Modulation-Based Speech Processing

MASTER'S THESIS

By

DANIEL SCHERAN

Signal Processing and Speech Communication Laboratory

GRAZ UNIVERSITY OF TECHNOLOGY

Supervisor:
Priv.-Doz. Pejman Mowlaee, PhD

Graz, August 2018

ABSTRACT

Speech is everywhere. As a result, there are countless applications where speech enhancement plays an important role, in particular for removing additive noise from speech. Typical applications are telephony, hearing aids, automatic speech recognition and hands-free devices. In the past, many single-channel speech enhancement algorithms were formulated in the acoustic frequency domain. Many of these algorithms are able to improve speech quality, but also decrease speech intelligibility. More recently, temporal spectral modulations of speech signals were found to be primarily responsible for intelligibility. More and more emphasis has been put on performing speech processing in the modulation domain, motivated by physiological and psychological findings on human speech perception.

In this thesis, we first elaborate the concept of a two-dimensional framework representing both acoustic frequency and modulation frequency. We discuss different kinds of modulation-based speech processing frameworks. Special attention is dedicated to the *Double Spectrum* domain, a pitch-synchronous modulation framework that is capable of preserving the harmonic structure of speech and describing how spectral components evolve over time. Secondly, speech and noise estimators are derived in Double Spectrum. The estimators in this domain are tested using objective criteria of speech quality, intelligibility and the degree of noise reduction. Lastly, further experiments are performed that demonstrate the potential of modulation-based speech processing. The various speech enhancement frameworks in modulation domain are tested thoroughly and compared with benchmark methods that were previously applied in the acoustic frequency domain. Results show that modulation-based algorithms are capable of significantly reducing noise and increasing speech quality while preserving intelligibility at the same time.

All experiments and algorithms that are used throughout this thesis were implemented in *MATLAB*.

ZUSAMMENFASSUNG

Sprache ist allgegenwärtig. Daher gibt es eine schier endlose Anzahl an Applikationen, in denen Sprachsignalverbesserung, im Speziellen die Unterdrückung von additivem Rauschen, eine entscheidende Rolle spielt. Beispiele hierfür sind etwa Telefonie, Hörgeräte, automatische Spracherkennung oder Freisprecheinrichtungen. In der Vergangenheit wurden Algorithmen zur Rauschunterdrückung hauptsächlich in der akustischen Frequenzdomäne implementiert. Viele dieser Algorithmen können zwar die Sprachqualität verbessern, verringern jedoch gleichzeitig die Sprachverständlichkeit. Jüngere Forschungsergebnisse zeigen, dass temporale Modulationen im Sprachsignal hauptsächlich für die wahrgenommene Verständlichkeit verantwortlich sind. Motiviert durch mehrere Studien über die menschliche Sprachwahrnehmung, wurde in den letzten Jahren Sprachsignalverarbeitung in der Modulationsdomäne vermehrt Beachtung geschenkt.

Zu Beginn dieser Arbeit wird das Konzept einer zweidimensionalen Signaltransformation, bei der sowohl die akustische Frequenz als auch die Modulationsfrequenz repräsentiert wird, vorgestellt. Besonderes Augenmerk wird hierbei auf das *Double Spectrum* gelegt, eine Umgebung, in der das Sprachsignal abhängig von seiner Grundfrequenz in seine spektralen Komponenten und deren temporale Modulationen zerlegt wird. Die harmonische Struktur von Sprache bleibt dabei erhalten. Des Weiteren werden statistische Schätzalgorithmen für Sprache und Rauschen vorgestellt. Die im Double Spectrum implementierten Algorithmen werden anhand objektiver Kriterien im Sinne von Sprachqualität, -verständlichkeit und Rauschunterdrückung evaluiert. Schlussendlich werden Experimente durchgeführt, um das Potential von Sprachsignalverarbeitung in der Modulationsdomäne aufzuzeigen. Verschiedene Frameworks in der Modulationsdomäne werden überprüft und mit Referenzmethoden in der akustischen Frequenzdomäne verglichen. Die Resultate zeigen, dass modulationsbasierte Algorithmen in der Lage sind, Rauschen signifikant zu reduzieren und dabei die Sprachqualität zu verbessern, während die Verständlichkeit erhalten bleibt.

Alle Experimente und Algorithmen im Rahmen dieser Arbeit wurden in *MATLAB* implementiert.

ACKNOWLEDGEMENTS

After a period of ten months, it is now time to thank all the people that have made the emergence of this thesis possible. Without their help and support, I would not be in the situation of writing these lines right now. Some people deserve particular appreciation for their contributions.

First and foremost, I would like to thank my supervisor Pejman Mowlaee for his excellent support, for the fruitful discussions and the helpful advices at the right time. I want to express my appreciation that the communication worked so well between Copenhagen and Graz. My most profound thanks for providing me with an employment in the *PhaseLab* research group and for giving me the chance to get insight into this fascinating field of research.

I want to thank my *PhaseLab* collegues Johannes Stahl and Sean Wood for their support and the interesting discussions at the office and also for the many funny moments.

Furthermore, I want to express my gratitudes to the staff of the SPSC Lab at TU Graz for allowing me to feel valued and being a part of the team from the beginning. I would also like to mention my Audio Engineering fellow students who shared the ups and downs of university life with me during the last years.

Last but not least, I want to thank my parents at that point for supporting me financially and mentally during my entire life and the past six years of my studies in particular. I never had to worry about anything existential and could and can always rely on being supported unconditionally at any time. Danke Mama und Papa!

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. The text document uploaded to TUGRAZonline is identical to the present Master's thesis.

DATE: ..... SIGNATURE: .....

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Speech enhancement is a technical way for improving aspects like quality and intelligibility of speech that has been degraded by some disturbances like additive noise, speech distortion or reverberation. The main purpose of single-channel speech enhancement is to improve these aspects when speech has been recorded only by a single microphone. In this thesis, we will focus on removing additive noise of different kinds from a speech signal. Algorithms for single-channel speech enhancement occur in many applications of every-day life. Some examples for the great variety of scenarios where noise reduction is relevant are voice communication over a cellphone, hands-free devices or hearing aids. Such systems typically suffer from all possible kinds of background noises, caused by traffic, babbling people, a noisy communication channel, etc. Speech enhancement algorithms thus can be used to improve the speech quality and intelligibility at the receiving end. Furthermore, *automatic speech recognition* (ASR) algorithms can benefit from such algorithms as well. Noisy speech is pre-processed by a speech enhancement algorithm before being fed to the speech recognizer [1].

1.1 Motivation and Goal of Research

Traditionally, speech enhancement algorithms are mainly performed in the short-time spectral domain [1, 2]. Many approaches are based on the Analysis-Modification-Synthesis (AMS) framework, where a short-time Fourier transform (STFT) is used as analysis stage in order to enhance speech in the spectral domain [3, 4]. Psychoacoustical and physiological studies have shown that the auditory system interprets sounds by considering the modulations of its spectral components [3]. This motivates to extend the framework to the modulation domain, where we can directly select and modify signal components by its spectral modulation content. As will be shown, this can provide some advantages, especially

in terms of speech intelligibility, over the conventional algorithms in short-time spectral domain.

In this thesis, we will present several modulation-based frameworks for speech enhancement. The main emphasis will be on a framework called *Double Spectrum* (DS). The fundamentals of Double Spectrum were proposed by Nilsson et al. in [5] and developed further and presented in detail by Blass, Mowlaee et al. in [6, 7]. Algorithms for noise reduction in this domain will be presented in detail and thoroughly tested. The main goal is to derive speech enhancement algorithms in Double Spectrum that are capable of jointly improve speech intelligibility and perceived speech quality, given a noisy observation of speech. Properties of the Double Spectrum domain will be exploited to achieve this goal. The results of the experiments will be compared with conventional speech enhancement algorithms both in spectral amplitude and spectral modulation domains.

1.2 Thesis Structure

The remainder of the work is organized as follows. In Chapter 2, the concepts of speech production and speech perception will be presented. The advantages of the modulation domain and its relation to psychoacoustics will be elaborated. We will provide a proof of concept and will present different modulation processing frameworks. In Chapter 3, different statistically based speech estimation algorithms in modulation domain will be presented. Chapter 4 will focus on noise estimation, again with a focal point on the Double Spectrum domain. In Chapter 5, several experiments will be presented and the proposed methods will be evaluated. What is more, different speech processing frameworks will be compared. Finally, conclusions will be drawn in Chapter 6.

CHAPTER 2

SPEECH SIGNALS AND THE MODULATION DOMAIN

This chapter will provide the theoretical background of modulation-based speech processing and an overview of previous research. The process of speech production will be elaborated as well as the relation of the modulation domain to psychoacoustics. Furthermore, a proof-of-concept, pointing out the advantages of modulation-based speech processing, will be presented. Finally, we will have a look on different processing frameworks in which speech enhancement can be performed.

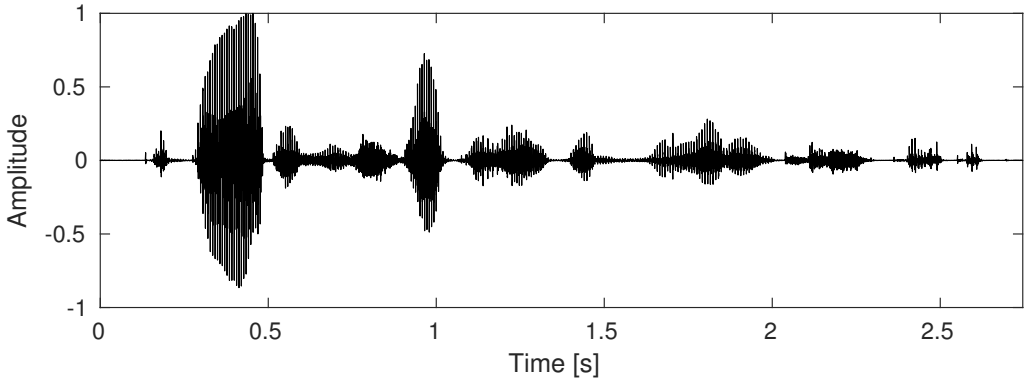
2.1 Speech Production and Perception

For the understanding of speech enhancement algorithms described in this thesis, it is important to comprehend what speech actually consists of and to have knowledge of the speech production process, as well as some understanding for the listener's speech perception.

2.1.1 Speech Production

A speech signal is a highly non-stationary process with a power spectrum that changes over time. However, when speech is divided into short periods of time in the range of 10-30 ms, the spectral characteristics can be assumed to be *wide-sense stationary* (WSS) [1]. The main physiological components that are involved in speech production are the following. The lungs save air during breathing and supply the glottis with air during speaking. The glottis, where the vocal folds are located, transforms the air flow into a time-varying velocity signal – the excitation. The vocal tract consists of the mouth and nasal cavity, respectively, and the lips. It shapes the excitation signal into perceptible speech sounds. The velum can be seen as a switch, selecting one of the two cavities or both [8].

(a) time signal representation



(b) spectrogram

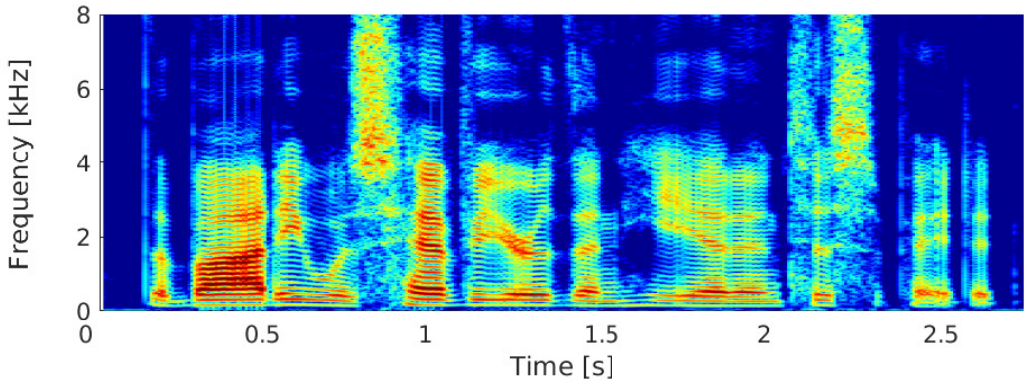


FIGURE 2.1. Speech signal *"The body was heavier than he had anticipated"*, spoken by a male speaker.

Speech sounds can roughly be classified into four categories. We refer to *voiced speech* when the vocal cords oscillate with a certain frequency. This fundamental frequency is also called the *pitch* (f_0). It typically lies within a range of $f_0 \in [50, 250]$ Hz for male speakers, and $f_0 \in [120, 500]$ Hz for female speakers, respectively [9]. At this voicing state, both excitation and speech are periodic and produce the excitation signal consisting of f_0 and its harmonics. Voiced sounds are for example vowels like /a/ or /e/. At *unvoiced speech*, the vocal cords are open. The excitation and the produced speech are noisy and do not have a pitch (e.g. /s/, /f/). When there is a transition from voiced to unvoiced speech, we talk about *mixed speech*. Both excitation and speech are noisy periodic signals. The last category is *plosive speech*. Examples for plosive sounds are /p/ or /t/. They are produced by complete closure of the vocal cords and vocal tract, followed by a sudden release of air pressure. Both excitation and speech are impulsive [8, 9].

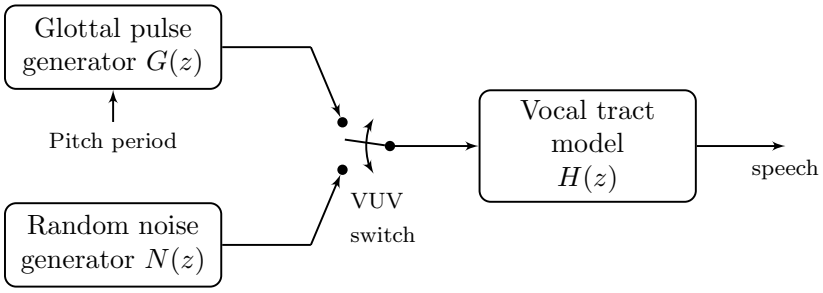


FIGURE 2.2. Source filter model of speech production (after [1]).

The vocal tract is a set of cavities, functioning as resonators. Harmonics produced by the vocal cords are emphasized depending on the current shape of the vocal tract. Speech is produced by changing the resonating behavior of the vocal tract over time. The range of frequencies in which there is an absolute or relative maximum in the sound spectrum is called a formant [10]. By moving the articulators (velum, tongue, lips and jaw), the shape of the vocal tract, and as a result the formants and the produced speech, can be dynamically controlled [11]. The speech signal is then emitted via the lips and carried by air to the listener's ears.

Speech signals can be represented visually in different terms. Figure 2.1 shows two examples. The top figure (a) shows the representation of the time signal, the bottom figure (b) shows the spectrogram that depicts the power of certain frequency regions over time. The distinction between voiced and unvoiced sounds as well as the harmonics of voiced speech are visualized in the spectrogram.

The speech production process can be approximated by a *source filter model* (SFM). It consists of two components: (i) The excitation source, constituting the influence of the lungs and the glottis, and (ii) a time-varying digital filter, representing the vocal tract [9]. Figure 2.2 depicts the block diagram of the source filter model. The following assumptions have to be made: (i) The excitation signal and vocal tract are independent, (ii) the vocal tract is assumed to be stationary for short periods of time, usually 20-40 ms. The SFM thus models the single WSS speech frames and not the entire signal [8]. The vocal cords can take on one out of two states: voiced speech is modeled by a periodic sequence of pitch pulses, unvoiced speech is modeled by a random noise generator. The current source is chosen by a switch. The vocal tract is approximated by a time-varying digital filter $H(z)$ [1, 9]. Multiplying the impulse responses of the excitation signal $G(z)$ or $N(z)$ with the vocal tract model $H(z)$ results in the output speech.

2.1.2 Speech Perception

After being emitted by a speaker, speech is received by the listener's ears. The ear is constituted of three sections: the outer ear, the middle ear and the inner ear. The outer ear consists of the pinna, the outer ear canal and the ear drum. The ear canal transmits the sound to the ear drum and has a single resonance frequency between 3 and 4 kHz [9]. The auditory system is thus most sensitive within this frequency range. Behind the ear drum, the middle ear is located. It is an air-filled cavity and contains the three auditory ossicles: hammer, anvil and stirrup. They connect the eardrum with the oval window and provide impedance adjustment. In the inner ear, the basilar membrane inside the cochlea performs a transformation of frequency to place by a traveling wave mechanism. Low frequencies stimulate hair cells of the organ of Corti near the centrum of the cochlea while high frequencies have the maximum near the oval window. This transformation can be interpreted as a spectral analysis using a non-uniform filterbank. The organ of Corti is located above the basilar membrane and senses vibrations with about 3,600 inner and about 26,000 outer hair cells. Stimulations of the hair cells are transferred into neural signals that are transmitted to the auditory center in the brain by the auditory nerve [9].

Speech is transferred by periodic fluctuations of air pressure. Since the human hearing is very sensitive and very robust at the same time, the range for hearing covers more than seven decades from the hearing threshold of 10^{-5} Pa up to 10^{2} Pa [9]. Therefore, it is reasonable to measure the sound pressure p in a logarithmic scale in dB:

$$L_p = 20 \cdot \log_{10}\left(\frac{p}{p_0}\right), \tag{2.1}$$

where L_p is the sound pressure level and the reference pressure p_0 is defined as 20 μPa. The perception of loudness is frequency-dependent. Figure 2.3 shows the curves that depict constant subjective loudness perception. The reference is a sinusoidal tone of $f = 1$ kHz. In listening tests, it was determined at which sound pressure level other frequencies are perceived with the same loudness as the reference. Thus, at the reference frequency, the loudness level is identical to the sound pressure level. The loudness curves are also called *equiphon-curves* as the pseudo-unit for the subjective loudness level is *phon* [9].

The frequency perception of the auditory system is closely related to the excitation patterns on the basilar membrane. Listening tests have indicated that in a frequency range up to 16 kHz, the auditory system is capable to distinguish between approximately 640 frequency steps. Further tests have also shown that the ear integrates the excitation over 24 frequency intervals that are called the *critical bands*. These bands can also be described as a filter bank with non-uniform spectral resolution [9].

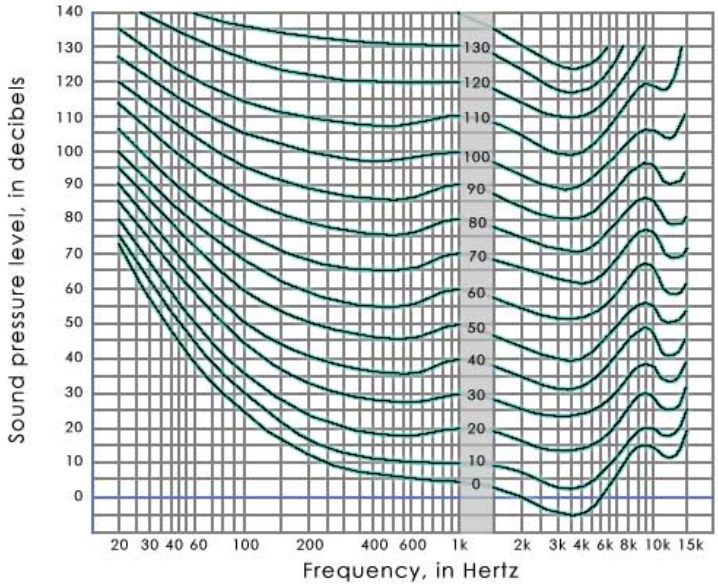


FIGURE 2.3. Contours of equal loudness level (equiphon-curves). The lowest curve depicts the hearing threshold, the values in the grey box denote the loudness level in phon. (Taken from *http://decoy.iki.fi/dsound/dsound-c-04*, accessed on 08/06/2018.)

In the last decades, research has revealed that not only loudness and acoustic frequency play a role in speech perception, but that also the temporal modulations are crucial for speech intelligibility [12, 13]. The next section will provide a brief overview of the relation of speech perception and the modulation domain.

2.2 Modulation Domain and Psycho-Acoustics

It was believed for a long time that the spectral envelope is the principal carrier of information in human speech [6]. In the last decades, growing psychoacoustical and physiological evidence was found that temporal modulations of the spectral envelope carry important information and highly correlate with speech intelligibility [14]. For example, when the modulators of a speech signal are replaced by envelopes of constant amplitude, speech becomes unintelligible, while it remains highly intelligible when the modulators are preserved but the carriers are replaced by white noise [15]. The role of temporal modulation has been investigated in a number of studies and experiments. A short summary will be provided below.

In 1940, Dudley [16] pictured a speaker as a sort of radio broadcast transmitter. An audible sound is needed to transmit the message to the listener. This is provided by a group of higher frequency waves in the audible range that is modulated by "message waves" to form the speech waves. According to this theory, the information is transmitted by an amplitude modulated carrier and received by the listener's ear and recovered by the listener's mind. That is, Dudley discovered that speech is actually a low-bandwidth process that modulates higher carriers [17].

Zwicker [18] was one of the first to systematically investigate the auditory system in terms of its sensitivity for amplitude modulations. He measured the just perceptible amplitude modulation and the just perceptible frequency modulation of sine tones of different pitch and loudness as a function of the modulation frequency. He discovered that the ear is capable of distinguishing between amplitude modulation and frequency modulation below a so-called phase-limit frequency, that is at about 30 Hz for low acoustic frequencies. Furthermore, it was shown that amplitude modulation is perceived much better than frequency modulation.

Similarily, Viemeister [19, 20] measured the *temporal modulation transfer function* (TMTF) of the auditory system. The amplitude and phase spectra are determined from masking curves obtained at several modulation frequencies. They ideally describe the attenuation characteristic of the temporal processing system and have a lowpass character. The results show that when modulation frequency is increased, the threshold amplitude of modulations remains constant up to approximately 10 Hz. For higher modulation frequencies, the threshold increases constantly up to 800 Hz. In 1989, Bacon and Grantham [21] showed in their experiments that there are channels in the auditory system that are tuned for the detection of modulation frequency, like there are channels tuned for the the detection of spectral frequencies. Following studies further investigated the ability of the human auditory system to detect amplitude modulations, such as Sheft and Yost (1990) [22] and Lee (1997) [23].

An important contribution to the development of modulation-based noise suppression algorithms is the work of Drullman et al. [12, 13]. In their studies, it was shown that modulation frequencies between 4 and 16 Hz are crucial for speech intelligibility. *Modulation frequencies* (MF) below 4 Hz and above 32 Hz can be reduced without severely impairing intelligibility. A special emphasis is put on modulation frequencies of 4-5 Hz that are shown to be most important. These adoptions are supported by the work of Arai et al. (1996) [24] where it is indicated that speech intelligibility is not severely negatively impacted when bandpass filtering speech modulations, as long as rates of change between 1 and 16 Hz are

preserved. This modulation frequency range approximately corresponds to the speed of movements of the vocal tract and the syllabic rate [25, 26].

In 2003, Atlas and Shamma [17] claimed very low-frequency modulation of sound to be the fundamental carrier of information in speech and of timbre in music. They presented a two-dimensional modulation spectral model for an analysis-modification-synthesis framework in order to be able to perform speech enhancement in the modulation domain.

A detailed summary of the modulation concept in psycho-acoustics can be found in Schimmel's dissertation (2007) [15].

The psychological and physiological findings described in this chapter motivated researchers to extend the speech signal processing framework from the short-time spectral amplitude to modulation based speech processing. Meanwhile, speech processing in modulation domain is applied in a variety of areas, like speech coding, speech recognition, speaker recognition, objective speech intelligibility evaluation as well as in speech enhancement [14].

2.3 Representations of Speech Signals

In this section, various spectral representations of speech signals will be presented. The basis of many speech enhancement algorithms is a spectral transformation. The main purpose of such transformations is to gain access to signal properties that are better depictable in a spectral representation than in time domain. Plus, it is desireable to perform a processing task with computational efficiency. Typically, such a representation consists of a model and its parameters and the resulting signal coefficients [27]. A compact and sparse signal representation is desired.

To a great extent, speech enhancement is taking place in an *Analysis-Modification-Synthesis* (AMS) framework. In the analysis stage, the signal is transfered into the desired framework. In the modification stage, the noisy signal is compensated for noise and distortion to get the modified spectrum. Finally, in the synthesis stage, inverse transforms are performed in order to obtain the enhanced time signal [28]. Four different AMS frameworks for speech enhancement will be presented: the *Short-Time Spectral Amplitude* (STSA) framework, the *Short-Time Spectral Modulation* (STSM) framework, the *Coherent Harmonic Demodulation* (CHD) and the *Double Spectrum* (DS).

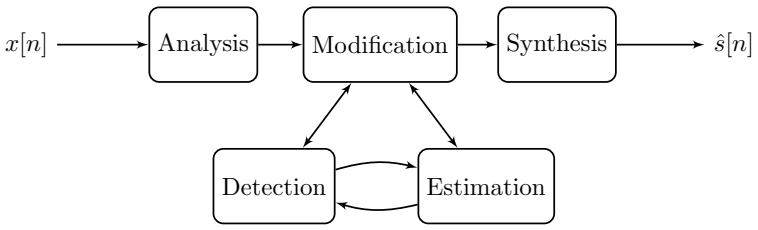


FIGURE 2.4. AMS framework in general.

2.3.1 Short-Time Spectral Amplitude Domain (STSA)

In the STSA domain, speech enhancement is taking place in the time-frequency domain. To get into this framework, a time domain signal $x(t)$ is transferred to its frequency domain representation using the Fourier transform. In general, the Fourier analysis of time continuous signals is defined as:

$$X(j\omega) = \int_{-\infty}^{\infty} x(t)\, e^{-j\omega t}\, dt, \tag{2.2}$$

where t is the time index and $\omega = 2\pi f$ denotes the radian frequency [9]. However, in digital signal processing, we do not have continuous time signals, but we have to deal with discrete-time sampled data. A discrete time signal is only defined at integer multiples of its sampling period T_s, with the sampling frequency $f_s = 1/T_s$, expressed in Hz [1]. The *Discrete Fourier Transform* (DFT) is designed to transfer signal segments of finite length to a discrete spectrum of the same length:

$$X(k) = \sum_{n=0}^{N-1} x[n]\, e^{-j\frac{2\pi kn}{N}}, \qquad 0 \le k \le N-1, \tag{2.3}$$

where n is the discrete time index, k is the frequency index and N is the length of the signal segment $x[n]$. The DFT provides a discrete spectrum with equally spaced frequency bins on the normalized frequency axis, with π representing $f_s/2$. The spacing of frequency bins is expressed by $\Delta f = f_s/N$.

Speech is a highly non-stationary signal. Therefore, it is not expedient in speech processing to perform the Fourier analysis on a long speech signal as a whole. On that account, speech is often processed in short time blocks with lengths of 10-30 ms. Within these short periods of time, the speech signal does not change very much (i.e., it is wide-sense stationary) [1]. A sliding analysis window function $w_A[n]$ is used to divide the signal into a number of time blocks that are analyzed separately by the DFT. The window function is of finite

length N, the samples outside the function are defined as zero:

$$w_A[n] = \begin{cases} w[n], & 0 \leq n \leq N-1 \\ 0, & \text{otherwise} \end{cases} . \tag{2.4}$$

At each analysis step, the window is shifted by Z samples. We refer to that block-wise analysis of a signal as *Short-Time Fourier Transform* (STFT) that is defined as:

$$X(k,m) = \sum_{n=0}^{N-1} x[n+mZ]\ w_A[n]\ e^{-j\frac{2\pi kn}{N}}, \tag{2.5}$$

where m is the index of the current time frame.

The coefficients that are obtained by any kind of the Fourier transform are complex-valued. This means that they can be separated into a magnitude and phase spectrum (polar representation), as

$$X(k,m) = |X(k,m)|\ e^{j\angle X(k,m)}, \tag{2.6}$$

with $|X(k,m)|$ being the magnitude and $\angle X(k,m)$ is the phase of an STFT coefficient. In many speech enhancement applications, only the magnitude part is being processed while the phase is left unchanged [29]. When doing so, we talk about *Short-Time Spectral Amplitude* (STSA) processing of speech. In more recent studies, more emphasis was placed upon phase-aware speech processing (see e.g. [30]). However, speech enhancement is performed on the magnitude, the phase or both in order to gain the modified speech spectrum $\hat{S}(k,m)$. The enhanced time signal can be obtained by applying the *inverse STFT* (iSTFT) on the enhanced spectrum, followed by an *overlap-add* (OLA) procedure:

$$\hat{s}[n] = \sum_m \left(w_A[n-mZ] \sum_{k=0}^{N-1} \hat{S}(k,m)\ e^{j\frac{2\pi kn}{N}} \right). \tag{2.7}$$

A block diagram of the AMS procedure in STSA domain is depicted in Figure 2.5.

