Graz University of Technology

Master Thesis

Single-Channel Speech Enhancement Using Double Spectrum

conducted at the
Signal Processing and Speech Communications Laboratory
Graz University of Technology, Austria

by
Martin Blaß, 0931987

Supervisors:
Dr. Pejman Mowlaee

Assessors/Examiners:
Dr. Pejman Mowlaee

Graz, June 2, 2016

Acknowledgements

First of all, I would like to thank my supervisor Dr. Pejman Mowlaee for his continuous support throughout the last year that I have been working on this thesis. I deeply appreciate his dedication and interest in research, his availability and most of all his helpful comments which guided me and kept me on the right track with my work. I am very thankful for this pleasant collaboration.

I want to thank Prof. Bastiaan Kleijn for his thoughtful advice at different stages of my work. I am happy to have had the opportunity of collaborating with him and to have benefited from his wisdom and experience.

Furthermore, I would like to thank my colleagues with whom I shared the pleasure of working on our master's theses. I guess the time we spent together always kept me motivated and focused, and helped me to accomplish my goals step by step.

Last but not least, my deepest gratitude goes to my fiancée, my family and my friends, who are always there for me and support me in any possible way.

Abstract

Single-channel speech enhancement plays an important role in mobile communications, automatic speech recognition, hearing aids and similar applications. In the past, many speech enhancement algorithms have been formulated in the Short-Time Fourier Transform (STFT) domain. However, the STFT, as predominant choice in speech enhancement, does not lead to a sparse signal representation for speech signals. As an alternative, several studies have reported advantages of using two-dimensional frequency domain representations comprised of acoustic frequency and modulation frequency. In human speech, temporal modulations convey information which has found to be relevant for the intelligibility. In this thesis, we propose the *Double Spectrum* (DS) obtained by combining a pitch-synchronous transform followed by a modulation transform. First, we discuss the relevance of temporal modulation in human speech and investigate different speech signal representations in a comprehensive literature review. Second, we explain the fundamentals of the DS analysis and synthesis procedures and present prominent properties of this domain. In experiments, we show the effectiveness of DS-based speech enhancement methods by comparing them to STFT-based and modulation-based benchmarks. The performance evaluation is conducted by means of objective measures for speech quality and intelligibility. In the course of this work, other applications relevant to speech signal processing, such as pitch and speech presence probability estimators, have been developed using the DS representation. All experiments and algorithms of this work were implemented in MATLAB.

Kurzfassung

Einkanalige Sprachsignalverbesserung spielt eine wichtige Rolle in Anwendungsgebieten wie Mobilkommunikation, automatischer Spracherkennung und Hörgeräten. In der Vergangenheit wurden viele Sprachverbesserungsalgorithmen in der Domäne der Kurzzeit-Fourier-Transformation (STFT) beschrieben. Die STFT, die als vorherrschende Transformationsdomäne für Sprachsignalverbesserung gilt, liefert jedoch keine kompakte Repräsentation für Sprachsignale. Als Alternative wird in diversen Studien über die Vorteile von zweidimensionalen Darstellungen im Frequenzbereich berichtet, die eine Kombination aus akustischer Frequenz und Modulationsfrequenz darstellen. Zeitliche Modulationen in Sprachsignalen beinhalten Informationen, die für die Sprachverständlichkeit als relevant befunden wurden. In der vorliegenden Arbeit präsentieren wir *Double Spectrum* (DS) als Kombination einer Transformation, welche auf der Periodendauer der Grundfrequenz basiert, und einer Modulationstransformation. Zunächst behandeln wir die Bedeutung zeitlicher Modulation in der menschlichen Sprache und untersuchen verschiedene Sprachsignaldarstellungen anhand einer Literaturrecherche. Anschließend beschreiben wir die Grundlagen von DS Analyse und Synthese und erläutern Eigenschaften dieser Domäne. In Experimenten demonstrieren wir die Effektivität von DS-basierten Sprachsignalverbesserungsalgorithmen durch Vergleiche mit STFT- und modulationsbasierten Referenzmethoden. Zur Evaluierung werden objektive Maße für Sprachqualität und Verständlichkeit herangezogen. Im Laufe dieser Arbeit wurden weitere Anwendungen im Feld der Sprachsignalverarbeitung entwickelt, wie z.B. ein Schätzer der Grundfrequenz und der Wahrscheinlichkeit von Sprachaktivität auf Basis von DS. Alle Experimente und Algorithmen dieser Arbeit wurden in MATLAB implementiert.

