respective grades and results are given in table 4.9.

| **Input SNR (dB)** | **0** | **5** | **10** | **15** |
|---|---|---|---|---|
| **Noisy speech grade** | 0.5 | 1.0 | 1.7 | 2.2 |
| **Weiner DD** | 1.9 | 2.1 | 2.6 | 2.8 |
| **SS using using over-subtraction and spectral floor** | **2.0** | **2.5** | **2.6** | **2.9** |
| **Mband** | **2.4** | **2.6** | **2.8** | **3.2** |
| **MMSE-STSA** | **2.5** | **2.7** | **3.0** | **3.1** |
| **MMSE-LSA** | **2.6** | **2.9** | **3.3** | **3.6** |
| **MMSE-STSA using SPU modifier** | **2.8** | **3.1** | **3.4** | **3.9** |
| **MMSE-LSA using SPU modofier** | **3.0** | **3.3** | **3.8** | **4.0** |
| **OM-LSA** | **3.4** | **3.8** | **4.3** | **4.5** |

Table 4.9: subjective test for speech quality evaluation

According to this quality subjective test, we can say:

- The quality of the noisy speech samples has been considerably improved after the enhancement by the implemented algorithms.
- The speech enhancement using Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method obtained the smallest grades for speech quality evaluation which confirm the annoying musical noise shown during the spectrograms visual examinations (random isolated dots).
- The speech enhancement using the OM-LSA algorithms provides the best speech quality.

#### 4.1.5.2 Subjective test for speech intelligibility evaluation

In this test, we asked our listeners to give the percentage of intelligibility (according to the number of words that can be identified correctly by a listener) in the same speech signals. The results are shown below in table 4.10.

| **Input SNR (dB)** | **0** | **5** | **10** | **15** |
|---|---|---|---|---|
| **Noisy speech percentage** | 10% | 30% | 47% | 58% |
| **Weiner DD** | 53% | 56% | 66% | 69% |
| **SS using using over-subtraction and spectral floor** | 53% | 60% | 67% | 72% |
| **Mband** | 56% | 67% | 70% | 75% |
| **MMSE-STSA** | 65% | 71% | 74% | 82% |
| **MMSE-LSA** | 67% | 73% | 75% | 85% |
| **MMSE-STSA using SPU modifier** | 68% | 73% | 76% | 88% |
| **MMSE-LSA using SPU modofier** | 71% | 74% | 77% | 90% |
| **OM-LSA** | 73% | 75% | 80% | 92% |

Table 4.10: subjective test for speech intelligibility evaluation

According to table 4.10 we can say:

- The implemented algorithms have considerably improved the intelligibility of the noisy speech signals.

- Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method provide the smallest percentages of intelligibility in comparison to the other implemented algorithms and that's due to the amount of distortions caused them.
- The optimally modified Log-Spectral Amplitude estimator (OM-LSA) algorithm shows the highest percentages of intelligibility.

### 4.1.6 Comments

The implemented algorithms performance evaluation based on visual examinations, objective and subjective tests show that the optimally modified Log-Spectral Amplitude estimator (OM-LSA) algorithm outperforms all the implemented algorithms (low signal distortion and the best amount of noise reduction). However, we would like to note the following:

- The global SNR is a poor estimator of subjective quality. A high SNR value, is thus, not necessarily indicative of good perceptual quality of the speech.
- The segmental SNR objective test is more related to the subjective tests.
- Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method show acceptable amounts of non-stationary noise reduction but produce some distortions in the shape of the enhanced speech signals.
- MMSE-STSA, and MMSE-LSA algorithms provide more non-stationary noise reduction and less distortions.
- Applying the SPU multiplicative modifier with MMSE-STSA, and MMSE-LSA algorithms increases the quality and the intelligibility of the enhanced speech signals.

## 4.2 Blind Multi-speaker speech signals separation

### 4.2.1 Implementation and performance evaluation of the DUET algorithm

The algorithm works by taking a two-channel wav-file, with one mixture per channel and also using a Hamming window of 1024 samples length and 512 overlapping samples.