Speech enhancement algorithms in STSA domain

As the STSA framework is the most commonly used one for speech enhancement, there is a great variety of algorithms for speech enhancement in this domain. To mention a few, without any claim to completeness:

- **Spectral subtraction:** This method is historically one of the first algorithms proposed for noise reduction in speech signals (see e.g. [31]). Many varieties of this algorithm

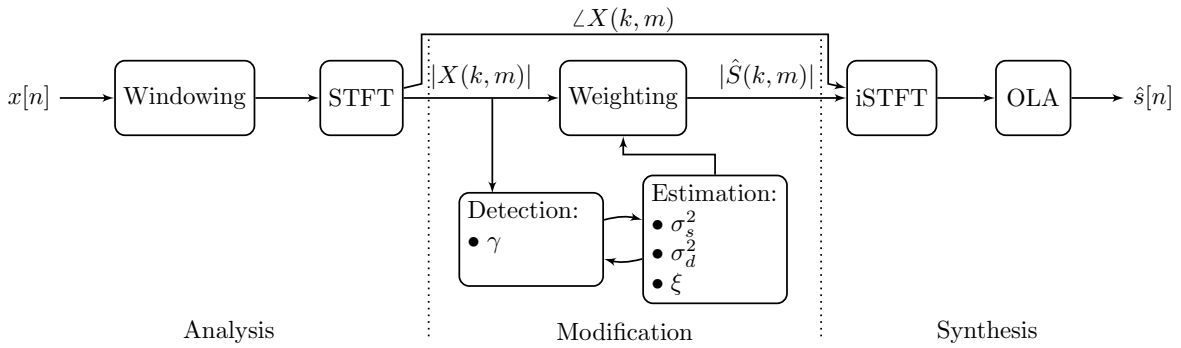


FIGURE 2.5. Block diagram of the AMS framework in STSA domain.

have been developed. The principle is simple. An estimate of the clean spectrum is obtained by subtracting an estimate of the noise signal from the noisy signal:

$$\left|\hat{S}(k,m)\right| = |X(k,m)| - \left|\hat{D}(k,m)\right|, \tag{2.8}$$

The noise spectrum $\left|\hat{D}(k,m)\right|$ is estimated and updated during segments where speech is absent. The method has to be applied carefully in order to avoid speech distortions caused by over-subtraction. Plus, this method is relatively susceptible for a phenomenon called *musical noise* that occurs due to inaccurate noise estimation and that is caused by random fluctuations of frequency bins [1].

- **Wiener filtering:** The Wiener filter (presented in [32]) can be derived by minimizing the *mean-square error* (MSE) between the output signal $\hat{s}[n]$ and the desired signal $s[n]$. For this purpose, a cost function is defined and the MSE is minimized. The approach will be presented in detail in Section 3.3.

- **Statistical-model-based methods:** In this approach, speech and noise underlie statistical models. The optimal solution is found by *maximum likelihood* (ML), *maximum a-posteriori* (MAP), *minimum mean-square error* (MMSE) or Bayesian estimators [1]. One famous example is the MMSE short-time spectral amplitude estimator from Ephraim and Malah (1984) [33].

- **Subspace algorithms:** This class of algorithms is based on linear algebra theory. On the basis of singular value decomposition or eigenvector-eigenvalue factorizations, additive and uncorrelated speech and noise signals can be decomposed into orthogonal speech and noise subspaces [1]. Subsequently, the noise subspace can be removed from the signal.

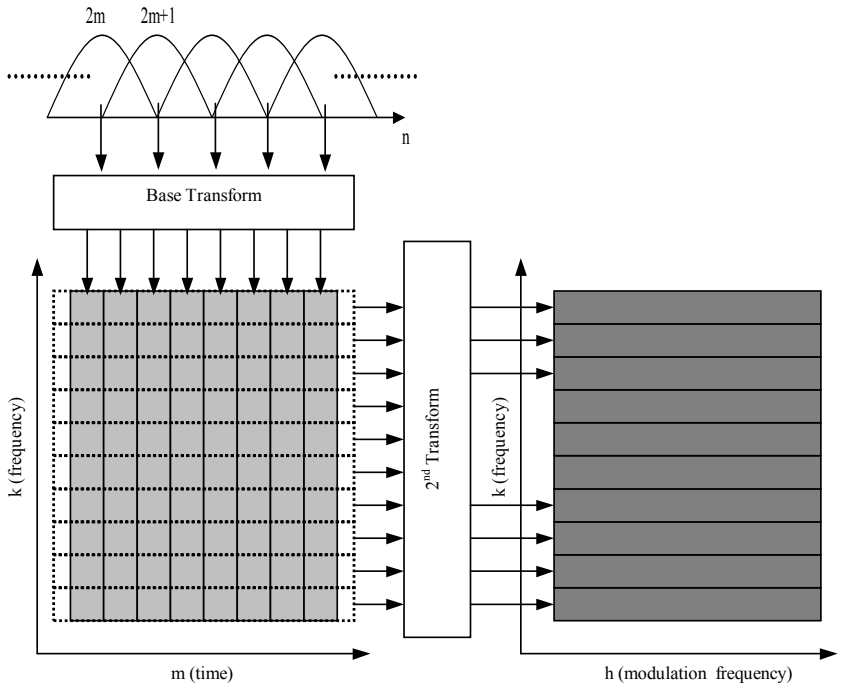


FIGURE 2.6. Two-dimensional spectral modulation framework (taken from [6], after [17]).

2.3.2 Short-Time Spectral Modulation Domain (STSM)

Motivated by the findings presented in Section 2.2, researchers aimed to include spectrotemporal modulations in an AMS framework. In 1950, Zadeh [34] firstly proposed a two-dimensional system function with two separate frequency dimensions. One was the acoustic frequency and the other one a transform of the time variation. The resulting bi-frequency representation was only defined, but not thouroughly analyzed. This was addressed by Kailath [35] in 1961 [17]. Influenced by these previous studies, a joint frequency-modulation frequency framework was proposed by Atlas and Shamma (2003) [17] in order to "[...] integrate a concept of sensory perception with signal processing methodology to achieve a significant improvement in the representation and coding of acoustic signals".

A definition of the short-time spectral modulation domain can be found in [28]:

> "[...] the modulation spectrum at a given acoustic frequency is the STFT of the time series of the acoustic spectral magnitudes at that frequency. The short-time modulation spectrum is thus a function of time, acoustic frequency and modulation frequency."

The structure of the two-dimensional framework is depicted in Figure 2.6. To get into the

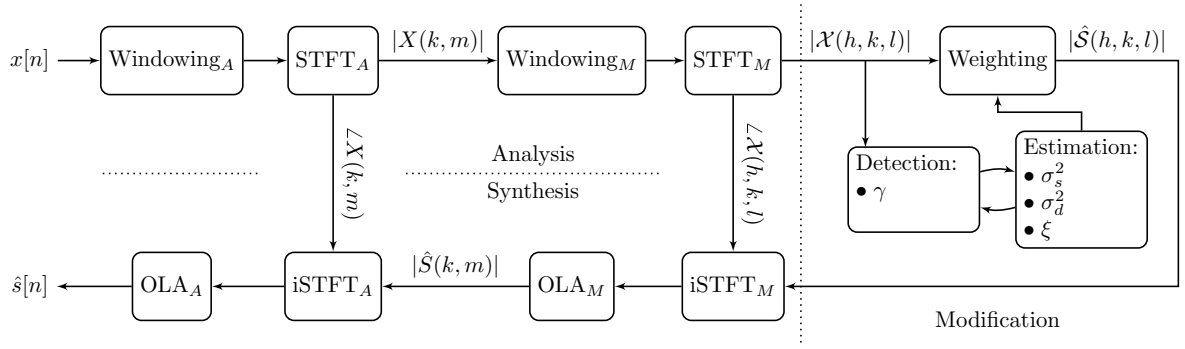


FIGURE 2.7. Block diagram of the AMS framework in STSM domain.

spectral modulation domain, we start with the spectral magnitude of the complex spectrum $|X(k,m)|$, obtained by the STFT in Equation (2.5). The time trajectories of each frequency bin are then transformed framewise using a second AMS procedure. The modulation spectrum, $\mathcal{X}(h,k,l)$ is then computed using a modulation window function $w_M(m)$ [28]:

$$\mathcal{X}(h,k,l) = \sum_{m=0}^{\mathcal{M}-1} |X(k,m+l\mathcal{Z})|\ w_M(m)\ e^{-j\frac{2\pi mh}{\mathcal{M}}}, \tag{2.9}$$

where h is the modulation band index, k is the acoustic frequency index, l is the index of the modulation frame, $\mathcal{M}$ is the *modulation frame duration* (MFD, number of time-frequency frames) and $\mathcal{Z}$ is the modulation frame shift. Like the time-frequency spectrum in (2.6), the modulation spectrum can also be represented in polar form as

$$\mathcal{X}(h,k,l) = |\mathcal{X}(h,k,l)|\, e^{j\angle\mathcal{X}(h,k,l)}, \tag{2.10}$$

where $|\mathcal{X}(h,k,l)|$ is the modulation magnitude spectrum and $\angle\mathcal{X}(h,k,l)$ is the modulation phase spectrum. Again, in many modulation based speech enhancement algorithms, only the magnitude is modified while the phase is left unchanged. In that case, we perform speech enhancement in *Short-Time Spectral Modulation* (STSM) domain. However, there is some evidence that the modulation phase contains important information of speech and is crucial for intelligibility [29, 36]. After modification, the inverse transform back to the enhanced time-frequency representation is performed analogous to Equation (2.7), followed by an OLA procedure. Subsequently, the iSTFT (2.7) is applied one more time to obtain the enhanced time signal $\hat{s}[n]$. A block diagram of the STSM framework is depicted in Figure 2.7.

Speech enhancement algorithms in STSM domain

The idea of performing speech enhancement in modulation domain has led to different algorithms with encouraging results. Many methods that had already been established in STSA domain have been adapted to STSM domain. For example, it is possible to perform spectral subraction [14], Wiener filtering or MMSE magnitude estimation [28] in modulation domain as well. These STSM based algorithms improve perceived quality with less musical noise and spectral distortions compaired to their counterparts in STSA domain [37].

2.3.3 Coherent Harmonic Demodulation (CHD)

Clark and Atlas (2011) [38] proposed a method for modulation filtering based on a product model where subband envelopes are determined from demodulation of the subband signal with a coherently detected subband carrier. This method is entiteled as Coherent Harmonic Demodulation. Any signal that is analytic and continuous, may be factorized into a continuous modulator-carrier pair [38]. CHD ensures distortion-free filtering of modulators while the harmonic structure of speech is retained [39]. With this method, the matter of fact that each modulator can be modified independently of the carrier and be recombined with its corresponding one separately, is exploited.

Speech can be considered as a discrete-time full-band signal which can be described by a signal product model as:

$$x[n] = \sum_{k=1}^{K} s_k[n] = \sum_{k=1}^{K} m_k[n] \circ c_k[n], \tag{2.11}$$

where $x[n]$ is the observed speech signal, $m_k[n]$ and $c_k[n]$ represent the k^{th} modulator and carrier waveforms of each subband $s_k[n]$, respectively. K is the number of subbands and the ($\circ$) operator denotes the *Hadamard product* (sample-by-sample multiplication). CHD can directly estimate the carrier from the input signal. By using harmonic demodulation, the spectral band is related to an integer multiple of the time-varying fundamental frequency f_0 of the speech signal. In regions where the input speech is unvoiced or weakly voiced, f_0 is interpolated between two valid detections. The primary difference to non-coherent demodulation techniques is that the harmonic instantaneous frequencies are smooth and bandlimited [39]. After computing the carrier, the modulator of each harmonic track can be extracted:

$$m_k = c_k^*[n] \circledast \sum_{p=0}^{n} h[n-p]x[p], \tag{2.12}$$

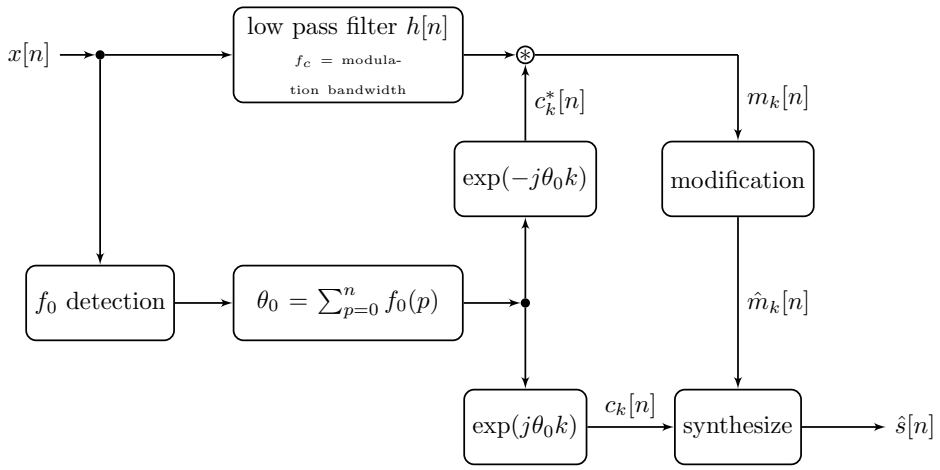


FIGURE 2.8. Block diagram of CHD technique (after [39]).

where $h[n]$ is the impulse response of a low-pass filter whose cut-off frequency is defined by the modulation bandwidth and $c_k^*[n]$ is the complex-conjugate of the carrier. $m_k[n]$ can now be transfered to frequency domain by the STFT and can be processed accordingly to obtain the enhanced modulators $\hat{m}_k[n]$. The block diagram of CHD technique is shown in Figure 2.8.

A speech enhancement algorithm in CHD domain was proposed by Samui et al. (2017) [39], where the MMSE estimator of Ephraim and Malah [33] was applied to noisy modulators in the CHD domain.

2.3.4 The Double Spectrum (DS)

Inspired by the works of *"A Canonical Representation of Speech"* by Nilsson et al. (2007) [5], the Double Spectrum was proposed by Blass, Mowlaee and Kleijn in [6, 7, 37]. The transformation procedure in DS consists of a pitch-synchronous transform followed by a modulation transform. We will refer to this procedure as *two-stage transform*. The key feature is the estimation of the fundamental frequency, providing a sparse speech-signal representation [37]. The two-stage transform approximates the *Karhunen-Loève Transform* (KLT) and maximizes the coding gain [5, 37]. In DS, it is possible to separate periodic and aperiodic components of the signal. These properties can be used to separate speech and noise and to enhance speech quality. Some important properties of the Double Spectrum are sparsity, linearity and real-valued coefficients [37].

The analysis stage of DS representation consists of two blocks [37]: (i) the *time block segmentation* (TBS) and (ii) the two-stage transform. The TBS divides the input speech

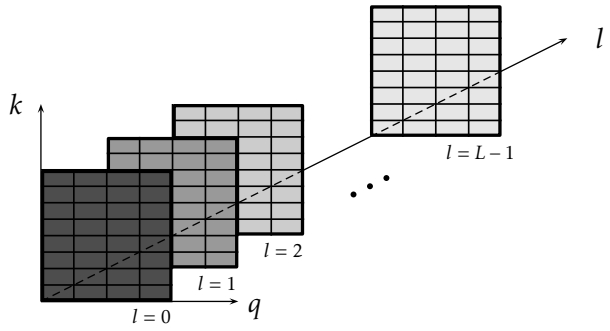


FIGURE 2.9. Illustration of the Double Spectrum Matrix $DS^{(l)}(q,k)$ (taken from [6]).

signal into L time blocks of variable length. The length of a time block is an integer multiple of the pitch period $P_0 = f_s/f_0$ in samples, with each block being further subdivided into frames of length P_0. Subsequently, each time block undergoes the two-stage transform. The first stage, i.e. the pitch-synchronous transform, is implemented as a *Modulation Lapped Transform* (MLT) [40], in the shape of a *Discrete Cosine Transform* (DCT-IV). The first-stage transform is defined as follows [37]:

$$f(k,l) = \sum_{n=0}^{2P_0-1} x_l[n]\ w[n] \sqrt{\frac{2}{P_0}} \cos\left(\frac{(2k+1)(2n-P_0+1)\pi}{4P_0}\right), \tag{2.13}$$

with $x_l[n]$ being the input signal, l denotes the l^{th} pitch-synchronous time block, k is the index of the frequency band and $w[n]$ is the window function. The MLT coefficients $f(k,l)$ of the output evolve slowly over time for voiced speech and rapidly for unvoiced speech. In order to analyze those temporal fluctuations, the modulation transform in shape of a DCT-II is applied to a number of consecutive frames of the MLT coefficients:

$$g(q,k,l) = \sum_{l=0}^{Q-1} f(k,l)c(q) \sqrt{\frac{2}{Q}} \cos\left(\frac{(2k+1)q\pi}{2Q}\right), \tag{2.14}$$

with $q = 0,1,...,Q-1$ being the modulation band index and $c(0) = 1/\sqrt{2}, c(q) = 1$ for $q \neq 0$. In the further notation, we will refer to the set of l subsequent DS frames as $DS^{(l)}(q,k)$ with K frequency bands as rows and Q modulation bands as columns. An illustration of DS frames is depicted in Figure 2.9 and the DS AMS framework is shown in Figure 2.10. Approaches for speech enhancement in Double Spectrum will be examined in detail in Chapter 3.

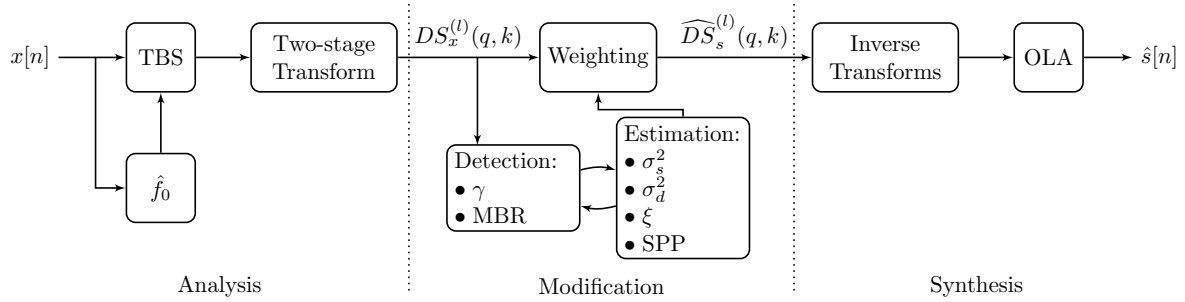


FIGURE 2.10. Block diagram of the AMS framework in DS domain (after [6]).

2.4 Proof of Concept: Ideal Channel Selection

Bringing this chapter to an end, the benefits of speech processing in modulation domain are shown in a proof of concept. Boldt et al. (2015) [25] evaluated enhanced speech obtained by ideal binary masking in terms of intelligibility and quality. The attempt aims to compare binary masking methods in time-frequency domain and in modulation domain. Three methods have been investigated by the authors for the purpose of showing the differences in intelligibility:

- *Ideal Channel Selection* (ICS),
- *Modulation Channel Selection* (MCS),
- *weighted Ideal Channel Selection* (wICS).

Ideal Channel Selection (ICS)

An ideal binary mask is applied in the time-frequency domain (see Section 2.3.1). The binary gain is computed having ideal knowledge of the clean speech and noise signals, respectively. When a frequency bin is declared as being dominated by speech energy, i.e. the *Signal-to-Noise Ratio* (SNR) is higher than a certain threshold θ, a gain of one is used, otherwise a gain of zero is used:

$$G_{\mathrm{ICS}}(k,m) = \begin{cases} 1, & \frac{|S(k,m)|^2}{|D(k,m)|^2} \geq \theta \\ 0, & \text{otherwise} \end{cases}, \tag{2.15}$$

where $G(k,m)$ is the binary gain of an acoustic frequency bin k at the time frame m. $|S(k,m)|^2$ and $|D(k,m)|^2$ are the spectral energies of speech and noise, respectively. The gain

is applied to the noisy spectrum in order to gain the enhanced spectrum:

$$\left|\hat{S}(k,m)\right| = G_{\mathrm{ICS}}(k,m)\cdot|X(k,m)|\,. \tag{2.16}$$

Modulation Channel Selection (MCS)

MCS can be seen as ICS performed in the modulation domain, in particular in STSM domain (see Section 2.3.2). As a result, the binary gain is calculated in the modulation domain as a function of modulation time frame l, acoustic frequency k *and* modulation frequency h. Like ICS, the binary gain in MCS is calculated with knowledge of the clean speech signal $s[n]$ and noise signal $d[n]$ before being mixed to the noisy signal $x[n]$:

$$\mathcal{G}_{\mathrm{MCS}}(h,k,l) = \begin{cases} 1, & \frac{|\mathcal{S}(h,k,l)|^2}{|\mathcal{D}(h,k,l)|^2} \geq \theta \\ 0, & \text{otherwise} \end{cases}, \quad \left|\hat{\mathcal{S}}(h,k,m)\right| = \mathcal{G}_{\mathrm{MCS}}(h,k,m)\cdot|\mathcal{X}(h,k,m)|\,. \tag{2.17}$$

Weighted Ideal Channel Selection (wICS)

In the first place, it is found that MCS is not able to outperform ICS in terms of quality and intelligibility, although it has higher computational complexity (see Figure 2.11 and Table 2.1). This is due to the modulation frame length of 256 ms that introduces time smearing in the MCS processed time signal. We have to use such a long window to ensure a decent resolution of modulation frequencies.

In order to test the hypothesis that the long modulation frames reduce the benefit from MCS processing, wICS is introduced to make binary masking in time-frequency domain and in modulation domain directly comparable. wICS is ICS with a binary decision at the same time resolution as in MCS, hence it is introducing "artificial time smearing". This is done by time weighting the magnitudes of noisy speech $|X(k,m)|$ with the window w_M that is also used for modulation analysis.

Test results indicate that MCS is able to outperform wICS. Hence, it is shown that keeping specific modulation frequencies leads to benefits for intelligibility and quality in comparison to the more simple decision based on the total energy of each time frame. [25]

Figure 2.11 and Table 2.1 show the improvements for percepted speech quality (ΔPESQ, [41]) and speech intelligibility (ΔSTOI, [42]). The height of the bars indicate the enhancement in comparison to noisy speech at a certain SNR and the errorbars represent the 95% confidence interval. For the evaluation, 100 phonetically balanced sentences from the *TIMIT* corpus [43] (sampling frequency f_s = 16 kHz), corrupted with factory noise from the *NOISEX-92* database [44], were used.

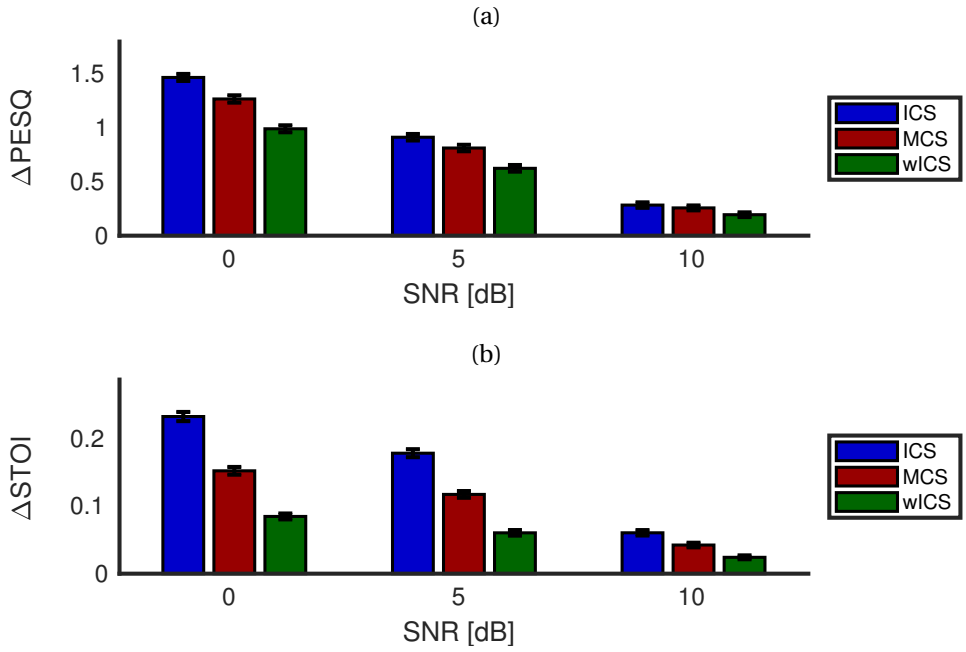


FIGURE 2.11. Improvements in (a) speech quality (ΔPESQ) and (b) speech intelligibility (ΔSTOI) of ICS, MCS and wICS over SNR, in *factory noise* environment.

	ΔPESQ			ΔSTOI		
SNR [dB]	0	5	10	0	5	10
ICS	1.47	1.27	0.99	0.23	0.15	0.09
MCS	0.91	0.81	0.63	0.18	0.12	0.06
wICS	0.28	0.26	0.19	0.06	0.04	0.02

TABLE 2.1. Mean improvements in speech quality (ΔPESQ) and speech intelligibility (ΔSTOI) of ICS, MCS and wICS over SNR, in *factory noise* environment.

CHAPTER 3

SPEECH ENHANCEMENT IN MODULATION DOMAIN

Speech enhancement algorithms aim to improve perceptual aspects of speech that has been degraded by additive noise. The emphasis is put on improving speech quality and intelligibility by removing as much noise as possible while the target speech should not be distorted. Speech enhancement algorithms therefore are also called *noise reduction algorithms* [1].

In single-channel speech enhancement, we deal with signals that are picked up by one single microphone. Therefore, we cannot take any advantage of spatial processing. The following additive signal model will be assumed throughout the whole thesis. The noise signal $d[n]$ is additive and uncorrelated, i.e. statistically independent of the speech signal $s[n]$. They mix up to the noisy signal $x[n]$:

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

The aim of speech enhancement algorithms is to gain an estimate $\hat{s}[n]$ of the clean speech signal from the observed noisy signal $x[n]$ without any prior knowledge of the involved parts $s[n]$ and $d[n]$. In the following, we will focus on single-channel speech enhancement in Double Spectrum domain. First, a review of already existing methods will be provided. Afterwards, new methods, like a DS-MAP estimator, will be proposed. In the end of the chapter, we will have a look at RASTA modulation filtering and the role of the sign of DS coefficients.

3.1 Double Spectrum Parameters

In Blass' Master's thesis (2016) [6] and the corresponding paper [7], several approaches for single-channel speech enhancement in Double Spectrum (DS) domain are presented.

Before describing these methods in detail, two important parameters that can be obtained from DS coefficients will be discussed: the *modulation band ratio* (MBR) and the *speech presence probability* (SPP).

3.1.1 Modulation Band Ratio (MBR)

The DS analysis stage requires pitch-synchronous segmentation of the input signal and hence an estimate of the fundamental frequency. Having a good estimate of f_0 leads to a sparse representation of speech with energy concentration in the low modulation bands. To obtain a good segmentation, the modulation band ratio is introduced. It is a measure of energy concentration in the lowest modulation band $q = 0$ that represents the periodic content of a signal. As a result, the MBR is a measure of the periodicity of a signal.

$$\mathrm{MBR}(K) = \frac{E_1}{E_{1:Q}}, \tag{3.2}$$

where $E_1 = \sum_k DS^2(0,k)$ represents the energy concentrated in the first modulation band $q = 0$ and $E_{1:Q} = \sum_q \sum_k DS^2(q,k)$ is the total energy in all modulation bands. For pitch estimation in periodic frames, the MBR is maximized, i.e. finding a block size K that maximizes the MBR leads to an optimal frame size [6]:

$$K_{opt} = \arg\max_K \mathrm{MBR}(K). \tag{3.3}$$

K_{opt} is searched in a range $K_{opt} \in [K_{min}, K_{max}]$ and it is a function of the fundamental frequency f_0. Using the relation between f_0 and K_{opt}, the an estimate of the fundamental frequency estimate can be found [6]:

$$f_{0,opt} = \frac{f_s}{K_{opt}}. \tag{3.4}$$

3.1.2 Speech Presence Probability (SPP)

The *speech presence probability* (SPP) measure can be used to distinguish between regions of speech presence and absence. It is possible to estimate the SPP from the noisy observation. The SPP in the DS domain can be computed from the MBR measure, which discriminates between voiced and unvoiced speech. MBR values close to one refer to voiced speech, values close to zero to unvoiced speech. The MBR itself hence is a good measure for SPP. [6]

The MBR as shown in Equation (3.2) results in a value for every frequency band K. In order to obtain an overall SPP measure, the MBR(K) values up to the 6^{th} harmonic are

averaged, i.e. $k \in [0, 11]$, since the spacing of frequency bins is $\Delta f_k = f_0/2$. This yields robust results for the SPP in a noisy environment. Following that rule, the SPP is defined as [6]

$$\mathrm{SPP} = \frac{\sum_{k=0}^{11} DS^2(0,k)}{\sum_{q=0}^{Q-1}\sum_{k=0}^{11} DS^2(q,k)}, \tag{3.5}$$

and can be computed for each time block.