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 sources used. The text document uploaded to TUGRAZonline is identical to the present Master's thesis.

date	(signature)

Contents

1 Introduction

In the last few decades digital signal processing of speech signals has grown of great importance in several fields of application. Speech transmission, coding and synthesis, as well as automatic speech recognition (ASR) and speech enhancement pose typical examples. Speech enhancement, in particular, is an essential part of communication networks, mobile telephony, hearing aids etc. and is becoming increasingly important. Since users expect these applications to work anywhere and at any time, this imposes heavy demands on the robustness of speech enhancement algorithms. Additionally, the perceived quality and intelligibility of speech are driving forces behind further scientific and technological development in the field of speech enhancement.

A common approach to improve speech signals is to reduce the impact of noise and acoustic disturbances, which may occur due to traffic, competing speakers, office environment etc. This can be done using single-channel or multi-channel noise reduction methods. In the single-channel case, the noisy speech signal is captured by only one microphone. While multi-channel methods, i.e. multiple microphones, often lead to a better performance than their single-channel counterparts, their usage is limited by computational complexity, spatial requirements and additional costs [1, 2].

In this thesis, we investigate single-channel speech enhancement using a novel transform domain, called Double Spectrum (DS). This domain combines a pitch-synchronous transform and a modulation transform, and leads to a compact representation designed for speech signals. In the course of this work we explore the fundamentals of the DS domain and use its beneficial properties for the design of a DS-based speech enhancement system. The performance of the proposed system will be compared to conventional speech enhancement methods. The thesis is structured as follows. Chapter 2 consists of a review about human speech production, temporal modulations in speech, digital signal processing and representations of speech signals. Chapter 3 introduces the novel transform domain, Double Spectrum, and explains its fundamentals. In Chapter 4, we derive a pitch estimation algorithm and a method for speech presence probability estimation using the DS transform. Chapter 5 presents different noise suppression rules for speech enhancement performed in the DS domain. In Chapter 6, we describe the experimental setup and report performance evaluation results of the proposed DS-based speech enhancement methods in terms of objective measures for speech quality and intelligibility. Throughout further experiments, we compare one of our methods to popular speech enhancement benchmarks to show its effectiveness. Chapter 7 concludes the work and gives a future outlook about the potentials and limits of DS-based speech enhancement.

2

Speech Signal Processing in the Modulation Domain

2.1 Speech Production

Human speech is the outcome of the interaction of many physiological components within the speech production system. The air flow that is needed for generating speech is originated from the lungs. The diaphragm pushes the ribcage to pump the air through the trachea straight to the larynx where the glottis is situated. The glottis denotes the space between the vocal cords and is understood as the sound source for speech. The larynx is responsible for converting the air flow into an acoustic excitation-source signal [3]. The vocal cords typically vibrate at a frequency of 50-250 Hz and 120-500 Hz for male and female, respectively [2]. This leads to an output of periodic pulses at the glottis, where the length of one period is called pitch period (P_0). The relation between the fundamental frequency f_0 (also known as *pitch*) and P_0 is

$$f_0 = \frac{1}{P_0}. \tag{2.1}$$

By colliding with themselves the vocal cords can produce a large number of harmonics, which are typically integer multiples of f_0. The vocal tract acts like a resonating filter modifying the excitation signal and produces formant frequencies. Depending on the shape of the vocal tract, up to five strong formants can be produced [3]. The speech signal is dynamically shaped and controlled by articulators along the path, such as velum, tongue, teeth and lips.