In order to investigate the applicability of the DUET, we have recorded five different speakers (male and female) in three languages Arabic, English, and French with sampling frequency 11025 Hz. After that we have formed artificial instantaneous speech mixtures composed of different number of speech signals (from two to five mixed signals). Two methods for DUET algorithm performance evaluation have been used: the first is based on listening tests and the second on distortion measures proposed by [55].

#### 4.2.1.1 Listening tests

In this test, we asked 10 people to listen twice to different artificial instantaneous speech mixtures. In the first time, we presented to them signals in their mixed form, whereas in the second time we presented to them the processed separated speech signals. After that, we asked our listeners to grade each signal heard on a scale from 1 to 5, based on how pleasant their listening experience was, the highest grade corresponding to the most pleasant one. Then we averaged the respective grades. The results are tabulated in table 4.11.

| Test Average grade<br>Mixture of speech signals | Before separation | After separation |
|---|---|---|
| Two speech signals mixture | **2.0** | **4.7** |
| Three speech signals mixture | **0.5** | **4.6** |
| Four speech signals mixture | **0.0** | **4.3** |
| Five speech signals mixture | **0.0** | **4.0** |

Table 4.11: listening test results for DUET algorithm performance evaluation

According to this test, we can say that, the quality and intelligibility of speech signals have been considerably improved after the separation using DUET algorithm.

**4.2.1.2 Objective tests for DUET algorithm performance evaluation**

To further validate the performance of the DUET algorithm a method based on distortion measures proposed in [55], is used.

The principle of the performance measures described is to decompose a given estimate $\hat{s}(t)$ of a source $s_i(t)$ as a sum:

$$\hat{s}(t) = S_{target}(t) + e_{interf}(t) + e_{artif}(t) \quad (4.2)$$

Where $S_{target}(t)$ is an allowed deformation of the target source $s_i(t)$, $e_{interf}(t)$ is an allowed deformation of the sources which accounts for the interferences of the unwanted sources, and $e_{artif}(t)$ is an "artifact" term that may correspond to artifacts of the separation algorithm such as musical noise, etc. or simply to deformations induced by the separation algorithm that are not allowed.

The distortion measures take into account: Source to Distortion Ratio (SDR), Source to Artifacts Ratio (SAR), and Source to Interference Ration (SIR). The measures can be expressed by the following equations.

$$SDR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}+e_{artif}\|^2}, \quad (4.3)$$

$$SAR = 10log_{10}\frac{\|s_{target}+e_{interf}\|^2}{\|e_{artif}\|^2}, \quad (4.4)$$

$$SIR = 10log_{10}\frac{\|s_{target}\|^2}{\|e_{interf}\|^2}. \quad (4.5)$$

The distortion measures were calculated using the same artificial instantaneous speech signals mixtures used in the listening test. The obtained results for every Blind speech signal separation are shown in tables from 4.12 to 4.15.

The time-domain separated speech signals and the weighted 2-D smoothed weighted histograms obtained from the different speech signals mixtures are shown in figures from 4.21 to 4.28.

| Two speech signals mixture | | | |
|---|---|---|---|
| Separated speech signals<br>Distortion Measures | **Speech signal 1** | **Speech signal 2** | **The average** |
| SDR (dB) | 9.77 | 9.42 | 9.60 |
| SIR (dB) | 22.15 | 41.92 | **32.04** |
| SAR (dB) | 10.06 | 9.42 | 9.74 |

Table 4.12: Distortion measures for the blind separation of two speech signals mixture.