3.2 Periodicity Enhancement

The first class of speech enhancement algorithms in Double Spectrum to be presented relies on periodicity enhancement. The aim of this kind of algorithms is to take advantage of the property of DS that the input signal is decomposed into periodic and non-periodic components. Restoring the periodicity of noise-corrupted speech can lead to an improvement of the perceived speech quality, for instance, the pitch perception [45]. Furthermore, noise signals do not have the same periodicity characteristics as speech in many cases. This statement already implicates that this method is only reasonable for non-harmonic noise conditions. In harmonic noise environments, periodicity enhancement does not lead to substantial noise reduction.

However, weighting of coefficients in the two-stage transform domain can be used for periodicity enhancement. Huang et al. presented *fixed* and *adaptive* weighting schemes in [45, 46]. The methods were adapted to the Double Spectrum framework by Mowlaee et al. in [7].

3.2.1 Fixed Weighting

Periodicity enhancement can be implemented by adjusting the energy balance of the respective modulation bands. The fixed weighting method applies more weight to the low modulation bands (referring to periodic signal components) and less weight to higher modulation bands [6]. One method to apply fixed weighting in Double Spectrum is to define a weighting function W_q. The weight applied to a certain modulation band only depends on the modulation band index q and therefore on the modulation frequency:

$$W_q = \max\left(1 - \frac{q}{3}, 0\right), \tag{3.6}$$

According to this rule, we get a set of simply linear-decreasing weights: $W_0 = 1$, $W_1 = 2/3$, $W_2 = 1/3$, $W_q = 0$ for $q \geq 3$ [45, 46]. The weighting function is used as a gain function

$G(q,k) = W_q$ and is applied directly to the noisy double spectrum DS_x to receive the enhanced double spectrum of speech $\widehat{DS}_s$:

$$\widehat{DS}_s(q,k) = G(q,k) \cdot DS_x(q,k). \tag{3.7}$$

Using this method, the harmonic structure of voiced speech can be effectively restored [45]. Problems that occur when using the fixed weighting method are the artificial harmonization of the signal during speech pauses and strong attenuation of unvoiced components and voiced-unvoiced (VUV) transitions [6].

3.2.2 Adaptive Weighting

By the use of fixed weighting enhancement, all segments of the signal, i.e. voiced, unvoiced and silent frames, are treated equally. To overcome this issue, adaptive weight adjustment, according to the energy level of the periodic components of the signal, was proposed in [45]. Dynamic weights are applied in order to restore the periodicity of voiced speech effectively, while the aperiodic components in unvoiced speech will not be over-suppressed. To obtain a measure of voicing, the relation

$$\tilde{E}_1 = \frac{E_1}{\zeta^2} \tag{3.8}$$

is used, where E_1 is the energy of the lowest modulation band and ζ is the *root-mean-square* (RMS) of the whole signal segment. As can be seen, this measure is very similar to the modulation band ratio described in Section 3.1.1. By means of this measure, a gain function applied to the noisy signal in DS domain can be derived. The adaptive weights are dynamically adjusted according to the energy concentration of DS coefficients [6].

In Blass' master thesis [6], a method for adaptive weighting along the k and the q axis called *Adaptive Double Spectrum Weighting* (ADSW), that makes use of the joint frequency and modulation frequency dimensions, has been proposed. The enhanced DS of speech is again obtained by applying a gain function to the noisy DS:

$$\widehat{DS}_s(q,k) = G(q,k) \cdot DS_x(q,k).$$

$G(q,k)$ is a cascade of of two weighting schemes; $W_e(q,k)$ and $W_q(q,k)$. $W_e(q,k)$ is a logarithmic mapping function that de-emphasizes coefficients which are likely to carry noise information and gives more weight to coefficients that could be attributed to the speech target. It is based on the relative energy of a DS coefficient, compared to the mean energy of the noisy DS:

$$W_e(q,k) = \log_{10}(E_{rel}(q,k) + 1) \tag{3.9}$$

where the relative energy $E_{rel}(q,k)$ is defined as

$$E_{rel}(q,k) = KQ\frac{DS^2(q,k)}{E_{1:Q}}. \tag{3.10}$$

The (decadic) logarithm of E_{rel} is taken due to its broad dynamic range. The purpose of $W_q(q,k)$ is to enhance the harmonicity and periodicity of noisy speech similar to the methods of Huang et al. [45, 46]. In contrast to the fixed adaptive weighting methods, an exponentially decaying modulation weight is used for DS where the MBR is the decay factor:

$$W_q(q,k) = e^{-\mathrm{MBR}_k \cdot q}. \tag{3.11}$$

Finally, to ensure selective noise suppression, the SPP is applied as a scaling factor. The complete gain function for ADSW is therefore given as

$$G(q,k) = \mathrm{SPP} \cdot W_e(q,k) \cdot W_q(q,k). \tag{3.12}$$

3.3 Wiener Filtering

Wiener filtering is a well studied approach for noise suppression that derives the enhanced signal by optimizing a mathematically tractable error criterion, the mean-square error (MSE). Since we do not have access to the true speech and noise signals, we have to obtain their estimate. A system shall be designed in such a way that the estimated output signal $\hat{s}[n]$ is as close to the desired signal $s[n]$ as possible. This can be done by computing the estimation error $e[n] = s[n] - \hat{s}[n]$ and minimizing it. The optimal filter that fulfills this constraint is called the Wiener filter. In prinicple, Wiener filtering can be done either in the time or in the frequency domain [1].

In [6], the Wiener filter for speech enhancement was implemented in Double Spectrum. The *minimum mean-square error* (MMSE) in DS is defined as

$$\epsilon(q,k) = \mathbb{E}\left\{\left|\left|DS_s(q,k) - \widehat{DS}_s(q,k)\right|\right|_2^2\right\}. \tag{3.13}$$

Wiener filters are designed to compute a statistical estimate $\hat{s}[n]$ of a target signal $s[n]$ using the MMSE optimization criterion and they are optimal for jointly Gaussian signals in the unconstrained case [9]. Following the well-known derivation of the Wiener filter (shown e.g. in [1, 9]), the Wiener gain function $G_{\mathrm{W}}(q,k)$ is re-formulated in DS in [6]:

$$G_{\mathrm{W}}(q,k) = \frac{\xi(q,k)}{1+\xi(q,k)}. \tag{3.14}$$

In order to apply this gain function, we need to introduce a few parameters. Two important measures for statistical speech estimation are the *a priori* SNR ξ and the *a posteriori* SNR γ. Following e.g. [33, 47], the a priori SNR is defined as the ratio between the clean speech and the noise *power (modulation) spectral densities* (P(M)SD) $\xi = \sigma_s^2/\sigma_d^2$. In DS, the PMSD is the square of the magnitude of a coefficient, i.e. $\sigma_x^2 = DS_x^2(q,k)$. As a short definition, we will write $DS_{xx}(q,k)$ instead of $DS_x^2(q,k)$ for the PMSD of the noisy signal. Now we can define the a-priori SNR in DS as follows:

$$\xi(q,k) = \frac{DS_{ss}(q,k)}{DS_{dd}(q,k)} \tag{3.15}$$

The PMSDs of clean speech DS_{ss} and of noise DS_{dd} cannot be directly accessed. As a result, ξ needs to be estimated. In 1984, Eprahim and Malah [33] presented the so-called decision-directed approach. An estimate of the a-priori SNR of a time frame l, $\xi^{(l)}(q,k)$, is obtained by recursive smoothing of the previous estimate of $\xi^{(l-1)}(q,k)$ and the instantaneous maximum likelihood of the a-priori SNR. Implemented in Double Spectrum domain, the estimator is defined as follows:

$$\xi^{(l)}(q,k) = \alpha \frac{\widehat{DS}_{ss}^{(l-1)}(q,k)}{\sigma_d^{2,(l)}(q,k)} + (1-\alpha)\max\left(\gamma^{(l)}(q,k) - 1, 0\right), \tag{3.16}$$

with the approximation of the posterior SNR

$$\gamma^{(l)}(q,k) = \frac{DS_{xx}^{(l)}(q,k)}{\sigma_d^{2,(l)}(q,k)}. \tag{3.17}$$

α is a forgetting factor or smoothing parameter for recursive averaging. $\sigma_d^{2,(l)}$ is the realization of the noise PMSD estimate – noise estimation in Double Spectrum will be discussed in detail in Chapter 4. The estimate of the instantaneous SNR is obtained by an approximation of the posterior SNR subtracted by 1. The instantaneous SNR is used to update the prior SNR estimate. Using the temporal smoothed approach for $\xi^{(l)}(q,k)$ leads to effective suppression of musical noise [33]. The degree of noise suppression depends on the choice of α, where a large value provides good elimination of musical noise, but, on the other hand, makes the estimate insensitive for rapid changes of the prior SNR as happening during speech onsets. Commonly, the forgetting factor is chosen in the range $\alpha \in [0.9, 0.98]$ (e.g. in [48, 49]).

3.4 MMSE Speech Estimator in DS

An MMSE based approach for estimating the spectral amplitudes of Fourier coefficients has been firstly proposed by Ephraim and Malah (1984) [33], applied in the STSA domain.

Therefore, it is often called the MMSE-STSA algorithm. In distinction to the Wiener filter, the MMSE-STSA approach is optimal in terms of estimating the spectral *amplitude*, directly derived from the noisy observation of speech. That is, the Wiener filter minimizes the MSE of the complex spectrum while the MMSE-STSA algorithm minimizes the MSE of the spectral amplitudes.

The method serves very commonly as a benchmark for newly proposed noise suppression approaches. Motivated by the performance of the MMSE-STSA method, Paliwal et al. (2011) [49] applied the algorithm as a spectral modulation magnitude estimator in the STSM domain.

The estimator shall now be reformulated in the DS domain. The proposed estimator minimizes the MSE between the amplitude of DS coefficients of clean and estimated speech, respectively:

$$\epsilon = \mathbb{E}\left\{\left(\left|DS_s^{(l)}(q,k)\right| - \left|\widehat{DS_s}^{(l)}(q,k)\right|\right)^2\right\}. \tag{3.18}$$

A closed form solution for this approach was derived in [33] under the assumptions that speech and noise are i.i.d. zero-mean Gaussian random variables and that speech and noise are additive in time domain. Exploiting the additivity property of the Double Spectrum, this constraint also holds in DS domain. Presuming the above assumptions, $\widehat{DS}_s$ can be estimated under the MMSE criterion as:

$$\widehat{DS_s}^{(l)}(q,k) = G^{(l)}(q,k) \cdot DS_x^{(l)}(q,k), \tag{3.19}$$

where $G^{(l)}(q,k)$ is the gain function given by

$$G^{(l)}(q,k) = \frac{\sqrt{\pi}}{2} \frac{\sqrt{\nu^{(l)}(q,k)}}{\gamma^{(l)}(q,k)} \Lambda\left[\nu^{(l)}(q,k)\right] \tag{3.20}$$

with

$$\nu^{(l)}(q,k) = \frac{\xi^{(l)}(q,k)}{1+\xi^{(l)}(q,k)} \gamma^{(l)}(q,k) \tag{3.21}$$

and

$$\Lambda[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta) I_0\left(-\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right]. \tag{3.22}$$

$I_0(\cdot)$ and $I_1(\cdot)$ are modified Bessel functions of zeroth and first order, respectively. The a priori SNR is estimated by recursive averaging using the decision-directed approach as defined in Equation (3.16). The a posteriori SNR is defined in Equation (3.17).

3.5 Derivation of a MAP Speech Estimator in DS

In this section, a maximum a posteriori (MAP) estimator for speech, based on the statistical distributions of speech and noise in Double Spectrum, will be derived. In order to do this, firstly a *probability density function* (PDF) will be fitted to a histogram of the DS coefficients of clean speech. This PDF with the corresponding parameters will be further used to derive a speech estimator by applying the MAP estimation rule.

3.5.1 Derivation

The histograms of the DS coefficients of clean speech were created using the training set of 72 different sentences from the *TIMIT* database [43], spoken by male and female speakers, sampled at 16 kHz. The DS coefficients were normalized by the square-root of the variance $\sigma_s^{2,(l)}(q,k)$, estimated via recursive averaging, using a variation of the decision-directed approach [50]:

$$\sigma_s^{2,(l)}(q,k) = \alpha\, \sigma_s^{2,(l-1)}(q,k) + (1-\alpha)\, \widehat{DS}_{ss}^{(l)}(q,k) \tag{3.23}$$

with the smoothing parameter chosen as $\alpha = 0.98$. Furthermore, the data was thresholded by a value of -40 dB so that silent parts of the speech file were not taken into account. The DS coefficients were evaluated separately for each modulation band q for the total number of modulation bands $Q = 4$. The samples that have been used to create the histograms therefore have the same mean and variance.

The histograms were fitted with three different kinds of distributions: Gaussian, Laplacian and Gamma. As a goodness-of-fit test, the one-dimensional *Kolmogorov-Smirnov* (K-S) test was used [48, 51]. Since the speech DS coefficients are symmetrically distributed around zero, it is sufficient to investigate the distribution of the absolute values

$$\left|\overline{DS}_s^{(l)}(q)\right| = \frac{\left|DS_s^{(l)}(q)\right|}{\sigma_s^{(l)}(q)}, \tag{3.24}$$

as the sign of the noisy double spectrum will be preserved, i.e. $\mathrm{sgn}(\widehat{DS}_s) = \mathrm{sgn}(DS_x)$. The fitting of the distributions was performed by the *MATLAB*-function `fitdist.m`. As depicted in Figure 3.1, the Gamma distribution has the closest fit to the histogram data. Based on that result, a Gamma PDF will be used as prior distribution of clean speech coefficients. It is noticeable that the distribution of DS coefficients is almost identical across all modulation bands q when speech is analyzed over a long time span.

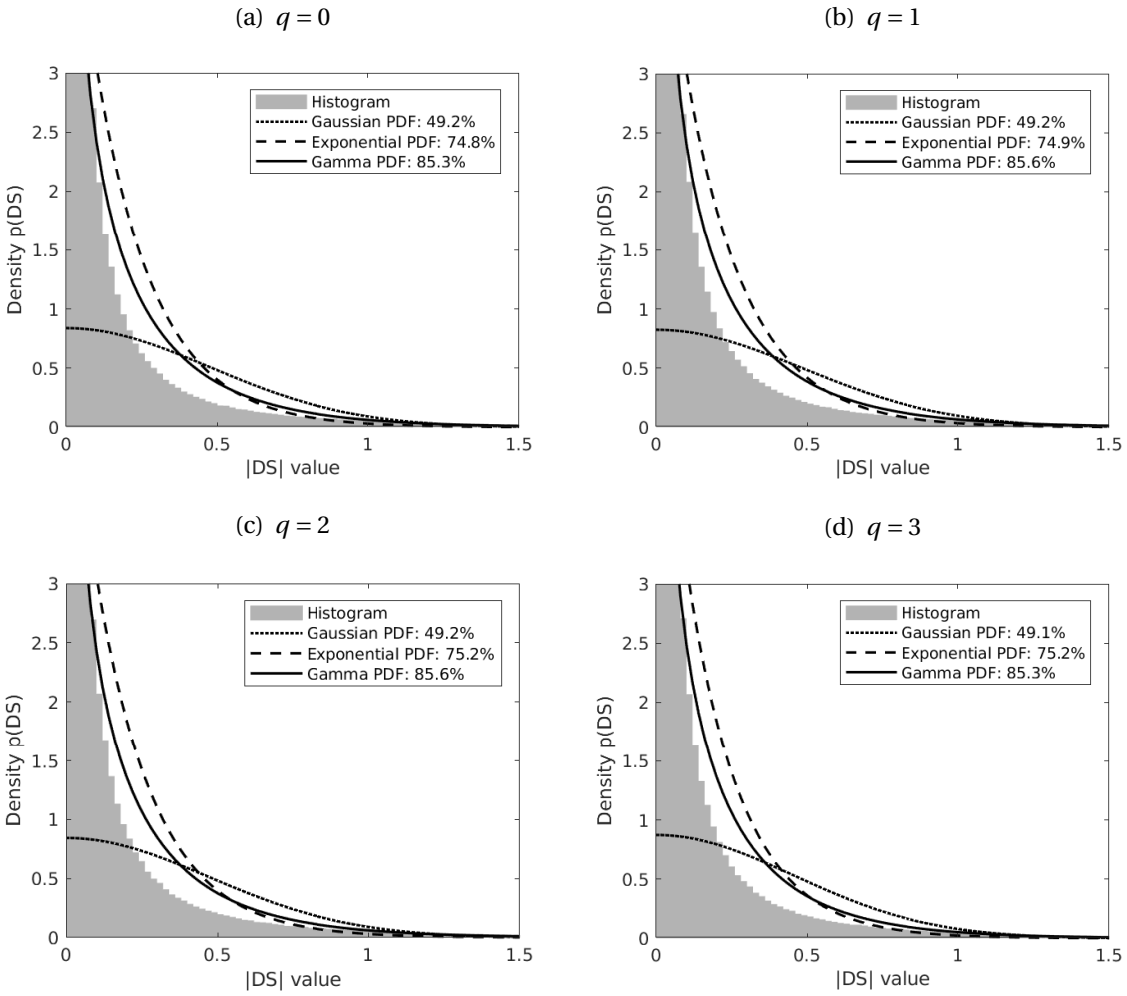


FIGURE 3.1. Histograms of normalized clean speech DS coefficients and fitted PDFs per modulation band q. The goodness-of-fit [%] is measured by the K-S test.

The Gamma PDF is defined as follows:

$$p(x|a,b) = \frac{1}{b^a \Gamma(a)} x^{a-1} \exp\left(-\frac{x}{b}\right), \tag{3.25}$$

where a and b are the scale and shape parameters of the Gamma PDF, respectively, and $\Gamma(\cdot)$ is the Gamma function. Table 3.1 shows the values for a and b that have maximized the result of the goodness-of-fit test.

We now introduce the statistical model for the speech and noise DS coefficients. As has been shown above, we use a Gamma PDF for the speech prior distribution. For the sake of readability, the time frame index l and the modulation band index q will be omitted in the following derivation. The calculations hold for each frame and modulation band separately.

Modulation band q	a	b
0	0.593	0.334
1	0.591	0.345
2	0.595	0.334
3	0.599	0.313

TABLE 3.1. Fitting parameters a and b for the Gamma PDF, for clean speech DS coefficients.

Furthermore, absolute values are assumed, i.e. $DS_s \geq 0$. According to the definition (3.25), the normalized speech prior distribution is defined as follows:

$$p\left(\frac{DS_s}{\sigma_s}\right) = \frac{1}{b^a \Gamma(a)} \left(\frac{DS_s}{\sigma_s}\right)^{a-1} \exp\left(-\frac{DS_s}{b\sigma_s}\right). \tag{3.26}$$

A distribution model for the noise DS coefficients is also required. According to the central limit theorem, it can be said that noise signals that are rather uncorrelated with themselves will converge towards a Gaussian distribution, regardless of the PDF of the time samples [52, 53]. We take this assumption and use the Gaussian distribution for the noise DS coefficients:

$$p(DS_d) = \frac{1}{\sqrt{2\pi\sigma_d^2}} \exp\left(-\frac{(DS_d - \mu)^2}{2\sigma_d^2}\right), \tag{3.27}$$

where μ is the mean value and σ_d^2 is the variance of the noise.

Based on the distributions of speech and noise, a MAP solution following

$$\widehat{DS}_s = \underset{DS_s}{\arg\max}\, p(DS_s|DS_x) = \underset{DS_s}{\arg\max}\, \frac{p(DS_x|DS_s)\,p(DS_s)}{p(DS_x)} \tag{3.28}$$

can be found. Since $p(DS_x)$ is independent from DS_s, we only need to maximize the numerator $p(DS_x|DS_s)\,p(DS_s)$ [53]. Using the distributions from (3.26) and (3.27), we obtain the likelihood

$$p(DS_x|DS_s)\,p(DS_s) = \frac{1}{b^a \Gamma(a)\sqrt{2\pi\sigma_d^2}} \left(\frac{DS_s}{\sigma_s}\right)^{a-1} \exp\left(-\frac{(DS_x - DS_s)^2}{2\sigma_d^2} - \frac{DS_s}{b\sigma_s}\right). \tag{3.29}$$

As the logarithm function is a monotonically increasing function, we can alternatively maximize the logarithm of (3.29) to simplify the term, so that the product of the polynomial and exponential converts into a sum [1, 53]:

$$\log\left[p(DS_x|DS_s)\,p(DS_s)\right] = -\log\left(b^a \Gamma(a)\sqrt{2\pi\sigma_d^2}\right) + (a-1)\log\left(\frac{DS_s}{\sigma_s}\right) - \frac{(DS_x - DS_s)^2}{2\sigma_d^2} - \frac{DS_s}{b\sigma_s}.$$

(3.30)

To obtain a maximum, the derivative of the expression (3.30) is set to zero. This leads to a quadratic function for DS_s:

$$\begin{aligned} \frac{\partial}{\partial DS_s}\log[p(DS_x|DS_s)p(DS_s)] &= \frac{a-1}{DS_s}+\frac{DS_x-DS_s}{\sigma_d^2}-\frac{1}{b\sigma_s}\stackrel{!}{=}0 \\ \Rightarrow \quad DS_s^2+\left(\frac{\sigma_d^2}{b\sigma_s}-DS_x\right)DS_s-\sigma_d^2(a-1)&=0 \end{aligned} \tag{3.31}$$

Solving (3.31) for DS_s by using the definitions of the a priori SNR $\xi=\sigma_s^2/\sigma_d^2$ and the a posteriori SNR $\gamma=DS_x^2/\sigma_d^2$, we finally obtain the MAP speech estimator as the product of a gain function G_{MAP} and the noisy Double Spectrum DS_x (the positive sign is used in front of the square root to ensure positive values of G_{MAP}):

$$\begin{aligned} &\widehat{DS_s}=G_{\mathrm{MAP}}\cdot DS_x, \\ &G_{\mathrm{MAP}}(\xi,\gamma,a,b)=\frac{u}{2}+\sqrt{\left(\frac{u}{2}\right)^2+\frac{a-1}{\gamma}}, \qquad u=1-\frac{1}{b\sqrt{\xi\gamma}}. \end{aligned} \tag{3.32}$$

3.5.2 Analysis of DS-MAP Gain Function

By having a closer look at G_{MAP}, we can see that there are combinations of a, b, ξ and γ possible where the argument under the square root becomes complex and so does the gain function, as a result. Since DS coefficients are real-valued, we need to avoid this case. Figure 3.2 shows the region on the ξ,γ plain where G_{MAP} becomes complex for $a=0.593, b=0.334$ (white area). The size of the complex area is a function of the parameters a and b.

In order to avoid complex values of G_{MAP}, several methods have been taken into account:

(i) Choose $a>1$ – the argument under the square root cannot turn negative anymore;

(ii) Take the absolute value of complex G_{MAP} and ignore the phase term;

(iii) Take the absolute value of complex G_{MAP} and use the projection of the phase on the real axis as sign of the gain function:

$$\mathrm{sgn}(G_{\mathrm{MAP}})=\mathrm{sgn}\left(\arctan\left(\angle G_{\mathrm{MAP}}\right)\right); \tag{3.33}$$

(iv) apply the lower gain bound $G_{\min}$ where G_{MAP} is complex.

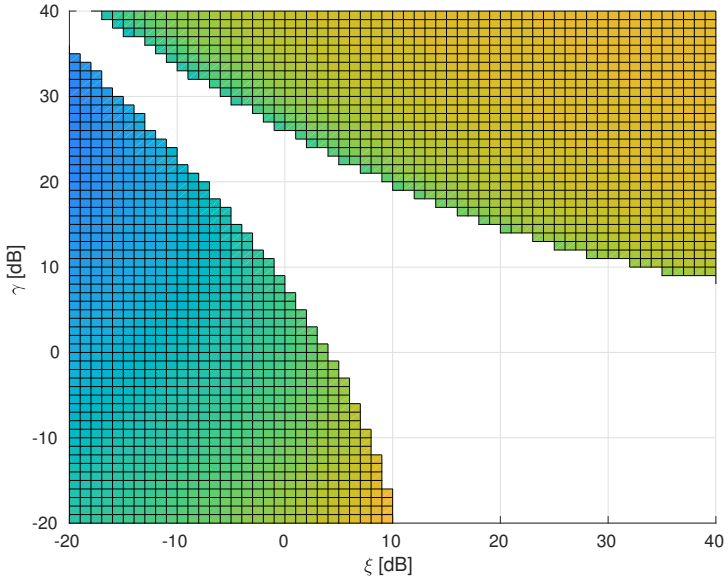


FIGURE 3.2. Visualization of $G_{\mathrm{MAP}}(\xi, \gamma, a = 0.593, b = 0.334)$; the white area corresponds to complex values of the gain function.

Method (i) would be the mathematically most solid solution, since the gain function does not have to be modified at all. On the other hand, by changing the value of a, the goodness-of-fit of the histogram is affected. Values of $a > 1$ lead to strong attenuation of large DS speech values and therefore affect intelligibility. Consequently, method (ii) was chosen empirically, as it provides the best results in terms of speech quality and intelligibility. As a result, we use the relation

$$\widehat{DS_s} = |G_{\mathrm{MAP}}| \cdot DS_x. \tag{3.34}$$

Figure 3.3(a) shows the absolute value of the MAP gain (3.32) as a function of the a priori SNR ξ and the a posteriori SNR γ. Figure 3.3(b) visualizes the corresponding phase of G_{MAP}. For reasons of comparison, the Wiener gain (3.14) function is depicted in Figure 3.3(c) and the MMSE gain (3.19) is shown in Figure 3.3(d). It is remarkable that in contrast to the other speech estimators, the Wiener gain is only a function of the a priori SNR ξ, i.e. it is independent of the posterior SNR γ.

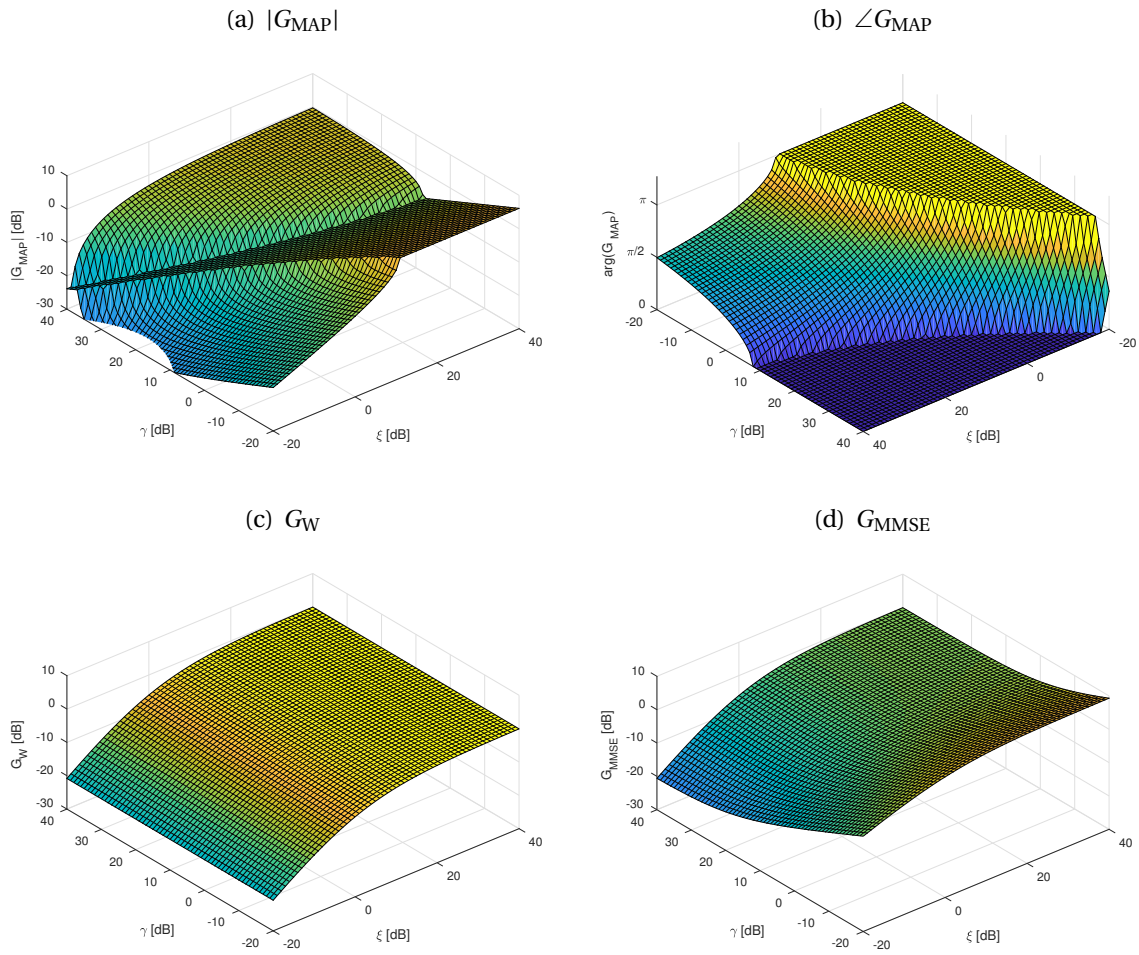


FIGURE 3.3. $G_{\mathrm{MAP}}(\xi, \gamma, a = 0.593, b = 0.334)$ (a) $|G_{\mathrm{MAP}}|$, (b) $\angle G_{\mathrm{MAP}}$, (c) Wiener gain and (d) MMSE gain function.