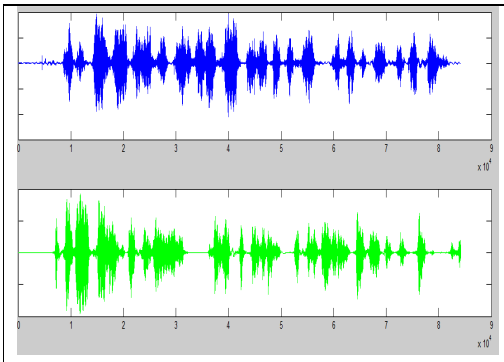

**Figure 4.21:** The two separated speech signals in time-domain.

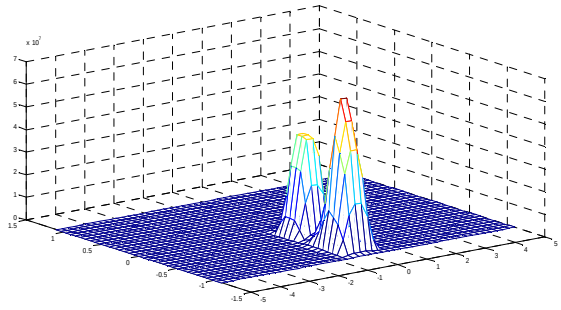

**Figure 4.22:** The two-D smoothed weighted histogram obtained from the 2 mixtures of two speech signals.

| Three speech signals mixture | | | | |
|---|---|---|---|---|
| Separated speech signals<br>Distortion Measures | **Speech signal 1** | **Speech signal 2** | **Speech signal 3** | **The average** |
| SDR (dB) | 6.91 | 3.72 | 5.87 | 5.50 |
| SIR (dB) | 26.43 | 12.06 | 25.57 | **21.34** |
| SAR (dB) | 6.97 | 4.67 | 5.93 | 5.86 |

Table 4.13: Distortion measures for the blind separation of three speech signals mixture.

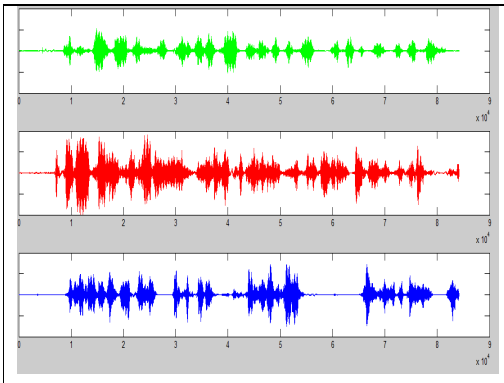

**Figure 4.23:** The three separated speech signals in time-domain.

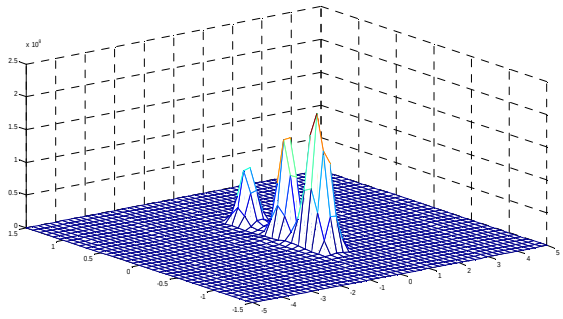

**Figure 4.24:** The two-D smoothed weighted histogram obtained from the 2 mixtures of three speech signals.

| Four speech signals mixture | | | | | |
|---|---|---|---|---|---|
| Separated speech signals<br>Distortion Measures | **Speech signal 1** | **Speech signal 2** | **Speech signal 3** | **Speech signal 4** | **The average** |
| SDR (dB) | 5.83 | 1.91 | 1.65 | 5.56 | 3.74 |
| SIR (dB) | 22.14 | 10.60 | 12.22 | 22.08 | **16.76** |
| SAR (dB) | 5.56 | 2.83 | 2.30 | 5.69 | 4.10 |

Table 4.14: Distortion measures for the blind separation of four speech signals mixture.

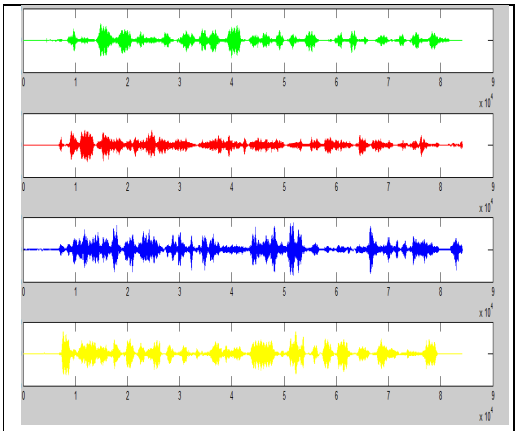

**Figure 4.25:** The four separated speech signals in time-domain.

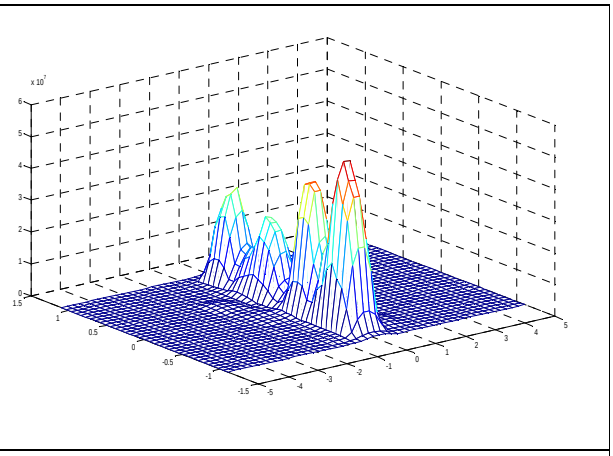

**Figure 4.26:** The two-D smoothed weighted histogram obtained from the 2 mixtures of four speech signals.

| Five speech signals mixture | | | | | | |
|---|---|---|---|---|---|---|
| Separated speech signals<br>Distortion Measures | **speech signal 1** | **speech signal2** | **speech signal 3** | **speech signal 4** | **speech signal 5** | **The average** |
| SDR (dB) | 3.24 | 1.0 | 1.0 | 1.50 | 4.01 | 2.15 |
| SIR (dB) | 15.80 | 8.10 | 8.72 | 11.65 | 20.22 | **16.76** |
| SAR (dB) | 3.61 | 1.0 | 1.0 | 2.23 | 4.15 | 2.40 |

Table 4.15: Distortion measures for the blind separation of five speech signals mixture.

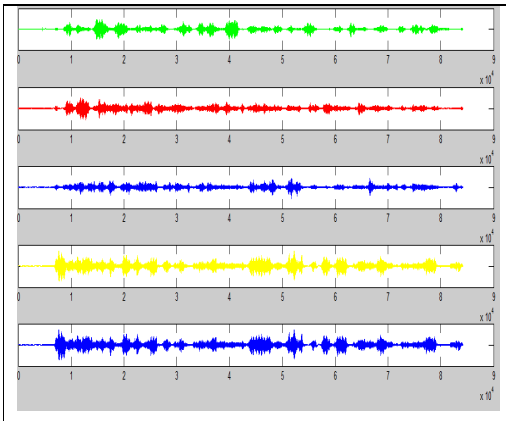

**Figure 4.27:** The five separated speech signals in time-domain.

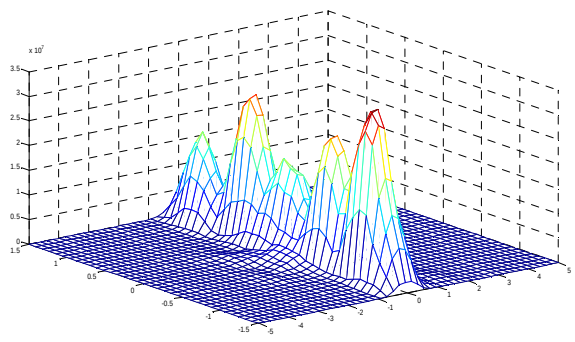

**Figure 4.28:** The two-D smoothed weighted histogram obtained from the 2 mixtures of five speech signals.

**Comments**

- The number of peaks in the two-D smoothed weighted histograms represents the number of speech signal sources to be separated.
- The separation using DUET algorithm provides a high Signal to interference ratios (SIR) and acceptable SDR, SAR ratios for all separated speech signals even when the number of signals is greater than 2 (the number of speech sensors or mixtures).
- The more we add a speech signal to the mixtures, the more SIR, SDR, and SAR values decrease.

The performance evaluation based on objective measures, observation of the waveforms, as well as subjective listening tests clearly show that the Degenerate Unmixing and Estimation Technique (DUET) for the Blind speech signal separation (artificial instantaneous mixtures) provides a good fidelity of recovering speech signals with low distortions.

# Conclusion and Future Work

# CONCLUSION AND FUTURE WORK

The work in this project addressed both of the problems: single-channel speech enhancement at the presence of highly non-stationary background noise and the blind multi-speaker speech separation of an arbitrary number of speakers given just two anechoic mixtures, as pre-processing stage for various speech applications.