3.6 Modulation Filtering / RASTA Filters

A different approach of speech enhancement in modulation domain is the so-called RASTA (RelAtive SpecTrA) filter. It belongs to the class of modulation filtering in STSA domain (see Section 2.3.1). It was presented by Hermansky (1994) [54] and was further developed in [55]. The RASTA filter takes advantage of the fact that non-linguistic components of a signal often have a slower rate of change than speech and it can be applied for denoising as well as for dereverberation. For this purpose, it does not need any noise PSD estimation, which makes the method simple and computationally efficient in comparison to stochastic techniques and independent of any voice activity detection.

The RASTA procedure is implemented as follows according to [54]. First, the noisy magnitude is squared to obtain the power spectrum. Subsequently, it gets compressed by a static nonlinearity. Several different nonlinearities can be used for this purpose. For dereverberation of speech, often the natural logarithm is used as compression stage since convolutive elements in time domain become additive in the logarithmic spectrum. From the linearity of the DFT it follows that applying RASTA to the logarithmic spectrum is equivalent to modulation-filtering of the cepstral coefficients [56]. However, uncorrelated additive noise components become signal dependent after the logarithmic operation on the spectral magnitude and therefore cannot be effectively removed by RASTA bandpass filtering. Alternatively, when using the RASTA filter for de-noising, the *lin-log* approach can be applied as compression stage:

$$y = \log(1 + Jx) \tag{3.35}$$

where J is a signal-dependent positive constant. The amplitude-warping transform is linear-like for $J \ll 1$ and logarithmic-like for $J \gg 1$ [54]. While the logarithmic compression is more suitable for processing convolutive distortions, the cubic root of the power spectrum has been found to provide the best results for additive noise cancellation [57, 58].

$$\left|X_{\text{comp}}(k,m)\right| = |X(k,m)|^{\frac{2}{3}}. \tag{3.36}$$

The time trajectory of each (compressed) frequency bin is filtered by the RASTA filter:

$$\left|X_{\text{comp,enh}}(k,m)\right| = \sum_{p=-\infty}^{\infty} h(k,m-p)\left|X_{comp}(k,p)\right|, \tag{3.37}$$

with the transfer function of the RASTA filter being defined as follows:

$$H(z) = 0.1z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.94z^{-1}}. \tag{3.38}$$

Finally, the filtered data is expanded by the complementary static nonlinearity to gain the enhanced spectral amplitude. The result is combined with the noisy sign to obtain the enhanced spectrum. In order to avoid complex spectral amplitudes, any negative output of the filter is substituted by zero (half-wave rectification) [55]. Please note that RASTA processing increases the dependence of the data on previous frames. Hence, some audible reverberation is introduced by the filter in the output speech file [54]. The AMS framework for RASTA filtering is depicted in Figure 3.5.

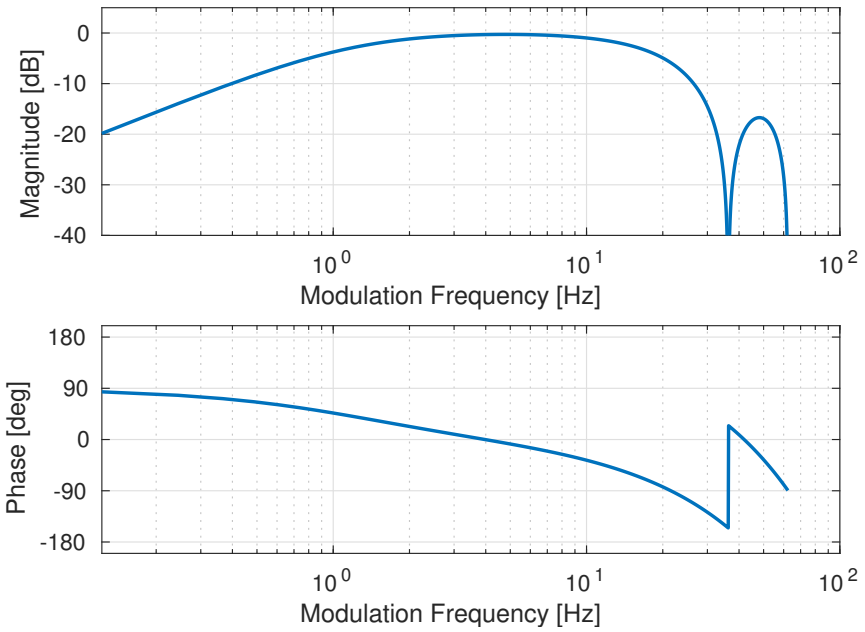


FIGURE 3.4. Frequency response $H(z)$ of the RASTA filter.

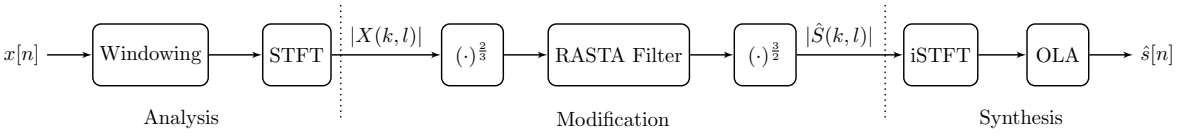


FIGURE 3.5. Block diagram of the AMS framework for RASTA filtering of speech.

3.7 Role of the Sign in DS

Up to this point, we have proposed several methods for enhancing the magnitude of Double Spectra, while the sign information has been neglected, since we have always combined the enhanced magnitude with the sign information of the noisy DS. It will be shown in this section that the sign carries important information of speech. As a result, by performing sign-aware DS speech enhancement, the above discussed methods can be further improved.

DS coefficients are always real-valued [6]. Therefore, there is no continuous phase information available. The sign of DS coefficients can be interpreted as projection of the phase information onto the real axis. Consequently, we have the following possibilities for the phase value: $\varphi = k \cdot \pi,\ \forall\ k \in \mathbb{Z}$.

3.7.1 Proof of Concept

The importance of the sign information can be pointed out by comparing enhanced speech obtained by the DS-MAP estimator (see Section 3.5) in combination with clean and noisy sign information. The clean and the noisy speech signals undergo the DS analysis stage using the same time block segmentation. The noisy DS magnitude is then enhanced with the DS-MAP estimator and combined with the noisy sign and with the sign obtained from the clean speech file, respectively. Figure 3.6 shows the potential of sign enhancement in factory noise environment. The blue bar indicates the improvements in speech quality (ΔPESQ) and intelligibility (ΔSTOI) of the enhanced speech obtained by the magnitude-only MAP estimator. The red bar shows the combination of the MAP enhanced magnitude with the sign of the clean speech DS coefficients (*oracle sign* scenario). Table 3.2 contains the mean values of ΔSTOI and ΔPESQ. As one can see, intelligibility as well as speech quality increase when the sign information of clean speech DS is preserved. These results motivate to implement an additional sign processing stage in the DS speech enhancement framework.

	ΔPESQ			ΔSTOI		
SNR [dB]	0	5	10	0	5	10
noisy sign	0.36	0.41	0.40	0.00	0.00	0.00
oracle sign	0.58	0.59	0.56	0.04	0.03	0.02

TABLE 3.2. Mean improvements in speech quality (ΔPESQ) and speech intelligibility (ΔSTOI) over SNR using DS-MAP for noisy and oracle sign scenario, in *factory noise* environment.

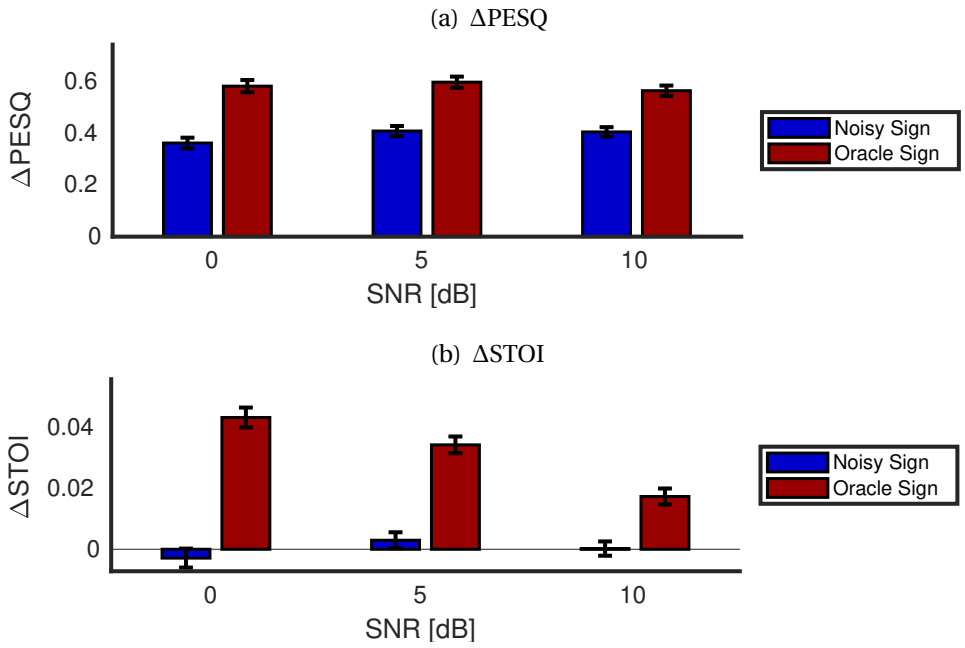


FIGURE 3.6. Proof of concept: improvements in (a) speech quality and (b) speech intelligibility of DS-MAP enhanced speech, over SNR, for noisy and oracle sign scenarios, in *factory noise* environment.

3.7.2 Smoothing of DS Trajectories

Figure 3.7 depicts the trajectory of the first modulation band $q = 0$ of the frequency band $k = 1$. In the pitch-synchronous DS framework, the frequency bin with the index $k = 1$ is always represents f_0. When speech is active, a great amount of speech energy is concentrated in this band as it represents the periodic content of the fundamental frequency band. The top plot (a) shows the temporal trajectory of DS values over time of a clean speech signal. The middle plot (b) shows the trajectory of the same speech signal, but corrupted with factory noise at an SNR of 5 dB. The DS values are normalized to $\max(DS) = 1$. Eventually, the bottom subfigure (c) shows the spectrogram and the MBR6th measure (see Section 3.1.2). As one can see, the trajectories of the clean speech DS values have a smooth course and a stable polarity during voiced frames. This *harmonic phase relation* [59] can be used for smoothing the trajectories. In simple terms, we can say that the polarity in those bands remains constant when voiced speech is active (i.e., the MBR remains on a high level). The speech presence probability can be determined by the use of the current MBR value. Stronger smoothing can be applied to regions where the MBR is high. Furthermore, the smoothness of the trajectories is dependent on the modulation band. The trajectories of clean speech

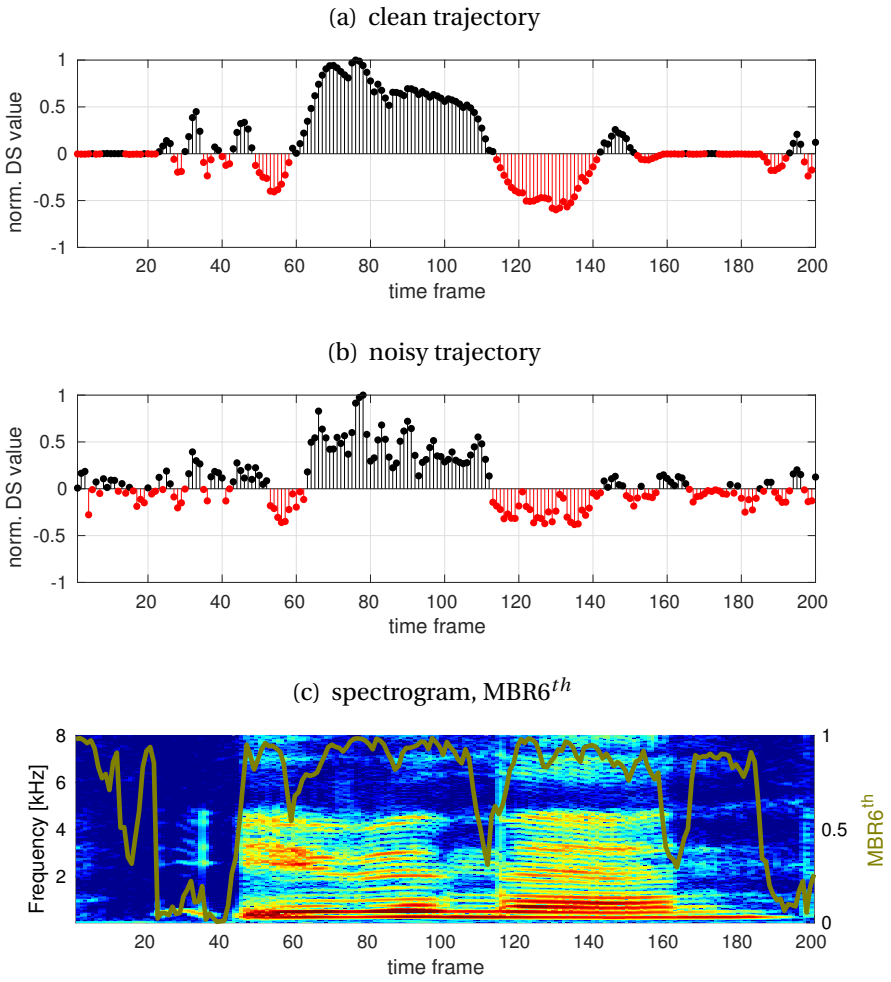


FIGURE 3.7. DS trajectories of $q = 0$ and $k = 1$ of (a) clean speech and (b) noisy speech (*factory noise* @ 5dB). (c) shows the spectrogram of clean speech and the MBR6th value.

DS values become less smooth when q increases. Thus, we will apply smoothing mainly to the lowest two modulation bands.

Interpolation filter

The DS trajectories can be smoothed using an interpolation filter. The method works as follows. First, DS-synthesis is performed on the output of the amplitude enhancement stage to obtain a time signal. This signal is subsequently re-analyzed to DS in order to gain a new representation based on the amplitude-modified speech. As can be seen in the proof of concept above, the trajectories of DS coefficients tend to have a stable polarity and a smooth course when there are stablely voiced areas. Therefore, it is worthwhile to make the amount

of smoothing dependent on the current MBR, since a high MBR indicates strongly voiced regions. To this end, we separate the DS frames into regions of stable polarity. The regions are classified into voiced and unvoiced ones; when the average MBR measure is greater than a threshold, the region is declared as *voiced*, otherwise as *unvoiced*. Subsequently, the regions are smoothed separately with a *Savitzky-Golay-Filter* [60], a polynomial interpolation filter of third order. Since DS trajectories of voiced areas, especially in low modulation bands, are assumed to be smoother, more data points are taken into account for smoothing. As a result, (i) the classification of a region with stable polarity and (ii) the modulation band index are determining the strength of smoothing, i.e., the filter coefficients.

One problem that occurs with this method is the binary VUV classification of regions. This leads to sudden changes of the amount of smoothing at VUV transitions and as a result to audible artifacts in the speech signal. To overcome this problem, an overlap is introduced to alleviate this effect; some additional DS frames of neighboring regions are taken into account for smoothing as well.

Recursive smoothing

The DS trajectory smoothing stage can also be performed by a *recursive smoothing* (RS) algorithm. When using this method, the VUV separation is not necessary. The RS algorithm is defined as follows:

$$\overline{DS}^{(l)}(q,k) = \alpha^{(l)}(q) \cdot \overline{DS}^{(l-1)}(q,k) + \left(1 - \alpha^{(l)}(q)\right) \cdot DS^{(l)}(q,k). \tag{3.39}$$

The smoothing factor $\alpha^{(l)}(q)$ has to be adaptive and dependent on the MBR: $\alpha^{(l)} = f\left(\mathrm{MBR}^{(l)}\right)$. This is realized with a logistic function (3.40). As before, regions with a high MBR are smoothed more strongly than ones with a low MBR. Simultaneously, $\alpha^{(l)}(q)$ also depends on the modulation band. Low modulation bands are smoothed more strongly than high ones. The logistic function is defined as follows:

$$\alpha^{(l)}(q) = \left(\frac{1}{1 + \exp\left(-(\mathrm{MBR}_{\max} - \mathrm{MBR}_{\min}) \cdot \mathrm{MBR}^{(l)}\right)} + \mathrm{MBR}_{\min} \right) \cdot \beta(q), \tag{3.40}$$

where $\beta(q)$ is a value between 0 and 1, determining the amount of smoothing in a certain modulation band q.

CHAPTER 4

NOISE ESTIMATION

So far, various statistically based speech estimators in Double Spectrum have been presented. Up to this point, we have assumed a reliable estimation of the noise *power spectral density* (PSD) being available. In reality, the noise PSD has to be estimated from the noisy signal. For a good performance of a speech enhancement algorithm, a reliable estimation of the noise PSD is essential since it highly affects the resulting speech quality and intelligibility. To give an example, the noise PSD is required to compute the a priori SNR ξ that is the key parameter for Wiener filtering and other statistically based speech estimators. When the estimate of the noise PSD is too low, annoying residual noise will be audible; if it is overestimated, speech will be distorted and this will possibly result in loss of intelligibility [1].

In this chapter, three different noise estimators in the Double Spectrum framework will be presented. The Double Spectrum is a modulation based framework; therefore we talk about the *Power Modulation Spectral Density* (PMSD) instead of PSD in this domain. The concept of recursive PMSD smoothing in DS and an adaptive forgetting factor will be proposed. At the end of the chapter, the noise tracking performance of the proposed noise estimators will be evaluated.

4.1 Recursive Smoothing of Noise PMSD

4.1.1 Interpolation of DS Frames

For the noise estimators that will be presented below, we require recursive smoothing of the noise PMSD in order to obtain a smooth estimate and to avoid musical noise artefacts. Since the Double Spectrum is a pitch synchronous framework, subsequent frames may have different dimensions, meaning that every DS frame has a different number of fre-

quency channels $K = \tilde{P}_0$, where $\tilde{P}_0 = f_s / f_0$ is the instantaneous pitch period in samples. Furthermore, the number of modulation bands Q is held constant by using a "static" time block segmentation (TBS) for the modulation transform, resulting in a constant number of frames per time block [6]. Having DS frames of different sizes makes temporal smoothing of the single *frequency-modulation frequency* (F-MF) bands difficult. Therefore, we need to "stretch" the respective DS frames to a uniform size K_{global}. We can now compute a smoothed global estimate of the noise PMSD η^2:

$$\eta^{2,(l)}(q,k') = \lambda\, \eta^{2,(l-1)}(q,k') + (1-\lambda)\, \mathcal{I}\left\{ DS^{(l)}_{xx,\text{global}}(q,k') \right\}, \tag{4.1}$$

where λ is a forgetting factor chosen depending on the stationarity of noise and $\mathcal{I}\{\cdot\}$ denotes an operator that performs cubic interpolation along the frequency axis k, for each modulation band q, in order to set the current DS frame to the required size $K_{\text{global}} \times Q$ of the global estimate. After the recursive smoothing stage, the estimate $\eta^{2,(l)}(q,k')$ is interpolated back to its original dimension $K \times Q$ for further processing:

$$\sigma_d^{2,(l)}(q,k) = \mathcal{I}^{-1}\left\{ \eta^{2,(l)}(q,k') \right\}. \tag{4.2}$$

Altough small errors are introduced to the noise estimate by this method, they are negligibly small compared to those caused by the noise estimation process itself [6].

4.1.2 Adaptive Forgetting Factor for Recursive Noise PMSD Smoothing

One method for a possible improvement of the noise tracking algorithms in terms of recursive smoothing is the use of an adaptive forgetting factor. It can be shown that the forgetting factor λ is of great importance for the resulting quality and intelligibility of enhanced speech. The higher the value of λ, the less musical noise is introduced, but on the other hand, the estimate is less sensitive to rapid changes of the noise environment.

Typically, static values between $0.9 < \lambda < 1$ are chosen for the forgetting factor, depending on the current estimate of the a-priori SNR and the stationarity of noise. As ξ changes over time, λ needs to be adapted accordingly. For that purpose, an adaptive forgetting factor using a logistic function in logarithmic domain is introduced:

$$\lambda(\xi) = \frac{\lambda_{max} - \lambda_{min}}{1 + \exp\left(-\frac{\xi_{max} - \xi_{min}}{180} \cdot \xi \right)} + \lambda_{min}. \tag{4.3}$$

Empirically, ranges of $\lambda_{min} = 0.95 < \lambda < \lambda_{max} = 0.999$ and $\xi_{min} = -25\ \text{dB} < \xi < \xi_{max} = 25\ \text{dB}$ are chosen. The adaptive forgetting factor gives more weight to the current update of the

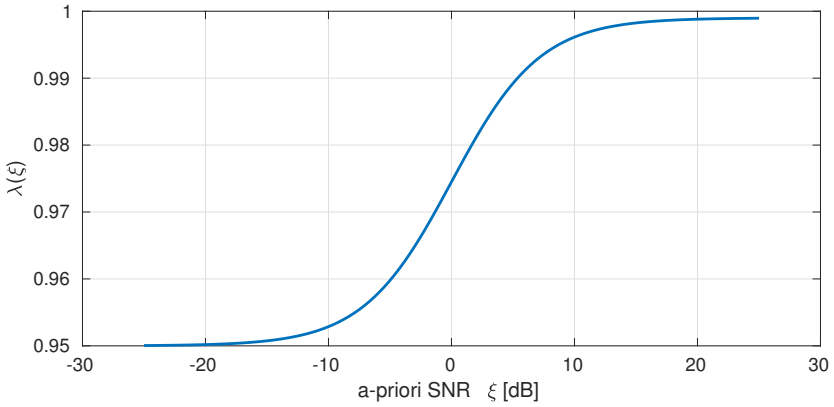


FIGURE 4.1. Logistic function of the adaptive forgetting factor $\lambda(\xi)$ for noise PMSD estimation.

noise PMSD, the lower the current estimate of the a priori SNR is. This leads to a dynamic noise PMSD update and to more flexibility of the noise tracking algorithm under noise conditions with varying stationarity. The function is depicted in Figure 4.1.

4.2 VAD based Noise Estimator

Voice activity detection (VAD) is a method for discriminating between regions with speech presence ($\mathcal{H} = 1$) and speech absence ($\mathcal{H} = 0$). VAD based algorithms typically determine voice activity based on some feature of the signal, e.g. the short-time energy or the zero-crossing rate of the input signal that is compared against a threshold value to make a binary decision for every frame. VAD based algorithms are popular due to their simplicity and thus their computational efficiency [1]. It should be noted that the majority of VAD based algorithms become unreliable under low-SNR conditions, especially in non-stationary noise environments, like babble or street noise.

A simple noise estimator has been implemented in DS based on the speech presence probability (SPP, see Section 3.1.2). The noise PMSD is estimated as follows. First, an initial estimate of the noise is computed from leading frames without speech activity. The estimate is then updated during speech absent frames using recursive smoothing (with the interpolation method from Equation (4.1)):

$$\sigma_d^{2,(l)}(q,k) = \begin{cases} \lambda\, \sigma_d^{2,(l-1)}(q,k) + (1-\lambda)\, DS_{xx}^{(l)}(q,k), & \mathcal{H} = 0 \\ \sigma_d^{2,(l-1)}(q,k), & \mathcal{H} = 1 \end{cases} . \tag{4.4}$$

$\sigma_d^{2,(l)}(q,k)$ denotes the smoothed noise PMSD estimate with dimension $K \times Q$. The decision whether the current frame is treated as speech present or speech absent frame is taken by an estimate of the SPP (see Section 3.1.2). A binary decision on voice activity is made, depending on the SPP value being above or below a certain threshold θ:

$$\mathcal{H} = \begin{cases} 0, & \text{SPP} < \theta \\ 1, & \text{SPP} > \theta \end{cases} \tag{4.5}$$

Based on this voice activity detection, the noise estimate is either updated or left unchanged at a certain frame l.

4.3 MMSE Noise Estimator using SPP

The VAD based noise estimator presented above can be improved by replacing the hard decision of VAD by a soft decision using an SPP estimator. A similar method was presented by Gerkmann and Hendriks (2011) [61] in the STSA domain. For this work, the estimator was adopted to Double Spectrum adequately. This MMSE based estimator for the noise power under speech presence uncertainty in DS is defined as follows:

$$\begin{aligned} \sigma_d^{2,(l)}(q,k) &= \mathbb{E}\left\{\sigma_d^{2,(l-1)}(q,k) \mid DS_x^{(l)}(q,k)\right\} \\ &= \mathbb{P}\left(\mathcal{H}=0 \mid DS_x^{(l)}(q,k)\right) \cdot DS_{xx}^{(l)}(q,k) + \mathbb{P}\left(\mathcal{H}=1 \mid DS_x^{(l)}(q,k)\right) \cdot \sigma_d^{2,(l-1)}(q,k), \end{aligned} \tag{4.6}$$

where $\mathbb{P}\left(\mathcal{H}=0 \mid DS_x^{(l)}(q,k)\right)$ indicates the probability of speech absence when the noisy DS is given, and $\mathbb{P}\left(\mathcal{H}=1 \mid DS_x^{(l)}(q,k)\right)$ is the probability of speech presence. $\sigma_d^{2,(l-1)}(q,k)$ is the estimated noise PMSD, obtained by recursive smoothing of the noise power of previous frames, as shown in Equation (4.4), while $\sigma_d^{2,(l)}(q,k)$ is the current noise estimate of the frame l. $\mathbb{E}\{\cdot\}$ is the operator for the expected value and $\mathbb{P}\{\cdot\}$ is the probability operator. As a result, the estimator is a weighted sum of the noisy observation $DS_{xx}^{(l)}(q,k)$ and the previous estimate of the spectral noise power $\sigma_d^{2,(l-1)}(q,k)$ [61].

4.4 Minimum Statistics

The *minimum statistics* (MS) approach for noise PSD estimation was firstly employed by Martin (2001) [62] in the STSA domain. The advantage of this method is that it does not require any kind of voice activity detection or speech presence probability estimation. It

tracks spectral minima of each DS coefficient without any distinction between speech activity and speech pause. The approach works as follows: the smoothed *periodograms* $\overline{DS}_{xx}$ of M subsequent frames are stored in a vector $\mathbf{T}$, using a sliding window:

$$\mathbf{T} = \left\{ \overline{DS}_{xx}^{(l-M+1)}, ..., \overline{DS}_{xx}^{(l)} \right\}. \tag{4.7}$$

An estimate of $DS_{dd}^{(l)}(q,k)$ is obtained by taking the minimum T_{min} out of the vector $\mathbf{T}$. Since it is assumed that the power level of each frequency and modulation band regularly falls back to the level of noise, minima tracking leads to an estimate of the noise PMSD even during speech active parts. Taking M sufficiently large, it is guaranteed that T_{min} originates from a speech absent frame. With additional recursive smoothing of the noise power (4.1), (4.2), the noise estimate $\sigma_d^{2,(l)}(q,k)$ is obtained. Since the algorithm always tracks minima, the estimate of the noise PMSD is pushed below its mean. Therefore, bias compensation is required. This was implemented according to [63]. The choice of the sliding window length is always a trade-off. On the one hand, it must be ensured that the window is long enough to ensure that the tracked minimum in actual fact belongs to a speech absent portion of speech. On the other hand, when the sliding window length is selected too long, the algorithm becomes highly insensitive to rapid changes of the noise PMSD and lags behind its actual current value. Thus, a careful choice of the sliding window length is crucial for the performance of the MS algorithm.