A set of six DFT-based single-channel speech enhancement algorithms have been implemented using highly non-stationary noise estimator, and each implemented algorithm has been evaluated using the NOIZEUS data base corrupted by 4 different SNR values (0, 5,10 and 15dB) in 6 colored noise environments (train, car, street, restaurant, train station, and babble) and a synthesized white noise.

The performance evaluation results establish the superiority of the Optimally-Modified Log-Spectral Amplitude estimator (OM-LSA) algorithm over all the implemented DFT-based single-channel speech enhancement algorithms with respect to perceptible quality and intelligibility improvements of the enhanced speech signals. Therefore, OM-LSA can be considered as good pre-processing technique for single-channel speech applications. MMSE-STSA, MMSE-LSA (using SPU multiplicative modifier) algorithms provide acceptable levels of speech intelligibility and quality in most cases and the second one behaves a little bit better than MMSE-STSA especially in reducing the musical noise. Weiner filter, Spectral Subtraction (using over-subtraction and spectral floor) method, and MBSS method show more distortions in the shape of the enhanced signals at low SNRs (0-5dB) range in most cases.

For the blind multi-speaker speech separation of an arbitrary number of speakers the focus was on applying the degenerate Unmixing and Estimation Technique (DUET) that uses the two-dimensional smoothed weighted histogram to estimate the mixing parameters and the time-frequency masks to separate the speech sources. We have implemented and tested the behavior of the DUET technique on artificial instantaneous speech mixtures of different number of speakers and the obtained results demonstrate the powerfulness of DUET approach as an efficient tool of the blind separation of different numbers of speech sources given just two anechoic mixtures provided the assumption which we call Approximate W-disjoint Orthogonality.

# CONCLUSION AND FUTURE WORK

In addition to all the obtained results we may say that, the most suitable technique for speech enhancement is the one which provides robustness to environmental noise contributing factors and robustness to acoustical inputs.

In the future, we plan to study the real effects of the implemented pre-processing techniques on the various speech communication applications.

The works on implementing the DFT-based techniques for single-channel speech enhancement and DUET algorithm for Blind Source Separation as pre-processing stages for various speech applications should definitely continue considering the good results we managed to achieve. Here is a short list of items that we think could be subjected to further studies:

- Investigating the speech enhancement using Laplacian-based MMSE estimator of the magnitude spectrum rather than MMSE estimator, which is based on a Gaussian model.
- The error between the processed signal and the clean speech signal can be strongly minimized if the estimate of the noise spectrum is more accurate. Hence, it is desirable to estimate the noise signal at every available instant to get a more accurate estimate of the noise spectrum.
- Implementing the DUET algorithm to separate the speech convolutive mixtures.
- Implementing the DUET algorithm to separate the audio stream mixtures.
- Investigating the combination of a DUET algorithm, with an algorithm operating in the time domain to check if the results can be improved.
- Experimenting DUET algorithm with moving sources because the speakers are not always at fixed locations.
- Someone requiring hearing aid equipped with BSS technology should be able to move around. Hence experimenting the DUET algorithm with moving microphones will be a good contribution.
- Single channel blind source separation is a challenging task. Hence working on this provides a good contribution to speech signals pre-processing techniques.

# References

# References

[1] Rangachari, S. and Loizou, P. (2006). A noise estimation algorithm for highly non-stationary environments. Speech Communication, 28, 220-231.

[2] S.China Venkateswarlu, Dr. K.Satya Prasad, Dr. A.SubbaRami Reddy., "Improve Speech Enhancement Using Weiner Filtering," Global Journals Inc. (USA), Volume 11 Issue 7 Version 1.0 May 2011.

[3] Ian McLoughlin, "Applied Speech and Audio Processing: With MATLAB Examples," Cambridge University Press, pages 41–42, 2009.

[4] Gong, Y., "Speech recognition in noisy environments: A survey", Speech Communication, Vol. 16, No.3, pp 261-291, April 1995.

[5] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. J. Acoustical Soc. America, 87(4): 1738–1752, 1990.

[6] Y. M. Cheng and D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. Proc. Int. Conf. on Acoustics, Speech and Signal Processing, Vol. 2 pages 961–963, 1991.