4.5 Evaluation of Noise Estimators

The tracking performance of the noise estimators was investigated using the logarithmic mean of a spectral distance measure, referring to Taghia et al. [64]. The (extended) *modulation periodograms* (MP) of a reference noise signal and the noise MP estimated from noisy speech were compared. The reference noise PMSD and the estimated noise PMSD are denoted by $\eta_d^{2,(l)}(q,k')$ and $\hat{\eta}_d^{2,(l)}(q,k')$, respectively. The evaluation measure for the tracking error is the average logarithmic distance $LogErr_{\text{mean}}$ between the estimated and reference noise PMSDs:

$$LogErr_{\text{mean}} = \frac{1}{QK_{\text{global}}L} = \sum_{q=0}^{Q-1} \sum_{k'=0}^{K_{\text{global}}-1} \sum_{l=0}^{L-1} \Delta_{\text{dB}}^{(l)}(q,k'), \tag{4.8}$$

with

$$\Delta_{\text{dB}}^{(l)}(q,k') = \left| 10 \cdot \log_{10} \left[\frac{\eta_d^{2,(l)}(q,k')}{\hat{\eta}_d^{2,(l)}(q,k')} \right] \right|. \tag{4.9}$$

$LogErr_{\text{mean}}$ measures the distance between the true and the estimated noise PMSD as a mean over all F-MF bins and time frames. The measure includes both overestimation and underestimation and it cannot differentiate between systematic errors (bias) and random fluctuations [64]. As a result, it is capable of showing the average accuracy of a noise estimation algorithm, but it does not necessarily provide information about the resulting speech quality and intelligibility. As an example, overestimating the noise reduces intelligibility more severely than underestimating it, but this is not expressed by $LogErr_{\text{mean}}$.

The results of $LogErr_{\text{mean}}$ are shown in Table 4.1 for a babble noise scenario and in Table 4.2 for a modulated pink noise scenario, at different SNR levels. 72 sentences from the *TIMIT* train corpus [43] were used for this procedure. It can be seen that the VAD estimator has the smallest average error among the tested noise estimators at low SNR values. At higher SNR levels and especially in the non-stationary babble noise scenario, the MMSE-SPP and MS estimators appear to have a lower average deviation. It will be shown in Chapter 5 that the MMSE-SPP and MS noise estimators provide better speech quality and intelligibility than the VAD estimator. This can be explained with the fact that VAD tends to overestimate the noise PMSD (see red line in Figure 4.2(a)), while this is not the case for MMSE-SPP and MS.

The tracking curves of the different noise estimators are shown in Figure 4.2 in a babble noise scenario, at an SNR of 15 dB for the F-MF band $q = 0,\ k = 1$. Please note that in order to show a pitch synchronous F-MF band, the extended noise PMSDs $\eta_d^{2,(l)}(q,k')$ were interpolated back to their original sizes K, i.e. the black curve and the red curve correspond to $\sigma_d^{2,(l)}(0,1)$ and $\hat{\sigma}_d^{2,(l)}(0,1)$, respectively.

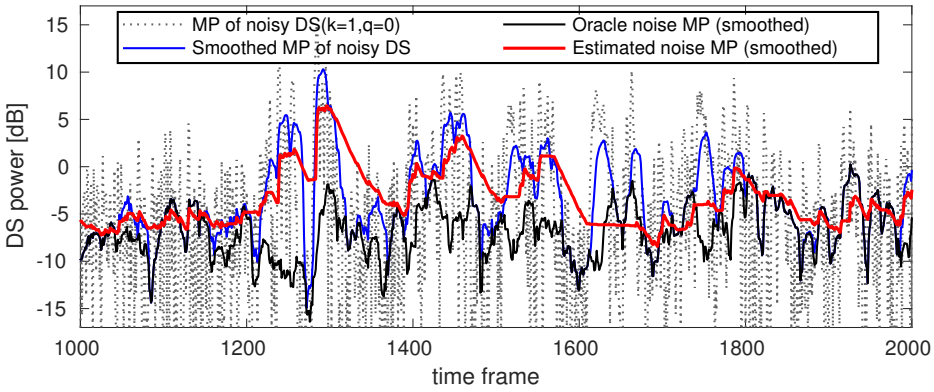
SNR [dB]	-5	0	5	10	15
VAD NE	4.09	5.09	6.29	7.54	8.58
MMSE-SPP NE	5.76	5.44	5.04	4.65	4.34
MS NE	5.38	5.03	4.70	4.52	4.68

TABLE 4.1. Logarithmic Mean Errors [dB] in *babble noise* environment of the different noise estimators.

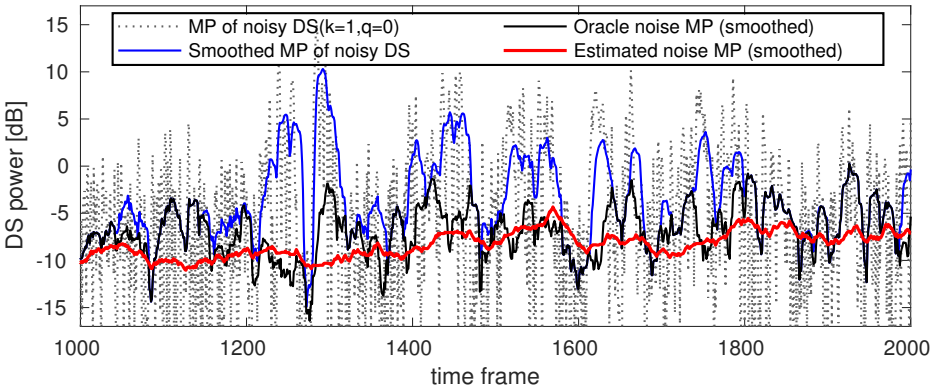
SNR [dB]	-5	0	5	10	15
VAD NE	2.93	3.12	3.49	4.24	5.36
MMSE-SPP NE	6.01	5.85	5.65	5.37	5.07
MS NE	5.17	5.03	4.84	4.60	4.34

TABLE 4.2. Logarithmic Mean Errors [dB] in *modulated pink noise* environment of the different noise estimators.

(a) VAD NE



(b) MMSE-SPP NE



(c) MS NE

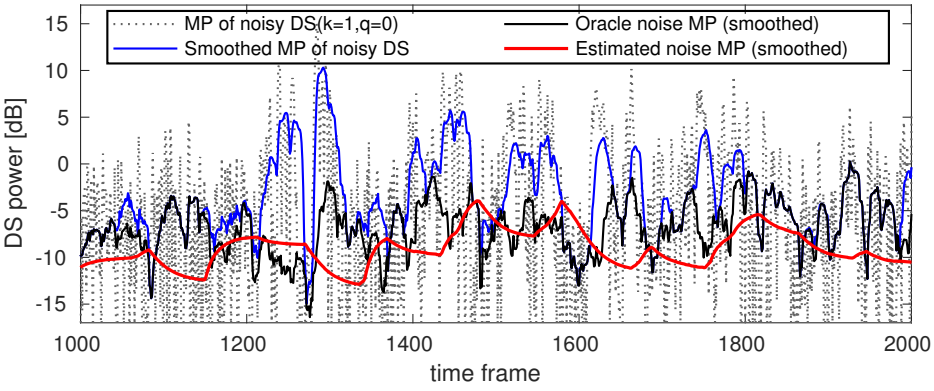


FIGURE 4.2. Noise tracking curves of the different noise estimators for $q = 0$, $k = 1$ (fundamental frequency band) in *babble noise* environment, SNR = 15 dB.

CHAPTER 5

EXPERIMENTS AND RESULTS

In this chapter, the methods for speech enhancement that were presented in Chapter 3 shall now be tested in terms of speech quality (PESQ), intelligibility (STOI) and amount of noise reduction, measured by the segmental SNR (SegSNR). We will evaluate the performance of the speech estimators in combination with the noise estimators presented in Chapter 4. Firstly, the DS-based speech estimators will be evaluated in various combinations with the noise estimators. A comparison of the speech processing frameworks that were presented in Section 2.3 will be provided using the Wiener filter. Furthermore, the impact of sign enhancement in DS will be examined and spectrograms will be presented. Finally, we will discuss the potential and limits of speech enhancement methods in the modulation domain.

5.1 Evaluation Methods

Speech and Noise Databases and Setup

In the following experiments, the *TIMIT* corpus of read speech [43] is used for the evaluation procedure. It has been designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of ASR systems. The corpus contains broadband recordings of 630 male and female speakers of eight major dialects of American English. Each speaker reads ten phonetically balanced sentences. That is, the database includes 6300 sentences in total; each of them having a resolution of 16 bit and a sampling rate of 16 kHz. *TIMIT* contains train and test subsets. For instance, the training subset was used to determine the coefficients of the speech prior PDF in Section 3.5 while the test subset will be used for evaluation in order to avoid unwanted correlations. The noise files are taken from the *NOISEX-92* database [44]. In particular, babble noise and factory noise are

used for evaluation. Additionally, white noise is used as an example for a stationary noise environment as well as modulated pink noise.

The experimental setup has been chosen as follows. Each evaluation procedure is based on 100 clean *TIMIT* speech sentences from five dialect regions that have been corrupted with babble, factory, modulated pink and white noise, at SNR levels of -5, 0, 5, 10 and 15 dB, respectively. In total, ergo 2000 different noisy speech signals have been taken into account for every evaluation process. Furthermore, every sentence is led by two seconds of noise only to give the noise estimator the chance to adjust.

PESQ

Perceptual Evaluation of Speech Quality (PESQ) will serve as measure for the percepted speech quality. It is a worldwide used ITU (International Telecommunication Union) standard for speech quality assessment, as experienced by a user of a telephony system. With PESQ, the original and degraded speech signals are mapped onto an internal representation using a perceptual model. With a cognitive model, the perceived speech quality of the degraded signal is predicted. This perceived listening quality is expressed in terms of *Mean Opinion Score*, an average quality score over a large set of subjects. Internal representations that are used by the PESQ cognitive model are calculated by using the psychophysical measures pitch and loudness [41]. The algorithm outputs a single number between 1.0 and 4.5, where 1.0 is denoting very poor quality and 4.5 excellent quality [65].

STOI

The *short-time objective intelligibility measure* (STOI) is an evaluation method that shows high correlation with the intelligibility of noisy and time-frequency weighted speech. It is based on a correlation coefficient between the temporal envelopes of the clean and degraded speech in short-time (384 ms), overlapping segments. The output of STOI is a scalar value between 0 and 1 that has a monotonic relation with the average intelligibility of a speech signal [42]. A STOI value of 1 indicates perfect intelligibility.

Segmental SNR

The *Segmental SNR* (SegSNR) algorithm will be used as a measure for the degree of noise reduction achieved by a speech enhancement algorithm. It computes the Signal-to-Noise-Ratio as an average of the SNR values of short segments of 15 to 20 ms. For this purpose, the enhanced signal $\hat{s}[n]$ is compared with the clean reference signal $s[n]$. The result is the

geometric mean value of all segments [1]. The algorithm is defined as follows:

$$\text{SegSNR} = \frac{10}{M} \sum_{m=0}^{M-1} \log_{10} \frac{\sum_{n=Nm}^{Nm+N-1} s^2[n]}{\sum_{n=Nm}^{Nm+N-1} (s[n] - \hat{s}[n])^2}. \tag{5.1}$$

5.2 Evaluation of Speech Enhancement Methods in DS

In this section, we will evaluate the following DS-based *speech estimators* (SE) presented in Chapter 3:

- ADSW (see Section 3.2.2),
- DS-Wiener (see Section 3.3),
- DS-MMSE (see Section 3.4),
- DS-MAP (see Section 3.5),

in combination with the *noise estimators* (NE):

- VAD based noise estimator (see Section 4.2),
- SPP-MMSE based noise estimator (see Section 4.3),
- MS noise estimator (see Section 4.4).

To have a reasonable comparison, the evaluation procedure has been accomplished for each combination of noise estimator and gain function (except of ADSW that does not require any noise PMSD estimation). We will now have a look on the parameter setup used for the evaluation and then compare the results and delta scores (improvements / degradations compared to the scores of the noisy speech) of PESQ, STOI and SegSNR for a stationary noise environment (white noise) and a highly non-stationary one (babble noise). The results obtained for the single sentences are averaged over all speakers and are presented as mean scores.

5.2.1 Parameter Setup

In this section, the parameters and implementation details for the evaluation of DS methods will be presented. Since all algorithms under test are implemented in Double Spectrum domain, the same input parameters are used for all of them, for the sake of comparability.

At DS analysis stage, the number of modulation bands was set to $Q = 4$. This was shown in [7] to be a reasonable trade-off between male and female speech in order to have a sufficient modulation frequency resolution.

Every method (except ADSW) requires an estimate of the noise PMSD and the a priori SNR. As described in the previous chapters, these values undergo recursive smoothing to avoid musical noise artifacts. The forgetting factor α for the decision-directed approach of a priori SNR estimation is empirically chosen as $\alpha = 0.90$. The forgetting factor $\lambda(\xi)$ for noise PMSD estimation is set adaptively depending on the instantaneous a priori SNR estimate, following the method presented in Section 4.1.2.

Another aspect that we take a great interest in is the minimum gain $G_{\min}$, i.e. the lower bound of attenuation of a DS coefficient. We know that speech energy in DS is mainly concentrated in the low modulation bands, while rapidly modulating signal content that is likely to contribute to noise is present in high modulation bands. As a result, it makes sense to allow higher attenuation in high modulation bands than in low ones. This can lead to better separation of speech and noise. In order to further improve the performance of DS based speech enhancement algorithms, $G_{\min}(q = 0,1)$ is chosen adaptively in a range between -8 and -20 dB using a logistic function. The value depends on the current estimate of the a priori SNR ξ. This method is similar to the adaptive forgetting factor λ, as presented in Section 4.1.2. Consequently, we have set the values for $G_{\min}(q,\xi)$ as follows:

$$G_{\min}(q,\xi)... \begin{cases} \in [-20,-8] \text{ dB}, & q \in [0,1] \\ = -25 \text{ dB}, & q \in [2,3] \end{cases} . \tag{5.2}$$

Informal listening revealed that especially the value of $G_{\min}(q = 0)$ is crucial for the performance of a speech enhancement algorithm as it contains the periodic signal content. When it is set too high, a buzzy noise is audible. We suppose that in that case some amplification of periodic signal content is the source of that effect. On the other hand, speech gets distorted and intelligibility decreases when this value is set too low. That is, it is important to find a good trade-off for the values of $G_{\min}$. The adaptive choice of $G_{\min}$ is one possible solution for this sensitive parameter.

The following parameters are required by the noise estimation algorithms. The threshold for VAD is set empirically to $\theta_{\text{VAD}} = 0.3$. The SPP-MMSE noise estimator requires a typical overall value of the a-priori SNR when speech is active. This value is set to $\xi_{opt} = 15$ dB. The minimum statistics noise estimator uses a sliding window. The length of this window highly affects the performance of the algorithm. The number of windows being averaged to obtain the noise periodogram is set to 12, while the number of windows which the minimum is tracked in is set to 64.

5.2.2 Results

At this point, the results of the different methods shall be presented, as outlined above. All the methods rely on a so-called *blind scenario*, i.e., only the noisy signal is known by the algorithm and there is no prior information of the pitch, noise type or noise power.

Speech Quality

Figure 5.1 shows the results for the evaluation of speech quality in terms of the PESQ score for babble noise environment. Figure 5.2 depicts the PESQ scores for white noise. The dashed black line indicates the scores of the noisy observation and the colored solid lines picture the results of enhanced speech of the respective methods. The corresponding mean delta scores are presented in Tables 5.1 and 5.2, respectively.

The performance of the estimators differs depending on the noise environment and the SNR. In babble noise environment, the ADSW estimator (the scores are the same for each NE since this method is independent of noise estimation) performs surprisingly well and outperforms the other methods in low SNR environments of -5 and 0 dB. One possible explanation is that ADSW solely relies on the structure of the DS where periodicity is enhanced by putting emphasis upon the low modulation bands. As a result, ADSW is still able to enhance speech where statistical noise estimators already are incapable of distinguishing between speech and noise at high noise levels. However, although the PESQ values are promising, intelligibility is severely degraded by the ADSW method. As the SNR increases, the DS-MAP estimator in combination with the MS NE shows to have a better performance both in babble noise and in white noise environment.

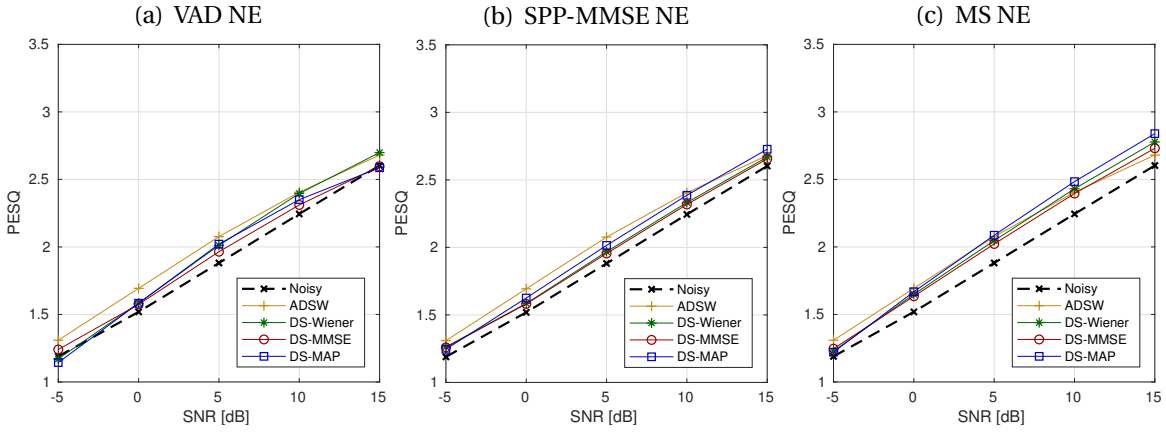


FIGURE 5.1. Plots of PESQ scores vs. SNR in *babble noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	**0.12**	**0.17**	0.20	0.16	0.08	**0.12**	**0.17**	0.20	0.16	0.08	**0.12**	**0.17**	0.20	0.16	0.08
DS-Wiener	-0.02	0.06	0.13	0.15	0.10	0.07	0.07	0.09	0.09	0.07	0.04	0.13	0.17	0.19	0.27
DS-MMSE	0.05	0.06	0.08	0.07	-0.01	0.07	0.06	0.07	0.07	0.05	0.06	0.12	0.14	0.15	0.13
DS-MAP	-0.05	0.06	0.14	0.11	-0.02	0.06	0.11	0.13	0.14	0.12	0.03	0.15	**0.21**	**0.24**	**0.24**

TABLE 5.1. Mean ΔPESQ scores in *babble noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

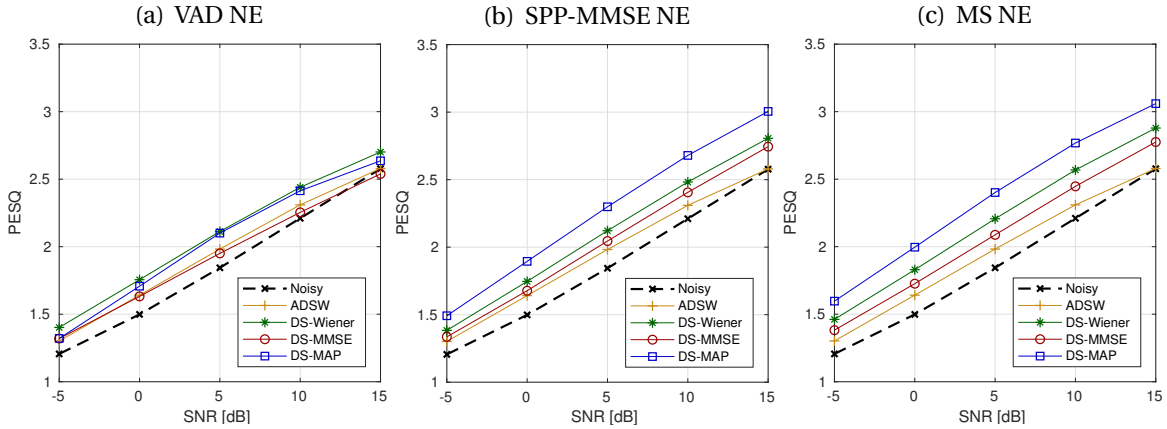


FIGURE 5.2. Plots of PESQ scores vs. SNR in *white noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	0.10	0.14	0.14	0.10	0.00	0.10	0.14	0.14	0.10	0.00	0.10	0.14	0.14	0.10	0.00
DS-Wiener	0.20	0.26	0.27	0.23	0.12	0.18	0.25	0.28	0.27	0.23	0.16	0.33	0.36	0.36	0.30
DS-MMSE	0.11	0.13	0.11	0.04	-0.04	0.13	0.18	0.20	0.20	0.17	0.18	0.23	0.24	0.24	0.20
DS-MAP	0.11	0.21	0.26	0.20	0.06	0.29	0.40	0.46	0.47	0.43	**0.39**	**0.50**	**0.56**	**0.56**	**0.48**

TABLE 5.2. Mean ΔPESQ scores in *white noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

Speech Intelligibility

Figure 5.3 shows the evaluated results in babble noise environment for speech intelligibility in terms of the STOI score. The results for white noise can be seen in Figure 5.4. The delta

scores are shown in Tables 5.3 and 5.4. Like other state-of-the-art speech enhancement algorithms, none of the algorithms presented is capable to improve intelligibility significantly. When performing noise reduction, the speech target will always be affected by some extent as well. The resulting distortion of speech decreases intelligibility. On the other hand, intelligibility at least can be preserved by the statistical methods when a more sophisticated noise estimator like MMSE-SPP or MS is used.

Unlike in speech quality, where ADSW showed promising results, intelligibility is strongly degraded by this method. The results indicate that the other methods are capable of preserving intelligibility in babble noise environment as well as in white noise. Furthermore, it is visible that in terms of STOI the SPP-MMSE and the MS noise estimator outperform the VAD estimator. Due to their higher complexity, they are able to provide a more accurate estimate of the noise PMSD than the simpler VAD based approach.

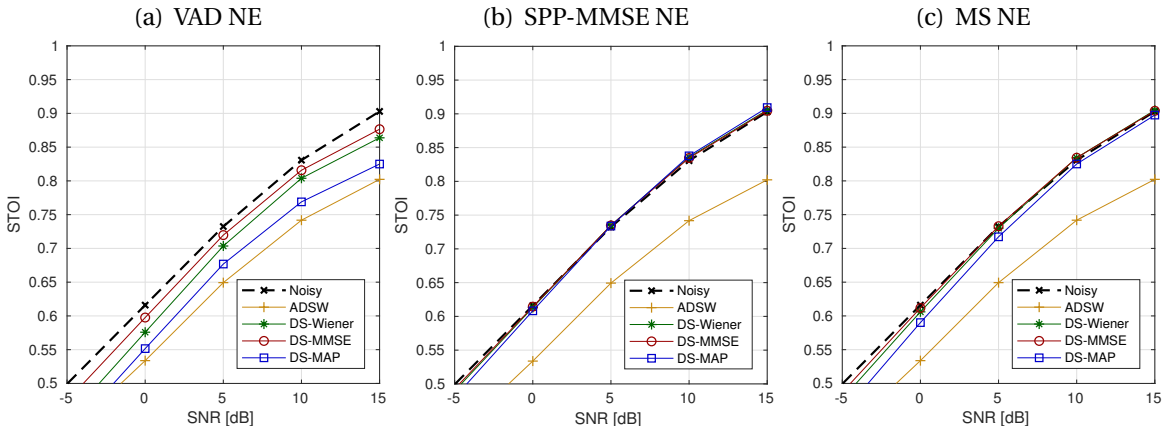


FIGURE 5.3. Plots of STOI scores vs. SNR in *babble noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	-0.07	-0.08	-0.08	-0.09	-0.10	-0.07	-0.08	-0.08	-0.09	-0.10	-0.07	-0.08	-0.08	-0.09	-0.10
DS-Wiener	-0.05	-0.04	-0.03	-0.03	-0.04	**-0.01**	**0.00**	**0.00**	**0.01**	0.00	-0.02	-0.01	**0.00**	0.00	0.00
DS-MMSE	-0.02	-0.02	-0.01	-0.01	-0.03	**-0.01**	**0.00**	**0.00**	0.00	0.00	**-0.01**	**0.00**	**0.00**	0.00	0.00
DS-MAP	-0.06	-0.06	-0.06	-0.06	-0.08	-0.02	-0.01	**0.00**	**0.01**	**0.01**	-0.04	-0.03	-0.02	-0.01	-0.01

TABLE 5.3. Mean ΔSTOI scores in *babble noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

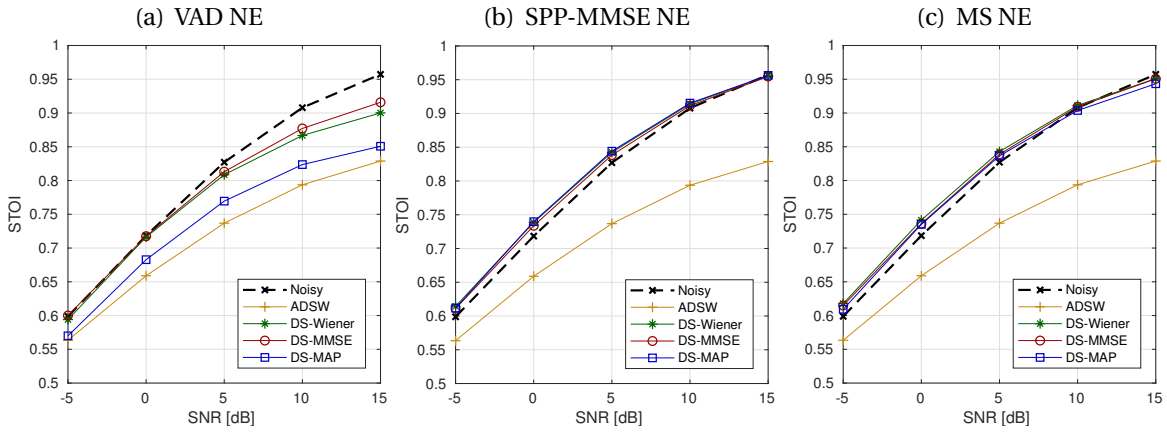


FIGURE 5.4. Plots of STOI scores vs. SNR in *white noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	-0.04	-0.06	-0.09	-0.11	-0.13	-0.04	-0.06	-0.09	-0.11	-0.13	-0.04	-0.06	-0.09	-0.11	-0.13
DS-Wiener	0.00	0.00	-0.02	-0.04	-0.06	**0.02**	**0.02**	**0.02**	**0.01**	**0.00**	**0.02**	**0.02**	**0.02**	0.00	-0.01
DS-MMSE	0.01	0.00	-0.01	-0.03	-0.04	0.01	**0.02**	0.01	0.00	**0.00**	**0.02**	**0.02**	**0.02**	0.00	-0.01
DS-MAP	-0.03	-0.04	-0.06	-0.08	-0.11	0.01	**0.02**	**0.02**	**0.01**	**0.00**	0.01	**0.02**	**0.01**	0.00	-0.01

TABLE 5.4. Mean ΔSTOI scores in *white noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

Noise Reduction

The degree of noise reduction is measured by the difference in the segmental SNR between the noisy and the enhanced signal. Figure 5.5 shows the results of SegSNR for babble noise and Figure 5.6 for white noise. The corresponding Tables 5.5 and 5.6 show the respective delta scores. In both stationary and non-stationary noise environments, the DS-MAP method together with MS noise estimation outperforms the other combinations.