[7] Y. M. Cheng and D. O'Shaughnessy, "Speech enhancement based conceptually on auditory evidence," IEEE Trans. Signal Proc., 39(9): 1943–1954, 1991.

[8] Beranek, L. L. and Ver, I. L. Noise and Vibration Control Engineering, John Wiley & Sons Inc, 1992.

[9] Brown, S. "On-Site Power Generation," A Reference Book, Chapter 23 Sound Attenuation, Third Edition, Electrical Generating Systems Association, 2000.

[10] E. Zwicker and H. Fastl. Psychoacoustics Facts and Models. Springer, Munich Germany, 2nd updated edition, January 1999.

[11] Edward W. Kamen and Bonnie S. Heck. Fundamentals of Signals and Systems using the Web and MATLAB. Prentice-Hall, 2000.

[12] S. V. Vaseghi, Advanced Signal Processing and Digital Noise Reduction. John Wiley and Sons, Ltd, 3rd Edition Book, pages 24, 246,247, 298, 2006.

[13] Virag, N., "Speech enhancement based on masking properties of the human auditory system", Master thesis, Swiss Federal Institute of Technology, 1996.

[14] Mukul Bhatnagar, BE "A modified spectral subtraction method combined with perceptual weighting for speech enhancement", Master of science thesis, university of Texas, pages18-20, August 2002.

[15] Junqua, J.C., “The influence of acoustics on speech production: a noise induced stress phenomenon known as the Lombard reflex”, ESCA-NATO Workshop on Speech under Stress, pp 83-90, Lisbon, September 1995.

[16] Lim, J.S. and Oppenheim, A.V., “Enhancement and bandwidth compression of noisy speech”, Proc. IEEE, Vol. 67, No.12, pp. 1586-1604, December 1979.

[17] Erhan Deger, “Noise thresholding with empirical mode decomposition for low distortion speech enhancement,” University of Tokyo, Master’s thesis, pages 2-8, Feb 2008.

[18] Eric Plourde, “Bayesian short-time spectral amplitude estimators for single-channel speech enhancement”, PHD thesis, McGill University Montreal, Canada, pp 30, October 2009

[19] J. Deller Jr., J. Hansen and J. Proakis, “Discrete-Time Processing of Speech Signals”, NY: IEEE Press, 2000.

[20] Soon Ing Yann, “Transform based speech enhancement techniques”, PHD thesis, Nanyang Technological University, pp.9-22, 2003.

[21] D. O’Shaughnessy, “Speech Communications: Human and Machine,”. IEEE Press, 2nd ed., 2000.

[22] A. S. Spanias, “Speech coding: A tutorial review,” Proc. IEEE, vol. 82, pp. 1541–1582, Oct 1994.

[23] Paul Coffey, “Enhancement of Speech in Noisy Conditions,” B.E. Electronic Engineering Project Report March 2009.

[24] Furui, Sadaoki, “Digital speech processing, synthesis, and recognition,” 2nd ed., rev. and expanded. Marcel Dekker, Inc. pp. 57–59. 2001.

[25] Doblinger, G., "Computationally efficient speech enhancement by spectral minima tracking in subbands," in EUROSPEECH'95, Madrid, Spain, Sept. 18-21, 1995, pp. 1513-1516.

[26] Cohen,I.andBerdugo,B.,"Noiseestimationby minima controlled recursive averaging for robust speech enhancement," IEEE Signal Proc. Letters,vol. 9, no. 1, pp. 12-15, January 2002.

[27] Boll, S.F., “Suppression of acoustic noise in speech using spectral subtraction”, IEEE Trans. on Acoust., Speech, Signal Proc., Vol. ASSP-27, No.2, pp.113-120, April 1979.

[28] S. F. Boll, “A spectral subtraction algorithm for suppression of acoustic noise in speech,” in Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing, (Washington, DC), pp. 200–203, Apr. 1979.

[29] R. J. McAulay and M. L. Malpass, “Speech enhancement using a soft-decision noise suppression filter,” IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-28, pp. 137–145, Apr. 1980.

[30] Joachim Thiemann, “Acoustic Noise Suppression for SpeechSignals using Auditory Masking Effects” Department of Electrical & Computer Engineering McGill University Montreal, Canada July 2001, page 43.

[31] M. Berouti, R. Schwartz and J. Makhoul, “Enhancement of speech corrupted by acoustic noise,” Proc. IEEE Int. Conf. on Acoust., Speech, Signal Procs., pp. 208- 211, Apr. 1979.

[32] Kamath, S. and Loizou, P. (2002). A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. Proc. IEEE Int. Conf. Acoust.,Speech, Signal Processing.

[33] Scalart, P. and Filho, J. (1996). Speech enhancement based on a priori signal to noise estimation. Proc. IEEE Int. Conf. Acoust. , Speech, Signal Processing, 629-632.

[34] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-32, pp. 1109–1121, Dec. 1984.

[35] D. Middleton and R. Esposito, “Simultaneous optimum detection and estimation of signals in noise,” IEEE Trans. Inform. Theory, vol. IT-14, pp. 434–444, May 1968 H. L.

[36] Van Trees, Detection, “Estimation and Modulation Theory,” New York: John Wiley & Sons, Inc., 1968.

[37] Y. Ephraim and D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” IEEE Trans. Acoustics, Speech, Signal Processing, vol. ASSP-33, No 2, pp. 443–445, April. 1985.

[38] I. Cohen “Optimal Speech Enhancement Under Signal Presence Uncertainty Using Log-Spectral Amplitude Estimator,” Lamar Signal Processing Ltd.2003.

[39] I. Y. Soon, S. N. Koh and C. K. Yeo, “Improved Noise Suppression Filter Using Self Adaptive Estimator of Probability of Speech Absence,” Signal Processing, vol. 75, pp. 151–159, 1999.

[40] R. Martin, I. Wittke and P. Jax, “Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech,”in Proc. Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP-2000, pp. 1479–1482.

[41] I. Cohen and B. Berdugo, “Speech Enhancement for Non-Stationary Noise Environments,” to appear in Signal Processing.

[42] Adel Belouchrani and Moeness G. Amin. “Blind Source Separation Based on Time-Frequency Signal Representations” IEEE transactions on signal processing, vol. 46, no. 11, november 1998

[43] CARDOSO, J.-F. Blind signal separation: statistical principles. Proceedings of the IEEE, 1998, vol. 86, no. 10, pp. 2009–2025.

[44] X. R. Cao and R. W. Liu, “General approach to blind source separation,” IEEE Trans. Signal Processing, vol. 44, no. 3, pp. 562–571, 1996.

[45] O.Yilmaz and S.Rickard, “Blind separation of speech mixtures via time-frequency masking” IEEE transactions on signal processing, Vol 52, no.7. pp.1830-1847, 2004.

[46] http://www.cs.northwestern.edu/~zra446/research.html#DUET_using_CQT_

[47] S. Rickard, “The duet blind source separation algorithm,” in Blind Speech Separation, S. Makino, T.W. Lee, and H. Sawada, Eds. Dor-drecht, The Netherlands: Springer, 2007, ch. 8, pp. 217–241.

[48] Makino S. et al. (Eds.), “Blind Speech Separation”, pp 217-224, July 2007 Springer.

[49] S. Rickard and O. Yilmaz, “On the approximate W-disjoint orthogonality of speech,” in IEEE International.

[50] A. Jourjine, S. Rickard, Ö. Yilmaz, ”Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures”, ICASSP 2000.

[51] Rickard S., Balan R., and Rosca J., “ Real-time time-frequency based blind source separation,” in 3rd International Conference on Independent Component Analysis and Blind Sources Separation (ICA 2001), Dec, 2001.

[52] Hu, Y. and Loizou, P. (2007). “Subjective evaluation and comparison of speech enhancement algorithms,” *Speech Communication*, 49, 588-601.

[53] Sovka, P., “Extended Spectral Substraction:Description and Preliminary Results”, [Research Report]. R95-2. Prague, CTU, Faculty of Electrical Engineering 1995. Pp 15.

[54] H. Hirsch, and D. Pearce (2000). “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions.” ISCA ITRW, ASR2000, Paris, France, September 18-20.

[55] C. F’evotte, R. Gribonval, and E. Vincent, “Bss eval toolbox user guide,” IRISA, Rennes, Franc¸a, Tech. Rep. 1706, 2005.