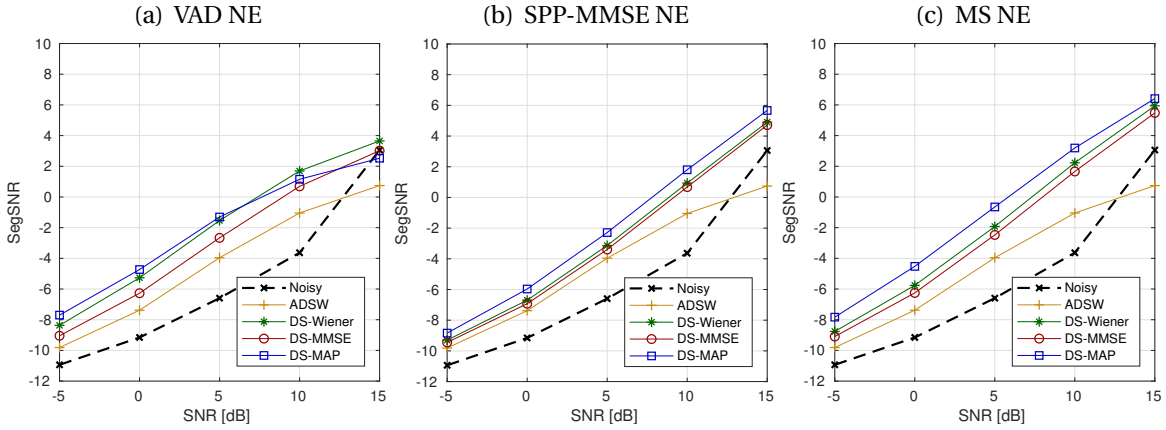


FIGURE 5.5. Plots of SegSNR scores vs. SNR in *babble noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	1.12	1.78	2.63	2.59	-2.32	1.12	1.78	2.63	2.59	-2.32	1.12	1.78	2.63	2.59	-2.32
DS-Wiener	2.57	3.89	5.06	5.33	0.59	1.64	2.46	3.46	4.57	1.85	2.17	3.40	4.67	5.86	2.88
DS-MMSE	1.89	2.88	3.93	4.33	-0.04	1.49	2.23	3.20	4.31	1.66	1.84	2.90	4.12	5.30	2.42
DS-MAP	**3.24**	4.43	5.29	4.79	-0.53	2.10	3.19	4.30	5.44	2.61	3.11	**4.63**	**5.94**	**6.83**	**3.35**

TABLE 5.5. Mean ΔSegSNR scores in *babble noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

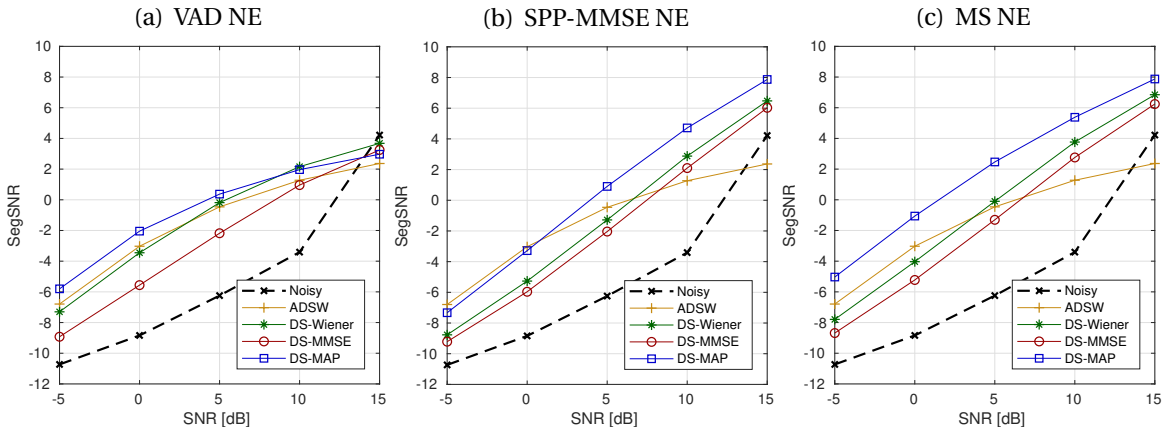


FIGURE 5.6. Plots of SegSNR scores vs. SNR in *white noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	3.94	5.81	5.78	4.68	-1.85	3.94	5.81	5.78	4.68	-1.85	3.94	5.81	5.78	4.68	-1.85
DS-Wiener	3.43	5.39	6.06	5.57	-0.54	1.96	3.57	4.96	6.29	2.26	2.93	4.81	6.15	7.17	2.63
DS-MMSE	1.80	3.28	4.06	4.37	-0.97	1.51	2.88	4.20	5.51	1.81	2.04	3.61	4.94	6.17	2.02
DS-MAP	4.93	6.80	6.61	5.38	-1.25	3.40	5.56	7.15	8.12	**3.65**	**5.70**	**7.78**	**8.71**	**8.79**	**3.65**

TABLE 5.6. Mean ΔSegSNR scores in *white noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

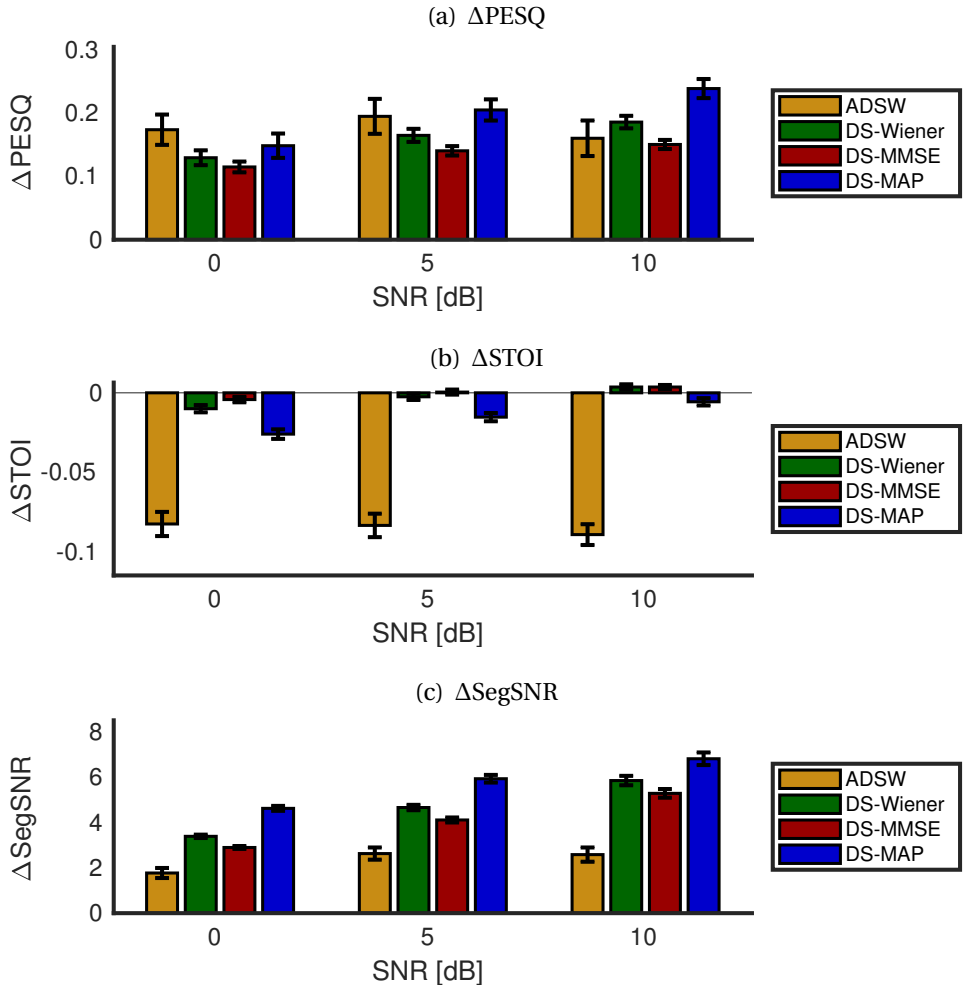


FIGURE 5.7. Barplots of delta scores in babble noise environment, using the MS noise estimator.

Conclusion

Naturally, a speech enhancement algorithm cannot only be assessed by a single value as provided by the evaluations above. It is the entirety of quality, intelligibility and pleasantness improvement of enhanced speech that determines the performance of an algorithm. The latter measure is highly subjective and can only be assessed by a listening test (that is out of the scope of this thesis). Having a look on the overall results, we can conclude that it is not the case that there is one single NE that works best in every situation. However, we can see that the simple VAD based one is inferior to the SPP-MMSE and MS based estimators by almost all means. This is due to the simple structure of the binary VAD whose capability of detecting speech activity correctly severely degrades in environments with high noise power. In terms of speech quality, the MS noise estimator provides the best results. Concerning intelligibility, no significant differences between the SPP-MMSE and the MS NE can be read out of the data. In the majority of the cases, MS tracked power spectra lead to better noise reduction than VAD and SPP-MMSE tracked ones. Based on these outcomes, we agree on using the minimum statistics noise estimator for further evaluations. Additionally, all evaluations have been performed with factory noise and modulated pink noise. The results are presented in the appendix.

As an conclusive example, the delta scores of the different speech estimators, in combination with the MS NE in babble noise environment, are depicted in Figure 5.7 for SNR levels of 0, 5 and 10 dB. The heights of the bars indicate the delta scores and the errorbars denote the 95% confidence interval.

5.2.3 Spectrograms

For visual comparison of the various DS speech enhancement methods, spectrograms of clean, noisy and enhanced speech are presented. One sentence uttered by a male and a female speaker, respectively, was chosen from the *TIMIT* database. The sentences are corrupted with white noise at an SNR of 5 dB. We use the minimum statistics approach as noise estimator (except for ADSW). The spectrograms are depicted in Figures 5.8 and 5.9, respectively. The corresponding Delta scores for PESQ, STOI and SegSNR are shown in Tables 5.7 and 5.8.

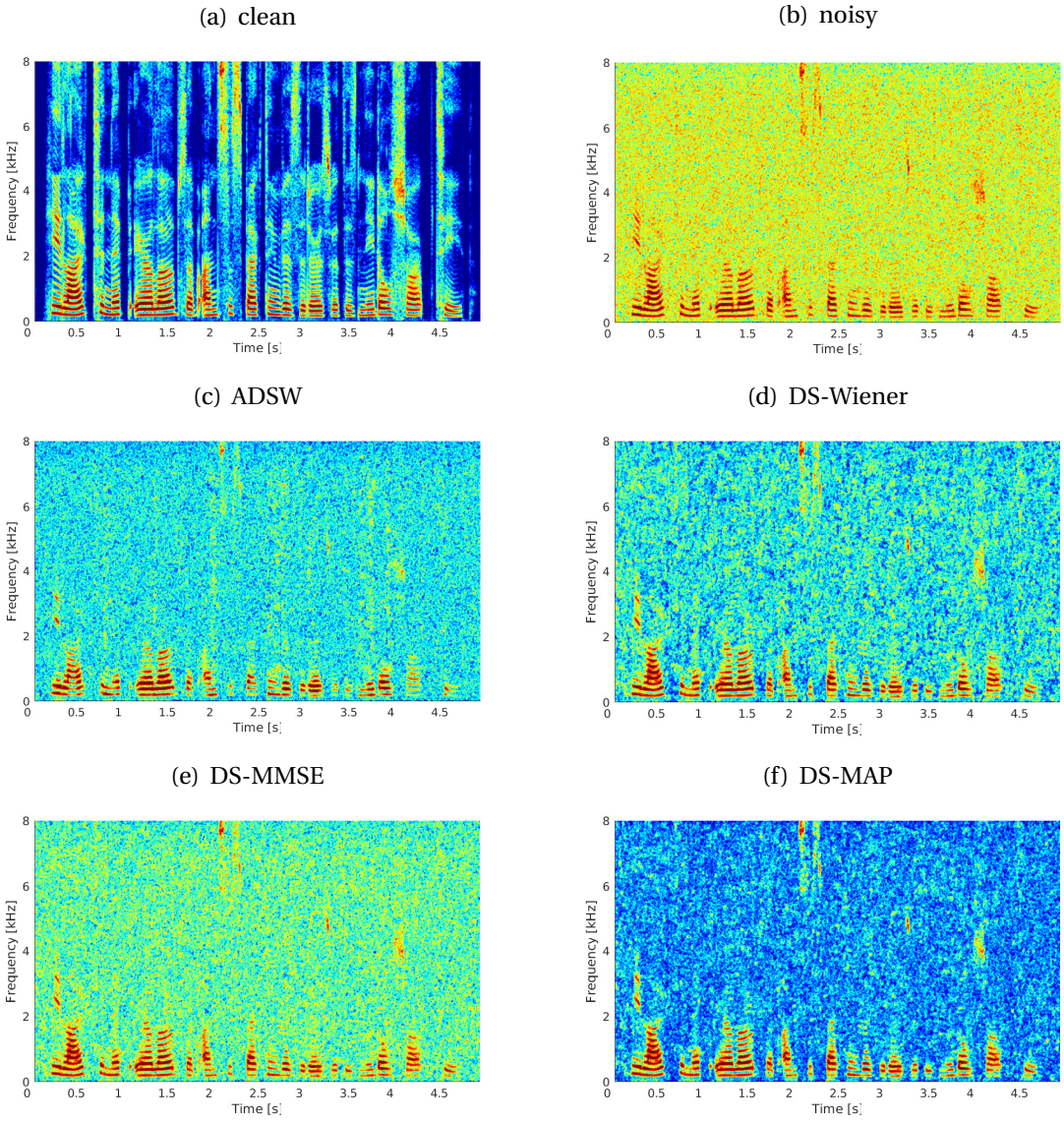


FIGURE 5.8. Spectrograms of clean, noisy (white noise @ SNR = 5 dB) and enhanced speech of the *TIMIT* sentence *'si633' "He liked to nip ear lobes of unsuspecting visitors with his needle sharp teeth"*, uttered by a female speaker.

	ΔPESQ	ΔSTOI	ΔSegSNR
ADSW	-0.10	-0.10	5.85
DS-Wiener	0.29	-0.02	7.01
DS-MMSE	0.13	-0.02	5.24
DS-MAP	0.45	-0.03	8.92

TABLE 5.7. Delta scores for enhanced speech of the *TIMIT* sentence *'si633'*.

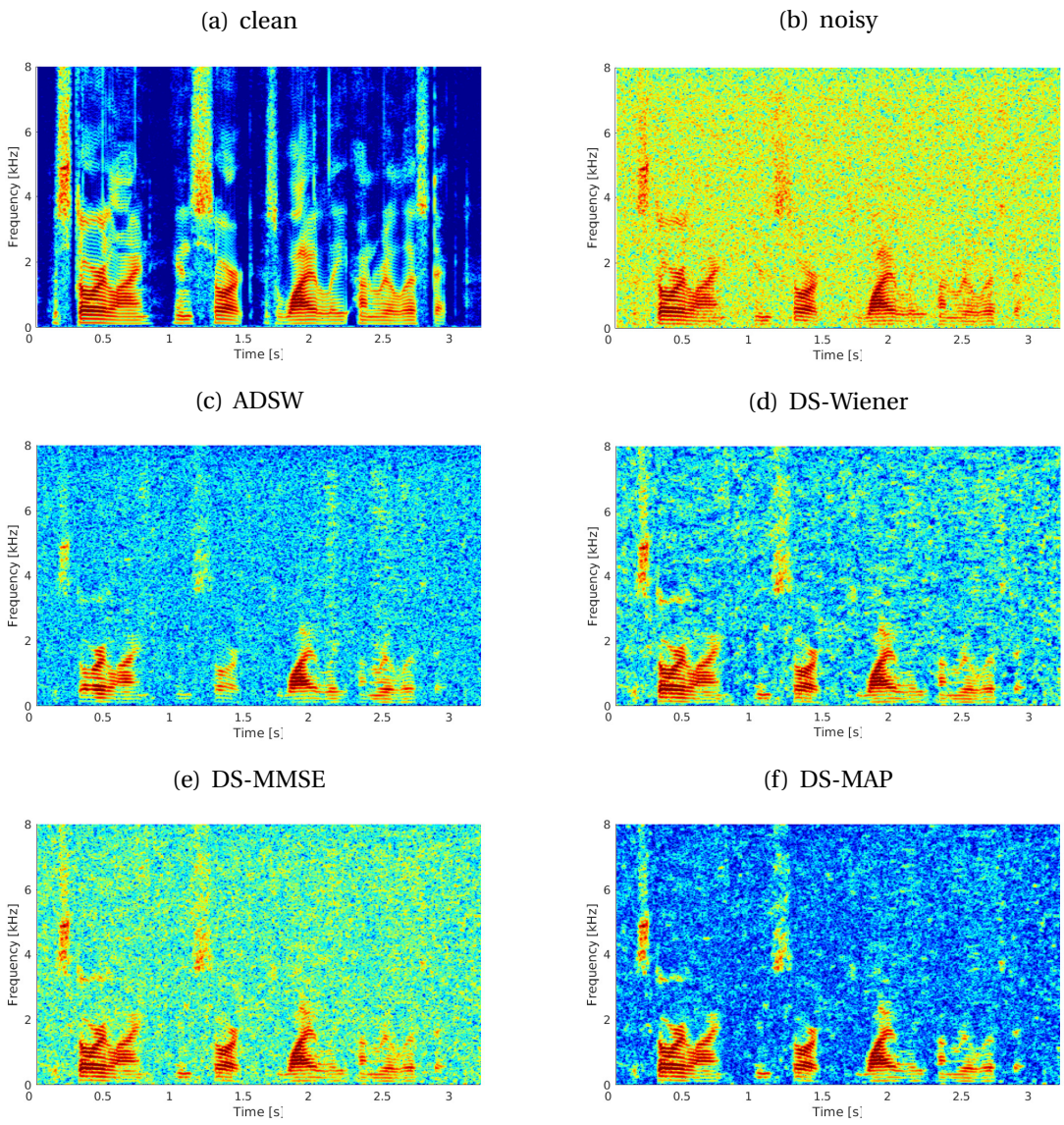


FIGURE 5.9. Spectrograms of clean, noisy (white noise @ SNR = 5 dB) and enhanced speech of the *TIMIT* sentence *'si617' "The storyline in sort is wildly unrealistic"*, uttered by a male speaker.

	ΔPESQ	ΔSTOI	ΔSegSNR
ADSW	0.06	-0.13	6.31
DS-Wiener	0.29	-0.01	7.54
DS-MMSE	0.13	-0.02	5.59
DS-MAP	0.39	-0.01	9.56

TABLE 5.8. Delta scores for enhanced speech of the *TIMIT* sentence *'si617'*.

5.2.4 Performance of f_0 Estimator

The Double Spectrum is a pitch synchronous framework and therefore dependent on a robust estimation of the pitch f_0, even under adverse acoustic environments. For that reason, the *blind* (f_0 estimated from the noisy signal) DS-Wiener and DS-MAP estimators were compared with a f_0 oracle scenario, where the fundamental frequency of clean speech is known during the DS analysis stage. The results of this comparison for factory noise at SNR levels of -5 to 10 dB are depicted as barplots in Figure 5.10. The results reveal that f_0 estimation works well even at low SNR levels and that the performance of the estimators is not degraded significantly by errors in pitch estimation. Similar results are reported for babble noise and white noise scenarios.

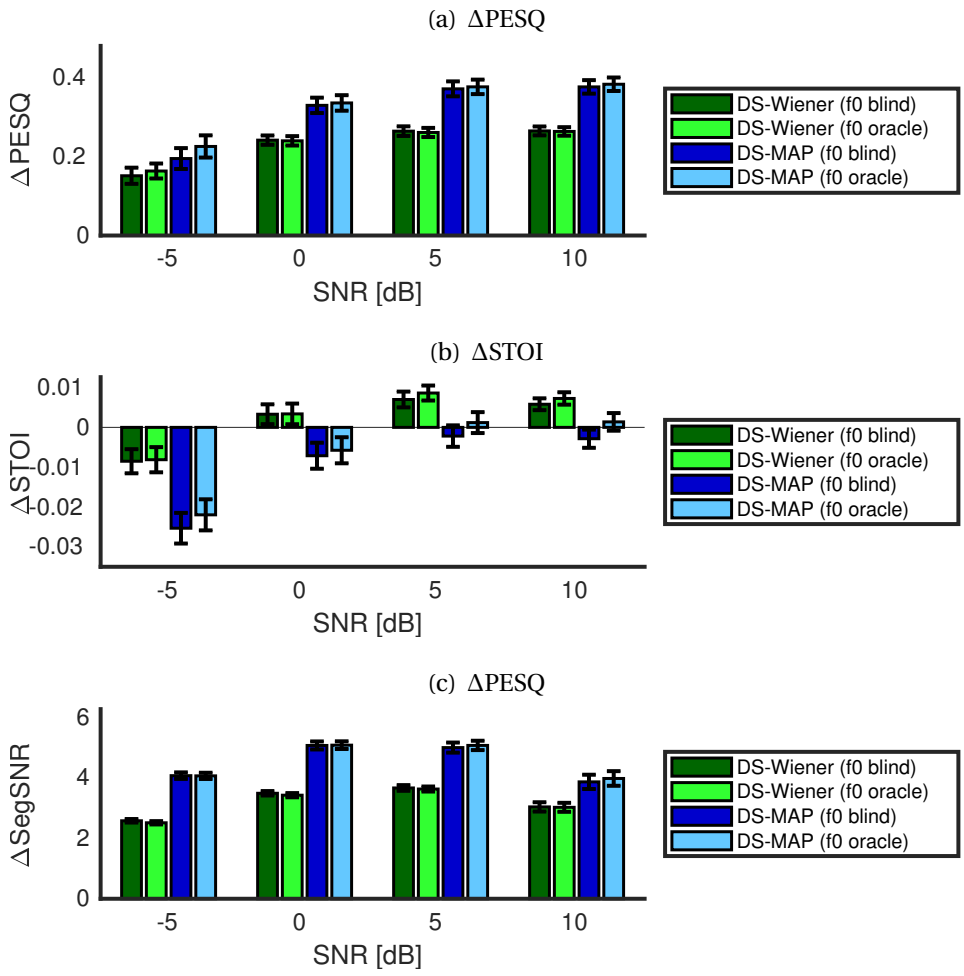


FIGURE 5.10. Barplots of delta scores in *factory noise* environment. f_0 estimation: blind vs. oracle scenario.

5.3 Evaluation of Speech Processing Frameworks

In this Section, an objective comparison between the presented speech processing frameworks in Section 2.3 will be presented. Namely, we compare the DS framework with the STSA, STSM and CHD frameworks. To have a reasonable base for this evaluation, the Wiener filter is applied as speech estimator in each framework. The minimum statistics approach is used for noise estimation.

5.3.1 Parameter Setup

The single frameworks have different principles and ideas for performing speech enhancement. Therefore, it is not surprising that it is impossible to make them directly comparable by all means. Nevertheless, at this point we will make an attempt to make this evaluation as reasonable as possible by finding a good set of parameters.

The following parameters have been set equally for all frameworks under test. The forgetting factor α for estimating the a priori SNR is chosen as $\alpha = 0.98$. The corresponding forgetting factor $\lambda(\xi)$ for noise P(M)SD estimation is set adaptively depending on the instantaneous estimate of the a priori SNR (see Section 4.1.2). We pick the value for $G_{\min}$ as -20 dB (statically for all modulation bands in DS this time, for the sake of comparability).

The window lengths for Fourier analysis are chosen as 32 ms for STSA, STSM and CHD. The corresponding window hop is set to 2 ms in STSM and CHD, and to 8 ms in the STSA framework. In the STSM framework, the window for the modulation transform has a length of 16 subsequent time frames and a hop size of 8 time frames. In CHD, additionally a modulation bandwidth of 250 Hz and a lower and an upper modulation cutoff frequency are required, which are set to 4 and 32 Hz, respectively, according to [39].

Finally, the values for the MS noise estimation sliding window are chosen depending on the time resolution of subsequent frames. As a result, 12 windows for constructing the periodogram and 64 windows for minima tracking are used in STSA, DS and CHD, while 16 and 4 windows, respectively, are taken for that purpose in STSM domain due to its lower time resolution.

The remaining parameters in DS domain are chosen equally as in Section 5.2.1.

5.3.2 Results

The results of the framework evaluation procedure in babble noise environment are shown in Figure 5.11. The respective delta scores are presented in Table 5.9 and depicted in Figure

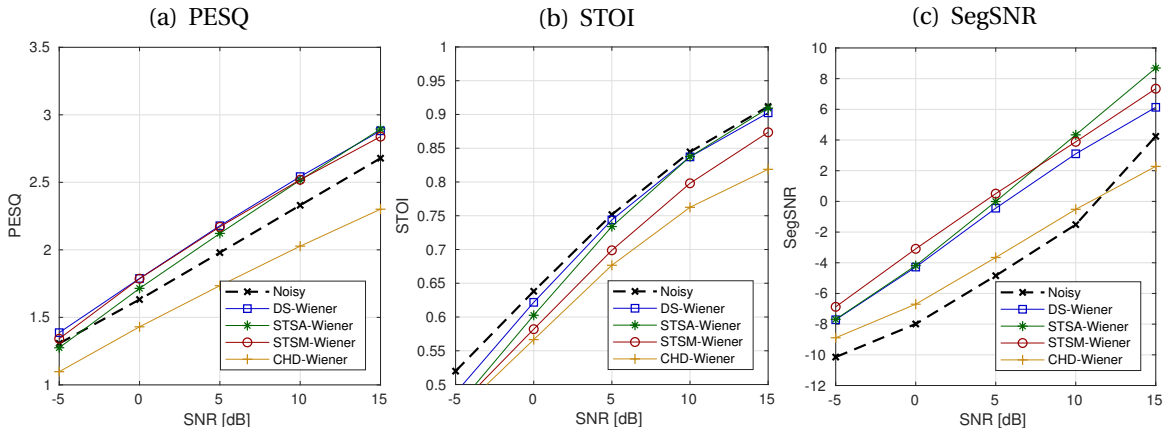


FIGURE 5.11. Plots of PESQ, STOI and SegSNR scores vs. SNR in *babble noise* environment. Evaluation of speech enhancement frameworks.

	ΔPESQ					ΔSTOI					ΔSegSNR				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
DS-Wiener	**0.08**	**0.16**	**0.20**	**0.21**	0.20	**-0.03**	**-0.02**	**-0.01**	-0.01	-0.01	2.43	3.71	4.40	4.62	1.89
STSA-Wiener	-0.03	0.08	0.14	0.19	0.22	-0.05	-0.04	-0.02	**-0.01**	**0.00**	2.46	3.81	4.83	**5.85**	**4.45**
STSM-Wiener	0.03	0.15	0.19	0.19	0.16	-0.06	-0.06	-0.05	-0.05	-0.04	**3.27**	**4.90**	**5.35**	5.40	3.11
CHD-Wiener	-0.21	-0.20	-0.25	-0.30	-0.38	-0.06	-0.07	-0.07	-0.08	-0.09	1.26	1.29	1.19	1.00	-1.96

TABLE 5.9. Mean delta scores of PESQ, STOI and SegSNR in *babble noise* environment. Evaluation of speech enhancement frameworks. Largest improvements per SNR level in **boldface**.

5.12. The results indicate that the pitch synchronous DS framework works well in enhancing the speech quality while preserving intelligibility at the same time. The best score for noise reduction at low SNR values is achieved by the STSM framework and at high SNR values by the STSA framework. On the other hand, intelligibility is degraded by a higher amount in the STSM framework. This might relate to the worse time resolution that introduces time smearing in the enhanced time signal. Since speech onsets are very important for intelligibility, this effect reduces the STOI score in STSM domain. In terms of quality, no significant difference between the STSA, STSM and DS frameworks is observable. The well-known Wiener filter in STSA domain achieves also good results, quite similar to those achieved by DS. In comparison to its counterparts, the CHD framework shows a very poor performance. This method is highly dependent on a robust estimate of the pitch. In the present implementation, f_0 is estimated by the PEFAC [66] algorithm. When the f_0 estimate is not accurate, all the harmonics of the speech signal will be distorted accordingly. We

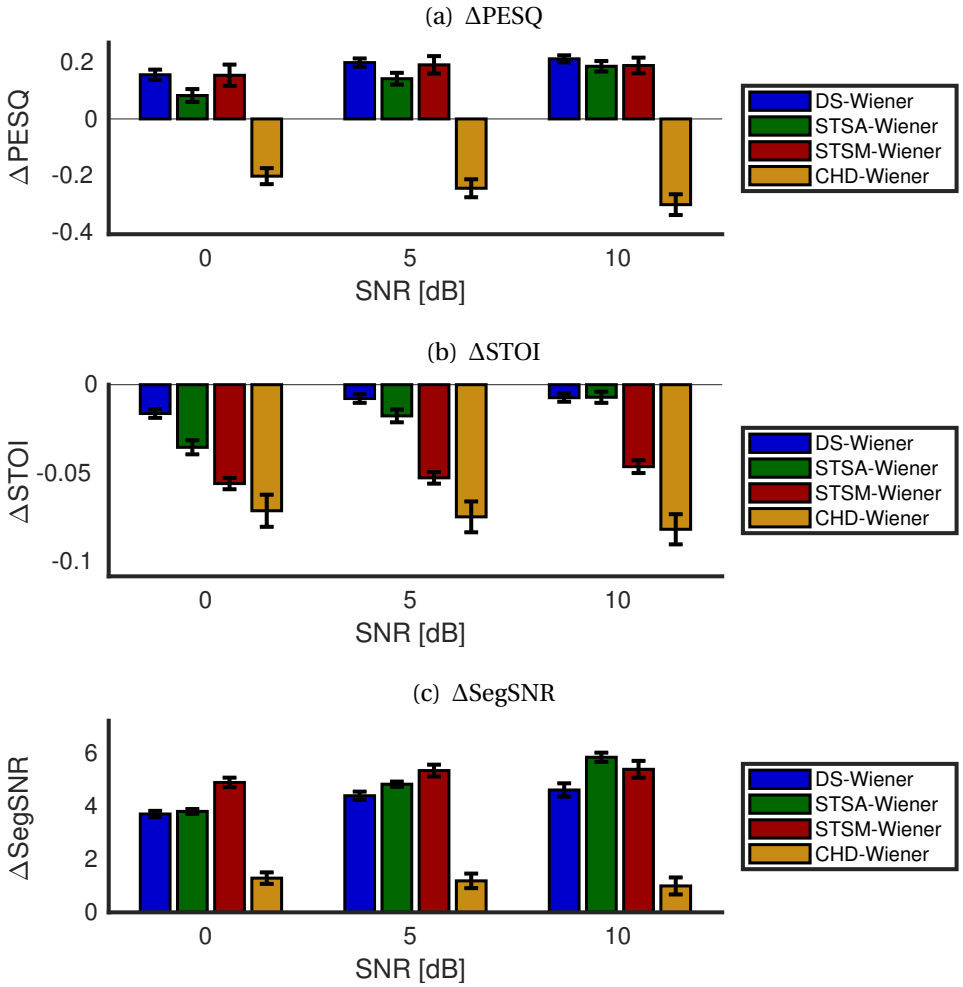


FIGURE 5.12. Barplots of delta scores in *babble noise* environment. Evaluation of speech enhancement frameworks.

suspect that this algorithm could be improved by implementing it with a more robust pitch estimator.

The STSA framework obviously performs best in white noise environment, especially in enhancing speech quality and in noise reduction. The STSM and DS frameworks achieve similar results in terms of quality, but STSM decreases intelligibility more. The results are depicted in Figure 5.13. The associated delta scores are shown in Table 5.10 and plotted in Figure 5.14.

One disadvantage of modulation based speech frameworks compared to frameworks that work in acoustic frequency domain is the higher computational complexity due to the extra modulation transform that has to be performed. The advantage of using a modulation

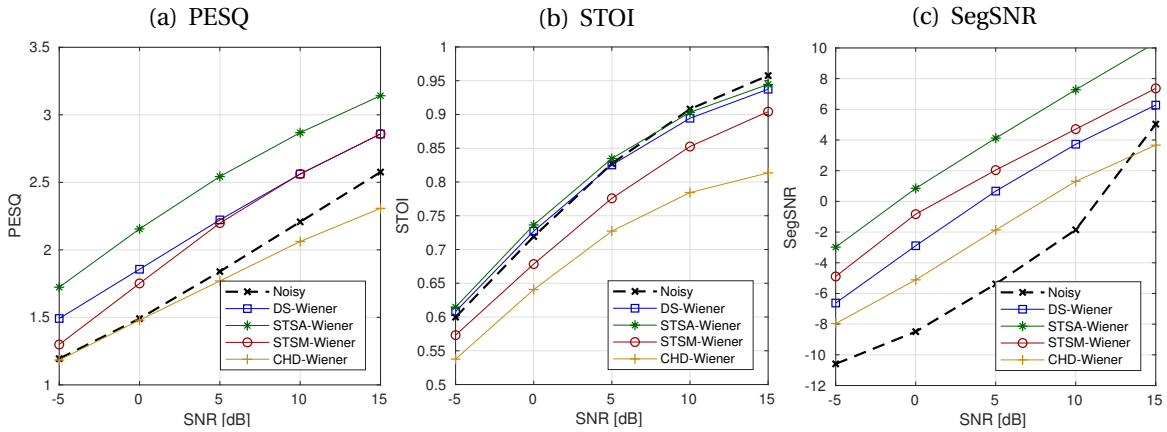


FIGURE 5.13. Plots of PESQ, STOI and SegSNR scores vs. SNR in *white noise* environment. Evaluation of speech enhancement frameworks.

	ΔPESQ					ΔSTOI					ΔSegSNR				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
DS-Wiener	0.30	0.37	0.38	0.35	0.28	**0.01**	0.01	0.00	-0.01	-0.02	3.97	5.60	6.05	5.58	1.24
STSA-Wiener	**0.53**	**0.66**	**0.70**	**0.66**	**0.57**	**0.01**	**0.02**	**0.01**	**0.00**	**-0.01**	**7.61**	**9.34**	**9.51**	**9.13**	**5.31**
STSM-Wiener	0.10	0.26	0.36	0.36	0.28	-0.03	-0.04	-0.05	-0.06	-0.05	5.70	7.66	7.42	6.56	2.33
CHD-Wiener	-0.01	-0.01	-0.07	-0.25	-0.27	-0.06	-0.08	-0.10	-0.12	-0.14	2.63	3.38	3.52	3.16	-1.37

TABLE 5.10. Mean delta scores of PESQ, STOI and SegSNR in *white noise* environment. Evaluation of speech enhancement frameworks. Largest improvements per SNR level in **boldface**.

based framework should be reasonably high to justify the higher effort in comparison to the computationally more simple STSA framework. The results obtained indicate that this is only partially the case.

5.3.3 Spectrograms

At this point we want to compare the spectrograms of clean, noisy and enhanced speech obtained in the different frameworks as described in the section above. The same set of sentences from the *TIMIT* database as in Section 5.2.3 was used for this comparison. Again, the clean sentences were corrupted with white noise at an SNR of 5 dB. The corresponding Delta scores for PESQ, STOI and SegSNR are shown in Tables 5.11 and 5.12. While it can be seen that speech enhancement in the DS, STSA and STSM frameworks provides improvements, the spectrograms visualize the problem with the CHD algorithm. The dependence on the f_0 estimate is clearly visible in the spectrograms, as well as the lack of noise suppression in the preserved frequency bands.

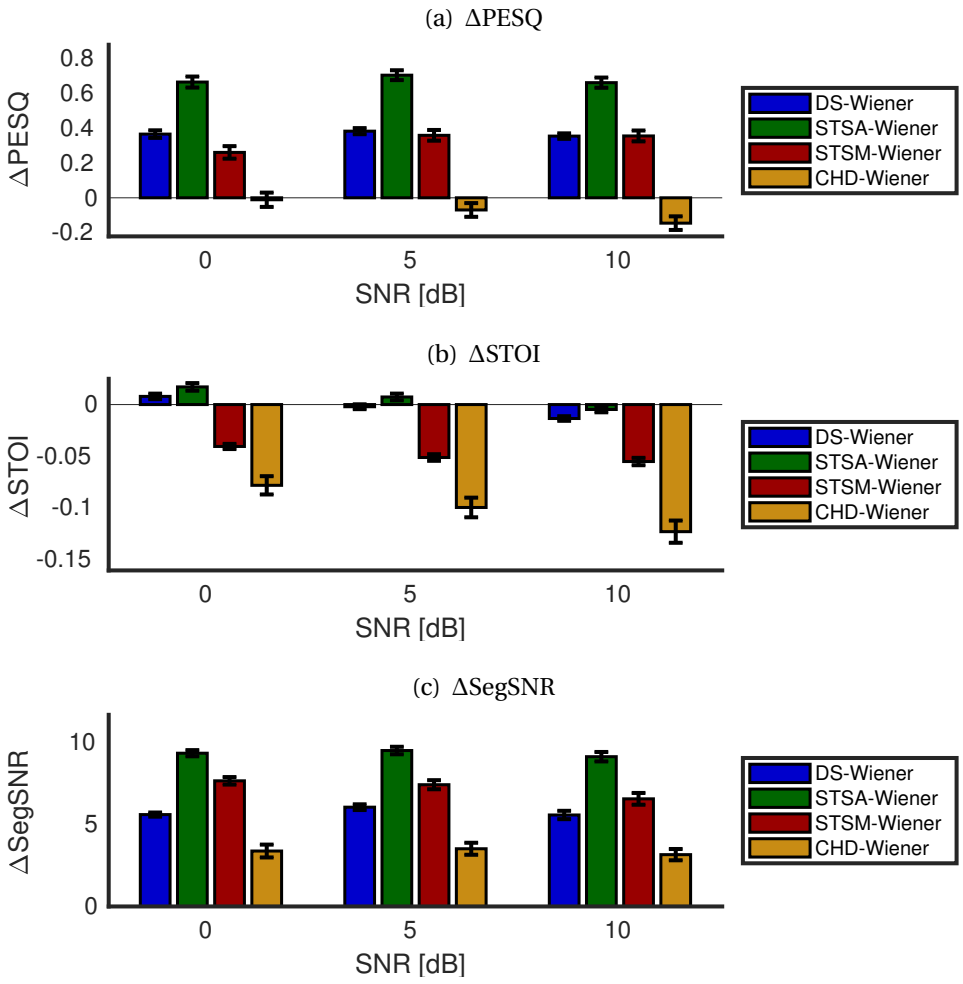


FIGURE 5.14. Barplots of delta scores in *white noise* environment. Evaluation of speech enhancement frameworks.

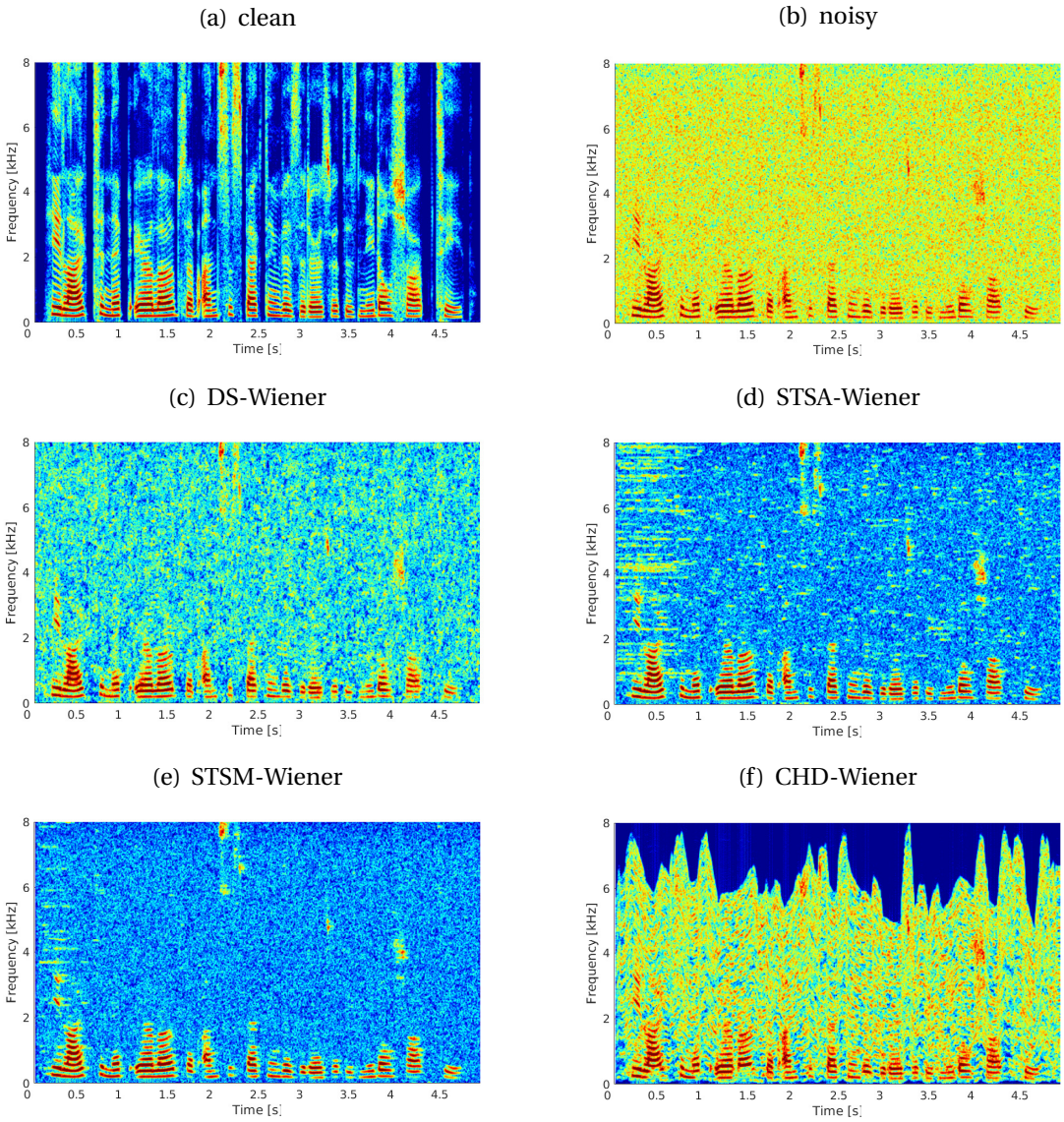


FIGURE 5.15. Spectrograms of clean, noisy (*white noise* @ SNR = 5 dB) and enhanced speech in different frameworks of the *TIMIT* sentence *'si633'*, uttered by a female speaker.

	ΔPESQ	ΔSTOI	ΔSegSNR
DS-Wiener	0.30	0.02	4.35
STSA-Wiener	0.70	0.00	9.36
STSM-Wiener	0.59	-0.03	9.14
CHD-Wiener	-0.03	-0.06	0.98

TABLE 5.11. Delta scores for enhanced speech in different frameworks, female speaker.

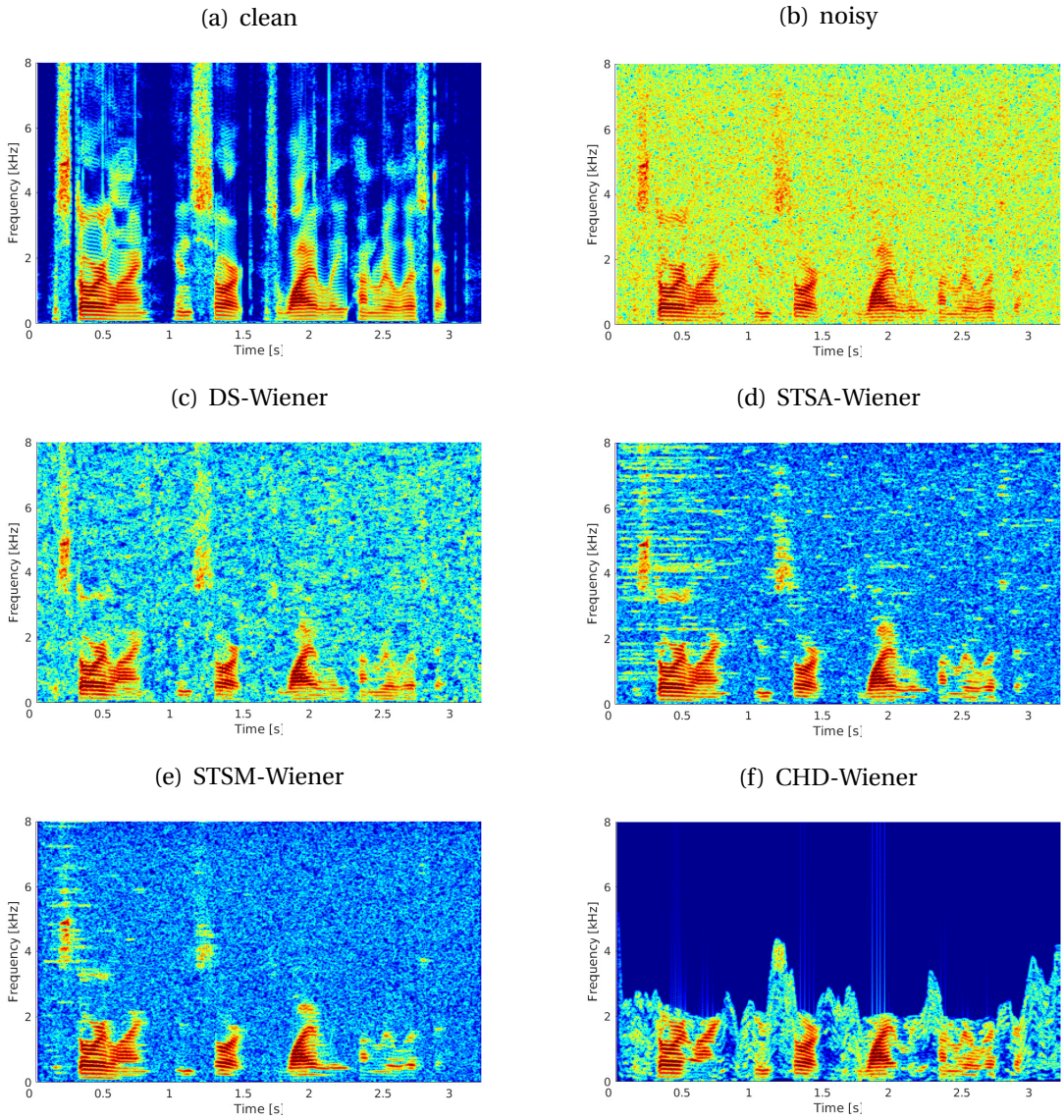


FIGURE 5.16. Spectrograms of clean, noisy (*white noise* @ SNR = 5 dB) and enhanced speech in different frameworks of the *TIMIT* sentence *'si617'*, uttered by a male speaker.

	ΔPESQ	ΔSTOI	ΔSegSNR
DS-Wiener	0.27	0.03	3.98
STSA-Wiener	0.48	-0.02	8.26
STSM-Wiener	0.33	-0.05	8.01
CHD-Wiener	0.18	-0.16	0.98

TABLE 5.12. Delta scores for enhanced speech in different frameworks, male speaker.

5.4 Impact of DS Trajectory Smoothing

Finally, we want to evaluate the effect of smoothing of the DS trajectories, as described in Section 3.7.2. We compare enhanced speech obtained by the DS-MAP estimator with noisy sign and smoothed trajectories. The interpolation filter and the recursive smoothing approach will be investigated. The results turn out to be very similar for all noise types. Therefore, only the results for factory noise will be presented at that point because it contains stationary noise as well as non-stationary elements caused by heavy machines.

5.4.1 Parameter Setup

When using the interpolation filter, the signal is separated into voiced and unvoiced frames before being filtered. Depending on the voicing state and the modulation band q, the number of windows taken into account for smoothing varies. The following values N_v for voiced frames and N_uv for unvoiced frames are chosen empirically:

$$N_\mathrm{v}(q) = \begin{cases} 9, & q=0 \\ 5, & q=1 \\ 3, & q=2 \\ 1, & q=3 \end{cases}, \quad N_\mathrm{uv}(q) = \begin{cases} 5, & q=0 \\ 3, & q=1 \\ 1, & q=2 \\ 1, & q=3 \end{cases}.$$

In order to avoid hard cuts between voiced and unvoiced areas, the overlap at transitions is set to 2 windows. The threshold MBR for VUV separation is chosen as 0.35.

In the recursive smoothing approach, the smoothing parameter $\alpha^{(l)}(q)$ is determined by a logistic function, as defined in Equation (3.40). The modulation-band dependent parameter $\beta(q)$ is chosen as

$$\beta(q) = \begin{cases} 0.7, & q=0 \\ 0.5, & q=1 \\ 0.5, & q=2 \\ 0.3, & q=3 \end{cases}.$$

5.4.2 Results

The results of the evaluation in factory noise environment for quality, intelligibility and segmental SNR are shown in Figure 5.17. Table 5.13 shows the delta scores that are also depicted graphically in Figure 5.18. It turns out that both proposed smoothing strategies do not lead to any significant improvement in the tested measures. The values of smoothed DS

trajectories are either remaining unchanged or even slightly degraded at high SNR levels. Since the proof of concept in Section 3.7.1 showed promising results when preserving the polarity of DS coefficients, a better method of taking advantage of the polarity property yet has to be derived.

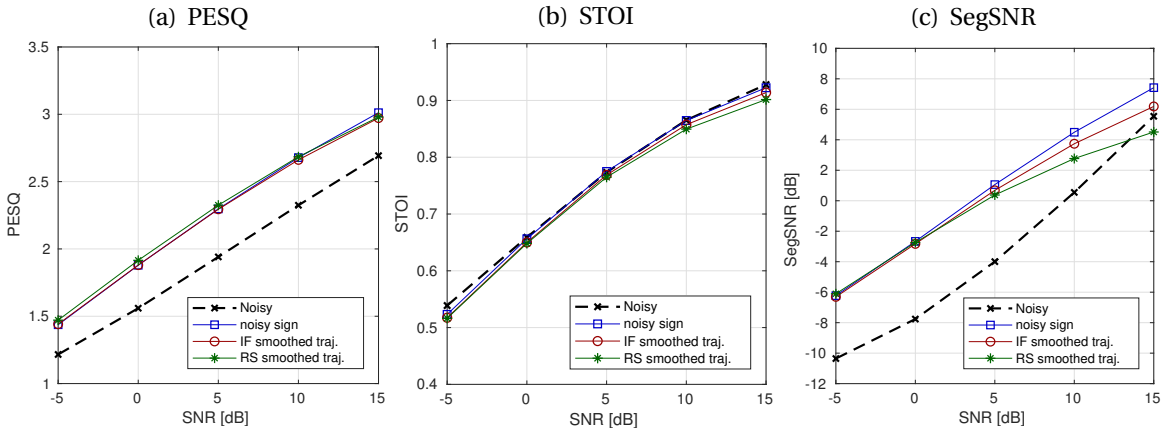


FIGURE 5.17. Plots of PESQ, STOI and SegSNR scores vs. SNR in *factory noise* environment. Evaluation of the impact of DS trajectory smoothing on DS-MAP enhanced speech.

	ΔPESQ					ΔSTOI					ΔSegSNR				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
noisy traj.	0.22	0.32	0.36	0.35	0.32	-0.02	0.00	0.00	0.00	0.00	4.13	5.10	5.06	3.95	1.88
IF smoothed	0.23	0.32	0.35	0.34	0.28	-0.02	-0.01	0.00	-0.01	-0.01	4.05	4.93	4.68	3.19	0.64
RS smoothed	0.26	0.36	0.38	0.36	0.29	0.01	0.01	-0.01	-0.02	-0.04	4.25	5.02	4.37	2.22	-1.03

TABLE 5.13. Mean delta scores of PESQ, STOI and SegSNR in *factory noise* environment. Evaluation of the impact of smoothing of the DS trajectories on DS-MAP enhanced speech.

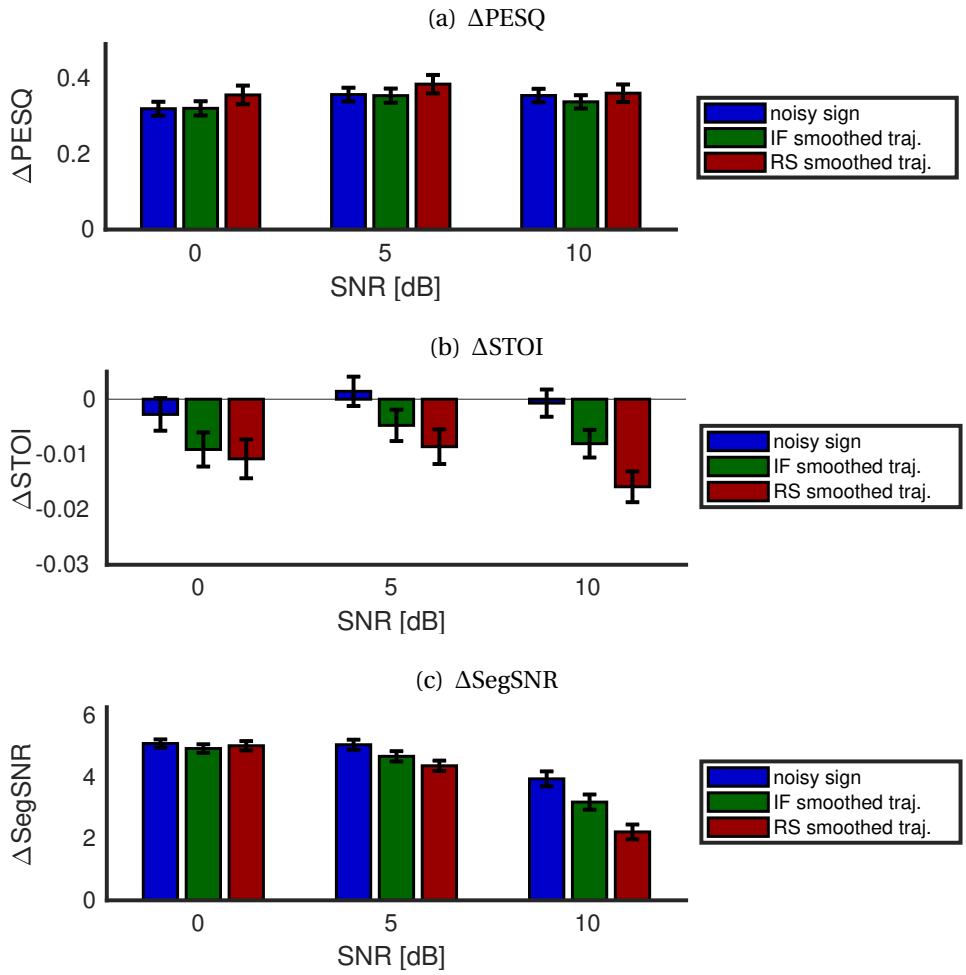


FIGURE 5.18. Barplots of delta scores in *factory noise* environment. Evaluation of the impact of smoothing of the DS trajectories on DS-MAP enhanced speech.

5.5 Potentials and Limits

Throughout this chapter, the potentials of modulation-based speech enhancement, especially in Double Spectrum domain, have been evaluated. Provided that a good combination of speech and noise estimators for DS speech enhancement is given, the proposed methods can keep pace with benchmark algorithms in other speech processing frameworks. The DS framework has the additional advantage of being a pitch-synchronous one that makes it capable of drawing advantages from the harmonicity of speech. This is for instance not possible in the STSM framework. This property can help to preserve speech intelligibility while reducing noise significantly and simultaneously increasing speech quality.

However, we believe that the potential of DS based speech enhancement has not been fully exploited yet. To name an example, the minimum statistics based noise estimation can be further improved by (i) introducing advanced bias compensation and (ii) adaptation of the size of the sliding window. Unlike in STSM domain, the TBS in DS leads to frames of variable lengths. Accordingly, when having a fixed sliding window length for minima tracking, this corresponds to a non-constant segment length in time domain. Implementing a variable window size depending on the time block lengths could resolve this issue.

There is potential of improving the proposed DS-MAP estimator. The prior distribution of DS speech coefficients was obtained by averaging speech DS coefficients of 72 sentences. The histogram turned out to be very heavy-tailed and is approximated by a Gamma distribution. Taking short-time dependencies into the model could help to preserve speech onsets and offsets better that are very important for intelligibility. What is more, the experiment in 5.4 shows that the proposed smoothing methods of DS trajectories do not lead to any improvements in the objective measures yet. Deriving a method that preserves the polarity of the coefficients better can lead to further advances of DS based algorithms.

Conclusively we can say that the modulation frameworks DS and STSM show high potential for speech enhancement, while the performance of CHD is very poor. Nevertheless, this might be due to an unideal implementation of the method, especially in terms of pitch estimation. It should be noted that in [39] promising results were obtained in this domain, in contrast to the present studies.

CHAPTER 6

CONCLUSIONS AND OUTLOOK

Throughout this thesis, we pointed out the benefits of modulation-based speech processing. Supported by physiological and psychological evidence, we showed that speech processing in modulation domain matches human speech perception better than the conventional spectral processing. Motivated by these findings, several approaches of speech processing in modulation domain were proposed and studied. Special emphasis was put on the Double Spectrum, a relatively novel pitch-synchronous speech enhancement framework in modulation domain. It provides a compact representation of speech allowing to describe how spectral components evolve over time [6].

The core of this thesis is the proposal of speech and noise estimators in DS. The Wiener filter and the MMSE-STSA approach by Ephraim and Malah [33] were reformulated in the DS framework. Furthermore, a DS-MAP speech estimator was derived based on the distributions of speech and noise DS coefficients. What is more, we presented the RASTA approach for modulation filtering in acoustic frequency domain. In terms of noise estimation, we presented a VAD based noise estimator as well as an MMSE estimator relying on the speech presence probability. What is more, the minimum statistics noise estimation approach was adopted to DS domain and appears to perform well in this domain. Furthermore, the role of polarity of the DS coefficients was investigated.

The presented methods were tested thoroughly by objective measures of speech quality, intelligibility and degree of noise reduction. The newly derived algorithms showed promising results. The DS framework was tested against STSA, STSM and CHD based speech processing frameworks. The DS domain proved to be effective in reducing noise and improving quality while preserving intelligibility at the same time. It is capable to keep up with state of the art algorithms in different frameworks in stationary and non-stationary noise environments at different SNR levels. A high number of utterances ensured validation of the results.

The investigation of the full potential of modulation-based speech processing is a current subject of research and development. Especially in the Double Spectrum domain there is still room for improvement. As an outlook, the computational efficiency of DS analysis, modification and synthesis could be optimized in order to implement a real-time DS speech enhancement environment. What is more, a model-based or data driven approach could be derived, e.g. a codebook-based approach might be implemented. In this thesis, the smoothing of DS trajectories has not lead to improvements yet. Algorithms that make better use of the polarity property could be derived in future work. Further interesting applications of DS could be its extension to a multi-channel speech enhancement framework (e.g. a beamformer as pre-processor) and performing single-channel source separation. Conclusively, implementing a subjective evaluation procedure in terms of a listening test would possibly reveal very interesting results.

APPENDIX A

Further Results of Evaluation of DS Methods

At this point, the results of the evaluation of speech enhancement methods in DS performed in Section 5.2 for factory noise and modulated pink noise will be presented.

Speech Quality

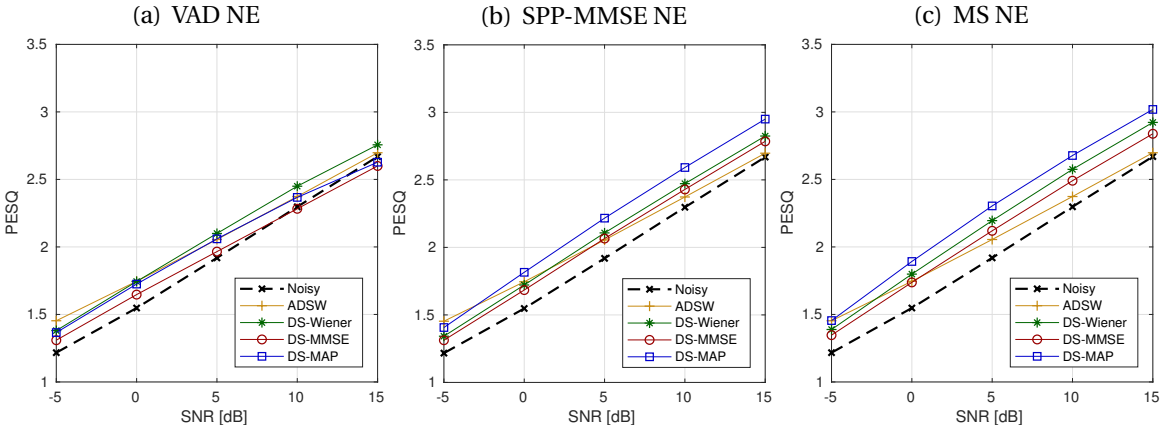


FIGURE A.1. Plots of PESQ scores vs. SNR in *factory noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	**0.24**	0.20	0.14	0.08	0.03	**0.24**	0.20	0.14	0.08	0.03	**0.24**	0.20	0.14	0.08	0.03
DS-Wiener	0.16	0.20	0.18	0.15	0.09	0.13	0.17	0.19	0.17	0.16	0.17	0.25	0.28	0.28	0.25
DS-MMSE	0.09	0.10	0.05	-0.01	-0.07	0.10	0.14	0.14	0.13	0.12	0.13	0.19	0.20	0.19	0.17
DS-MAP	0.15	0.18	0.14	0.07	-0.04	0.19	0.27	0.30	0.29	0.28	**0.24**	**0.34**	**0.38**	**0.38**	**0.35**

TABLE A.1. Mean ΔPESQ scores in *factory noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

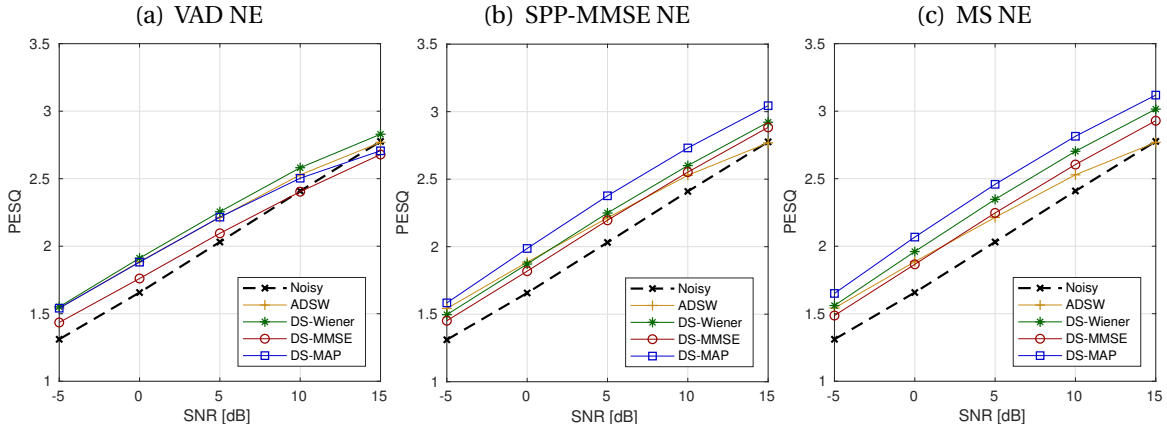


FIGURE A.2. Plots of PESQ scores vs. SNR in *modulated pink noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	0.23	0.23	0.18	0.12	-0.01	0.23	0.23	0.18	0.12	-0.01	0.23	0.23	0.18	0.12	-0.01
DS-Wiener	0.24	0.25	0.22	0.17	0.05	0.19	0.21	0.22	0.19	0.15	0.25	0.30	0.32	0.29	0.24
DS-MMSE	0.12	0.10	0.06	-0.01	-0.10	0.14	0.16	0.16	0.14	0.11	0.18	0.21	0.22	0.20	0.15
DS-MAP	0.23	0.22	0.18	0.09	-0.07	0.27	0.33	0.35	0.32	0.27	**0.34**	**0.41**	**0.43**	**0.41**	**0.34**

TABLE A.2. Mean ΔPESQ scores in *modulated pink noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

Speech Intelligibility

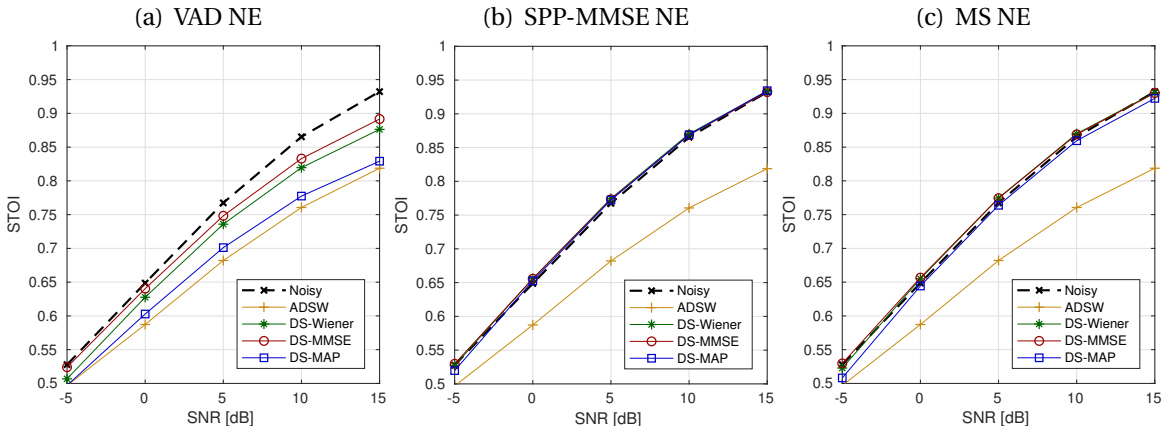


FIGURE A.3. Plots of STOI scores vs. SNR in *factory noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	-0.03	-0.06	-0.08	-0.10	-0.11	-0.03	-0.06	-0.08	-0.10	-0.11	-0.03	-0.06	-0.08	-0.10	-0.11
DS-Wiener	-0.02	-0.02	-0.03	-0.05	-0.06	**0.00**	**0.01**	**0.01**	**0.01**	**0.00**	-0.01	**0.01**	**0.01**	0.00	**0.00**
DS-MMSE	**0.00**	0.00	-0.02	-0.03	-0.04	**0.00**	**0.01**	**0.01**	0.00	**0.00**	**0.00**	**0.01**	**0.01**	0.00	**0.00**
DS-MAP	-0.03	-0.05	-0.07	-0.09	-0.10	-0.01	0.00	0.00	0.00	**0.00**	-0.02	0.00	0.00	-0.01	-0.01

TABLE A.3. Mean ΔSTOI scores in *factory noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

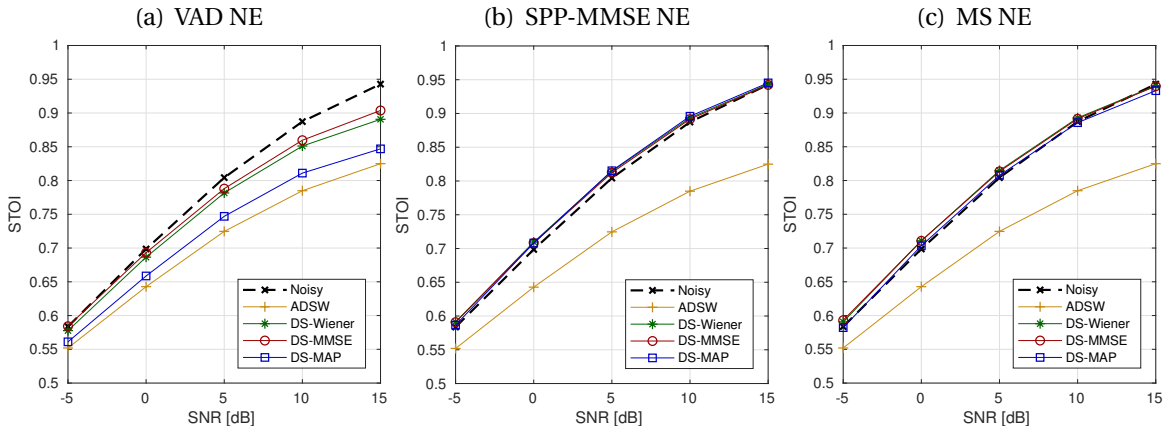


FIGURE A.4. Plots of STOI scores vs. SNR in *modulated pink noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	-0.03	-0.06	-0.08	-0.10	-0.12	-0.03	-0.06	-0.08	-0.10	-0.12	-0.03	-0.06	-0.08	-0.10	-0.12
DS-Wiener	-0.01	-0.01	-0.02	-0.04	-0.05	**0.01**	**0.01**	**0.01**	0.00	**0.00**	**0.01**	**0.01**	**0.01**	**0.01**	**0.00**
DS-MMSE	0.00	-0.01	-0.02	-0.03	-0.04	**0.01**	**0.01**	**0.01**	0.00	**0.00**	**0.01**	**0.01**	**0.01**	0.00	**0.00**
DS-MAP	-0.02	-0.04	-0.06	-0.08	-0.10	0.00	**0.01**	**0.01**	**0.01**	**0.00**	0.00	0.00	0.00	0.00	-0.01

TABLE A.4. Mean ΔSTOI scores in *modulated pink noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

Noise Reduction

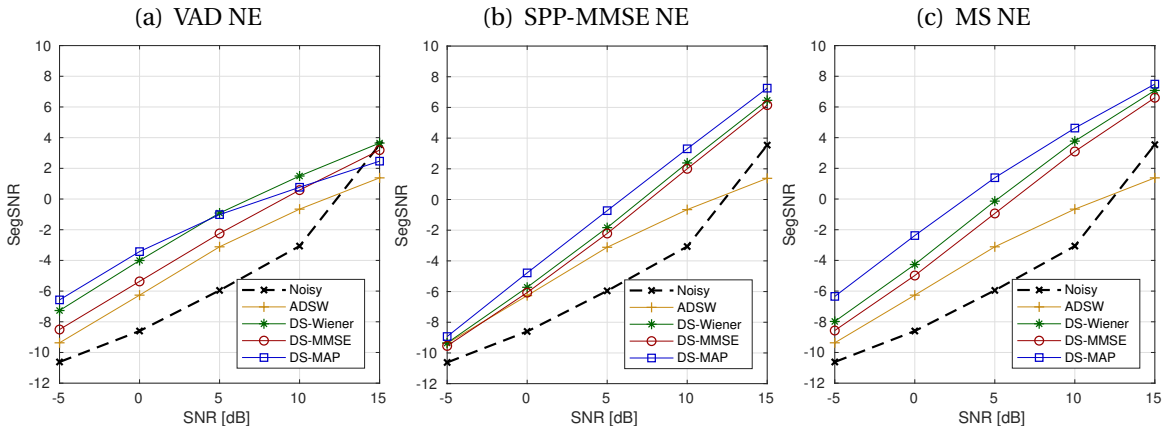


FIGURE A.5. Plots of SegSNR scores vs. SNR in *factory noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	1.25	2.34	2.84	2.41	-2.16	1.25	2.34	2.84	2.41	-2.16	1.25	2.34	2.84	2.41	-2.16
DS-Wiener	3.35	4.60	5.04	4.57	0.10	1.27	2.88	4.12	5.45	2.92	2.64	4.34	5.82	6.84	3.53
DS-MMSE	2.12	3.23	3.71	3.65	-0.36	1.08	2.53	3.74	5.07	2.61	2.05	3.62	5.01	6.16	3.06
DS-MAP	4.05	5.18	4.94	3.83	-1.08	1.69	3.82	5.23	6.37	3.71	**4.27**	**6.22**	**7.34**	**7.69**	**3.95**

TABLE A.5. Mean ΔSegSNR scores in *factory noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

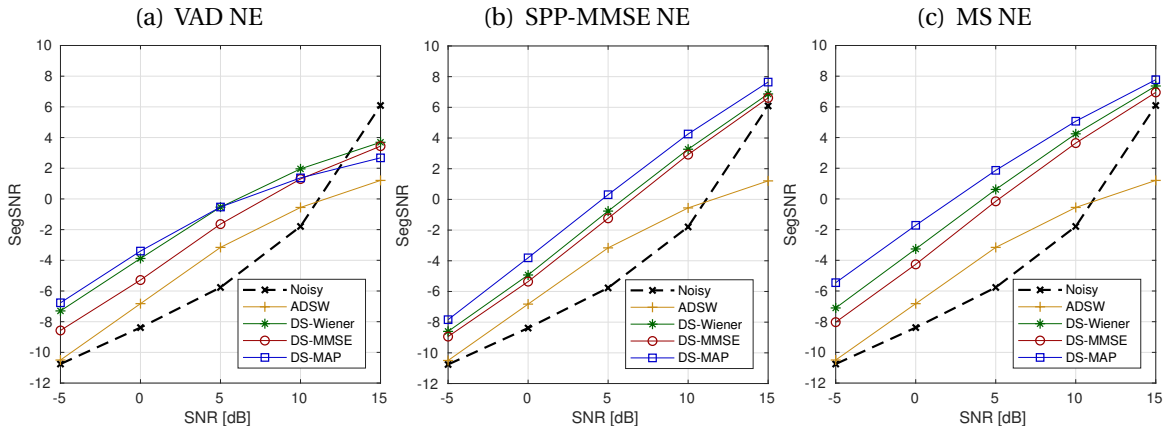


FIGURE A.6. Plots of SegSNR scores vs. SNR in *modulated pink noise* environment, using different noise estimators.

	VAD					SPP-MMSE					MS				
SNR [dB]	-5	0	5	10	15	-5	0	5	10	15	-5	0	5	10	15
ADSW	0.26	1.56	2.61	1.24	-4.88	0.26	1.56	2.61	1.24	-4.88	0.26	1.56	2.61	1.24	-4.88
DS-Wiener	3.46	4.51	5.23	3.75	-2.40	2.15	3.46	5.01	5.06	0.78	3.65	5.12	6.39	6.04	1.27
DS-MMSE	2.19	3.10	4.13	3.10	-2.65	1.82	3.03	4.53	4.71	0.52	2.73	4.12	5.61	5.43	0.85
DS-MAP	3.99	4.99	5.24	3.16	-3.41	2.92	4.58	6.07	6.05	1.56	**5.30**	**6.67**	**7.63**	**6.86**	**1.68**

TABLE A.6. Mean ΔSegSNR scores in *modulated pink noise* environment, using different noise estimators; VAD (left), SPP-MMSE (middle), MS (right). Largest improvements per SNR level in **boldface**.

List of Abbreviations

ADSW	Adaptive Double Spectrum Weighting
AMS	Analysis-Modification-Synthesis
ASR	Automatic Speech Recognition
CHD	Coherent Harmonic Demodulation
dB	Decibel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DS	Double Spectrum
F-MF	Frequency-Modulation Frequency
ICS	Ideal Channel Selection
i.i.d	independent identically distributed
iSTFT	inverse Short-Time Fourier Transform
KLT	Karhunen-Loève Transform
K-S test	Kolmogorov-Smirnov test
MAP	Maximum a posteriori
MBR	Modulation Band Ratio
MCS	Modulation Channel Selection
MF	Modulation Frequency
MFD	Modulation Frame Duration
ML	Maximum Likelihood
MLT	Modulation Lapped Transform
MMSE	Minimum Mean Square Error
MP	Modulation Periodogram
MS	Minimum Statistics
MSE	Mean Square Error
NE	Noise Estimator
OLA	Overlap-add

PDF	Probability Density Function
PE	Periodicity Enhancement
PESQ	Perceptual Evaluation of Speech Quality
PMSD	Power Modulation Spectral Density
PSD	Power Spectral Density
RASTA	Relative Spectra
RMS	Root-Mean-Square
RS	Recursive Smoothing
SE	Speech Estimator
SegSNR	Segmental Signal to Noise Ratio
SFM	Source Filter Model
SNR	Signal to Noise Ratio
SPP	Speech Presence Probability
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility Measure
STSA	Short-Time Spectral Amplitude
STSM	Short-Time Spectral Modulation
TBS	Time Block Segmentation
VAD	Voice Activity Detection
VUV	Voiced / Unvoiced
wICS	weighted Ideal Channel Selection
WSS	wide-sense stationary

List of Symbols

f_0	fundamental frequency, pitch
f_s	sampling frequency
Δf	spacing of frequency channels
ω	radian frequency
t	time-continuous time index
n	time sample index
m	time frame index in STSA domain
l	modulation frame index in STSM domain and DS
N	number of samples per frame in time domain
$\mathcal{M}$	number of windows per modulation frame in STSA domain
Z	frame shift in time domain
$\mathcal{Z}$	frame shift in STSA domain
k	frequency channel index
h	modulation band index in STSM domain
q	modulation band index in DS
$s[n]$	clean speech signal
$\hat{s}[n]$	enhanced speech signal, estimate of $s[n]$
$d[n]$	noise signal
$x[n]$	noisy signal
$m[n]$	modulator signal
$c[n]$	carrier signal
$w_A[n]$	analysis window function
$w_M(m)$	modulation window function
$G(\cdot), \mathcal{G}(\cdot)$	gain function
$x(t)$	time-continuous noisy signal
$X(j\omega)$	spectral representation of $x(t)$
$x[n]$	discrete-time noisy signal
$X(k,m)$	spectral representation of $x[n]$
$\mathcal{X}(h,k,l)$	modulation domain representation of $x[n]$

DS_s	clean double spectrum
$\widehat{DS}_s$	enhanced double spectrum, estimate of DS_s
DS_d	noise double spectrum
DS_x	noisy double spectrum
DS_{xx}	noisy double spectrum PMSD
σ_s^2	clean speech P(M)SD
σ_n^2	noise P(M)SD
η	noise PMSD of uniform size
$\mathcal{I}\{\cdot\}$	interpolation function
ξ	a priori SNR
γ	a posteriori SNR
α	forgetting factor for a priori SNR estimation
λ	forgetting factor for noise PSD estimation
$\mathbb{E}\{\cdot\}$	expectancy value operator
$\mathbb{P}\{\cdot\}$	probability operator
$\Gamma(\cdot)$	gamma function
a	shape parameter of gamma distribution
b	scale parameter of gamma distribution

BIBLIOGRAPHY

[1] P. Loizou, *Speech Enhancement Theory and Practice*.
CRC Press, 2 ed., 2013.

[2] J. Lim and A. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proceedings of the IEEE*, vol. 67, pp. 1586–1604, Dec. 1979.

[3] J. Bosco and E. Plourde, "Speech Enhancement using both Spectral and Spectral Modulation Domains," in *IEEE 30th Canadian Conference on Electrical and Computer Engineering*, 2017.

[4] J. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-25, pp. 235–238, June 1977.

[5] M. Nilsson, B. Resch, M. Y. Kim, and W. B. Kleijn, "A canonical representation of speech," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, pp. IV–849–IV–852, Apr. 2007.

[6] M. Blass, "Single-Channel Speech Enhancement Using Double Spectrum," Master's thesis, TU Graz, June 2016.

[7] P. Mowlaee, M. Blass, and W. B. Kleijn, "New results in modulation-domain single-channel speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 2125–2137, Nov. 2017.

[8] A. Koutrouvelis, R. Hendriks, and R. Heusdens, "Digital Audio and Speech Processing: Lecture 2, Speech Production and Modeling." Power Point Presentation, TU Delft, 2015.

[9] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding And Error Concealment*.
John Wiley & Sons, 2006.

[10] Standards Secretariat, Acoustical Society of America. ANSI S1.1-1994 (R2004) American National Standard Acoustical Terminology.

[11] B. Schwerin, *Modulation Domain Based Processing for Speech Enhancement.* PhD thesis, Signal Processing Laboratory, Griffith School of Engineering Science, Environment, Engineering and Technology, Griffith University, Brisbane, Australia, 2012.

[12] R. Drullman, J. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 5, pp. 2670–2680, 1994.

[13] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.

[14] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Communication*, vol. 52, no. 5, pp. 450 – 475, 2010.

[15] S. M. Schimmel, *Theory of Modulation Frequency Analysis and Modulation Filtering, with Applications to Hearing Devices.* PhD thesis, University of Washington, 2007.

[16] H. Dudley, "The carrier nature of speech," *The Bell System Technical Journal*, vol. 19, pp. 495–515, Oct. 1940.

[17] L. Atlas and S. A. Shamma, "Joint Acoustic and Modulation Frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, pp. 668–675, June 2003.

[18] E. Zwicker, "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones," *Acta Acustica*, vol. 2, no. 3, pp. 125–133, 1952.

[19] N. F. Viemeister, "Temporal modulation transfer functions for audition," *The Journal of the Acoustical Society of America*, vol. 53, no. 1, p. 312, 1973.

[20] N. F. Viemeister, "Modulation thresholds and temporal modulation transfer functions," *The Journal of the Acoustical Society of America*, vol. 60, no. S1, p. S117, 1976.

[21] S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *The Journal of the Acoustical Society of America*, vol. 85, no. 6, pp. 2575–2580, 1989.

[22] S. Sheft and W. A. Yost, "Temporal integration in amplitude modulation detection," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 796–805, 1990.

[23] J. Lee, "Amplitude modulation depth discrimination of a sinusoidal carrier: Effect of stimulus duration," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3688–3693, 1997.

[24] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 4, pp. 2490–2493, Oct. 1996.

[25] J. B. Boldt, A. T. Bertelsen, F. Gran, S. Jørgensen, and T. Dau, "Single channel speech enhancement in the modulation domain: New insights in the modulation channel selection framework," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5748–5752, Apr. 2015.

[26] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, vol. 77, pp. 1069–1077, Mar. 1985.

[27] M. Nilsson, *Entropy and Speech*.
PhD thesis, Sound and Image Processing Laboratory, KTH Electrical Engineering, 2006.

[28] K. Paliwal, B. Schwerin, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Speech Communication*, vol. 54, no. 2, pp. 282 – 305, 2012.

[29] K. Paliwal, B. Schwerin, and K. Wojcicki, "Role of modulation magnitude and phase spectrum towards speech intelligibility," *Speech Communication*, vol. 53, no. 3, pp. 327–339, 2011.

[30] P. Mowlaee, J. Kulmer, J. Stahl, and F. Mayer, *Single Channel Phase-Aware Signal Processing in Speech Communication - Theory and Practice*.
John Wiley & Sons, 2017.

[31] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[32] N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*.
MA: MIT Press, 1949.

[33] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, pp. 1109–1121, Dec. 1984.

[34] L. A. Zadeh, "Frequency Analysis of Variable Networks," *Proceedings of the IRE*, vol. 38, pp. 291–299, Mar. 1950.

[35] T. Kailath, "Channel characterization: Time-variant dispersive channels," in *Lectures on Communication System Theory*, pp. 95–123, 1961.

[36] S. Greenberg and T. Arai, "The Relation Between Speech Intelligibility and the Complex Modulation Spectrum," in *7th European Converence on Speech Communication and Technology*, 2001.

[37] M. Blass, P. Mowlaee, and W. B. Kleijn, "Single-Channel Speech Enhancement Using Double Spectrum," *Interspeech 2016*, pp. 1740–1744, 2016.

[38] P. Clark and L. Atlas, "Time-Frequency Coherent Modulation Filtering of Nonstationary Signals," *IEEE Transactions on Signal Processing*, vol. 57, pp. 4323–4332, Nov. 2009.

[39] S. Samui, I. Chakrabarti, and S. K. Ghosh, "Speech enhancement based on modulation domain processing using coherent harmonic demodulation technique," *Electronics Letters*, vol. 53, no. 24, pp. 1611–1613, 2017.

[40] H. S. Malvar, "Lapped Transforms for Efficient Transform/Subband Coding," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 969–978, June 1990.

[41] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II - Psychoacoustic model," *Journal of the AES*, vol. 50, pp. 765–778, 2002.

[42] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 7, 2011.

[43] W. F. John Garfolo, Lori Lamel, "TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1." Philadelphia: Linguistic Data Consortium, 1993.

[44] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92 database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[45] F. Huang, T. Lee, and W. B. Kleijn, "Transform-domain Speech Periodicity Enhancement with Adaptive Coefficient Weighting," *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–5, Dec. 2011.

[46] F. Huang, T. Lee, and W. B. Kleijn, "A method of speech periodicity enhancement based on transform-domain signal decomposition," in *18th European Signal Processing Conference (EUSIPCO)*, pp. 984–988, Aug. 2010.

[47] R. McAulay and M. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 137–145, Apr. 1980.

[48] M. Singh, S. Y. Low, S. Nordholm, and Z. Zang, "Bayesian noise estimation in the modulation domain," *Speech Communication*, vol. 96, pp. 81–92, 2018.

[49] K. Paliwal, B. Schwerin, and K. Wojcicki, "Single channel speech enhancement using MMSE estimation of short-time modulation magnitude spectrum," in *Interspeech 2011*, pp. 1209–1212, ISCA, 2011.

[50] T. Gerkmann and R. Martin, "Empirical distributions of DFT-domain speech coefficients based on estimated speech variances," in *International Workshop on Acoustic Signal Enhancement*, 2010.

[51] F. J. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[52] J. E. Porter and S. F. Boll, "Optimal estimators for spectral restoration of noisy speech," *IEEE ICASSP*, pp. 18A.2.1–18A.2.4, 1984.

[53] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 7, pp. 1110–1126, 2005.

[54] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 578–589, Oct. 1994.

[55] H. Hermansky, E. Wan, and C. Avendano, "Speech enhancement based on temporal processing," *Proceedings of ICASSP*, vol. 1, pp. 405–408, June 1995.

[56] M. Hasegawa-Johnson, "Landmark-based speech recognition."
University of Illinois at Urbana-Champaign, USA, PowerPoint presentation.

[57] T. E. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*.
Prentice Hall PTR, 2002.

[58] J. Lyons and K. Paliwal, "Effect of Compressing the Dynamic Range of the Power Spectrum in Modulation Filtering Based Speech Enhancement," in *Interspeech 2008*, pp. 387–390, 2008.

[59] I. Saratxaga, D. Erro, I. Hernaez, I. Sainz, and E. Navas, "Use of Harmonic Phase Information for Polarity Detection in Speech Signals," in *Interspeech 2009*, pp. 1075–1078, 2009.

[60] A. Savitzky and M. J. E. Golay, "Smoothing and Differentiation of Data by Simplified Least Squares Procedures," *Analytical Chemistry*, vol. 36, no. 8, pp. 1627–1639, 1964.

[61] T. Gerkmann and R. Hendriks, "Noise power estimation based on the probability of speech presence," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2011.

[62] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.

[63] R. Martin, "Bias compensation methods for minimum statistics noise power spectral density estimation," *Signal Processing*, vol. 86, pp. 1215–1229, June 2006.

[64] J. Taghia, J. Taghia, N. Mohammadiha, J. Sang, V. Bouse, and R. Martin, "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4640–4643, May 2011.

[65] "P.862 Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs."

[66] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 518–530, Feb. 2014.