Besides *voiced* sounds, which require a periodic excitation signal, speech can be categorized into *unvoiced* and *plosive* sounds. *Unvoiced* sounds, e.g. /ss/ or /f/, are made from a non-periodic turbulent air flow as the vocal cords remain relaxed. In this case the excitation signal shows properties of white Gaussian noise and has no f_0 and thus no harmonic structure [3]. *Plosives*, such as /b/ or /p/, are caused by a sudden release after air pressure was built up somewhere in the vocal tract. The released flow may create a voiced or an unvoiced sound, or even a mixture of both, depending on the constellation of the articulators [2]. Figures 2.1 and 2.2 show exemplary waveforms of a voiced and an unvoiced sound, respectively.

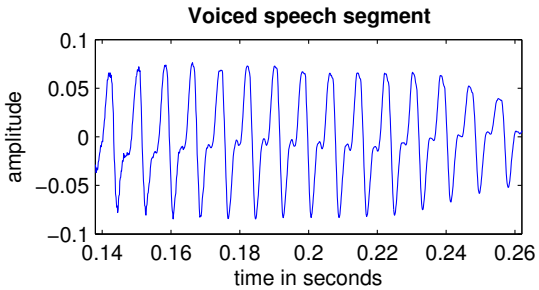


Figure 2.1: Signal segment of a voiced sound.

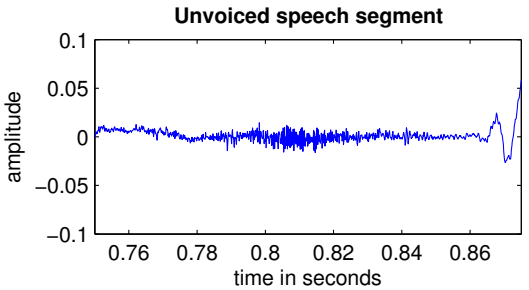


Figure 2.2: Signal segment of an uvoiced sound.

2.2 Importance of Temporal Modulation in Speech

For a long time it was believed that the spectral envelope is the principal carrier of information in human speech. More recently, several studies report that temporal modulations of the spectral envelope contain information which is relevant to the speech intelligibility [4]. The relevance of modulation in speech has been shown by a number of physiological and psychoacoustical studies:

Early research on modulations in human speech was conducted by Dudley (1940) who pictured a speaker as a sort of radio broadcast transmitter producing amplitude modulation (AM) signals with an audible carrier and inaudible "message-bearing waves". These waves originate from muscular wave motions of the vocal tract occurring at syllabic rates. He suggested that the message stored in the modulations is then recovered by the listener's mind [5].

In 1952, Zwicker investigated the auditory system with respect to its sensitivity for modulation frequencies. It was found that the human ear can distinguish between both amplitude and frequency modulation, and is sensitive to AM at modulation frequencies below 30 Hz for low acoustic frequency ($f_0 < 250$ Hz) [6]. Follow-up experiments of Viemeister (1979) and Bacon (1989) computed modulation transfer functions and modulation thresholds to characterize the ear's sensitivity to modulations in more detail. The results showed that the auditory systems features channels which are specifically tuned for the detection of modulation frequency [7, 8]. Similar findings were also obtained by Sheft and Yost (1990) [9].

Schreiner and Urbas (1986) examined the ability of cortical neurons to follow AM of tones in mammals and showed that a neural representation is preserved through all levels of the auditory system [4]. Further physiological research was conducted by Kowalski et. al (1996) who investigated neural responses with respect to their temporal and spectral properties. He found that neural cells are best driven by a combination of spectral and temporal modulations [10]. Similarly, Mesgarani and Shamma (2005) reported that neurons in the auditory cortex decomposed the spectrogram into its spectro-temporal modulation content and further used this knowledge to derive a noise suppression algorithm [11].

Interesting results about the relationship between temporal modulations and intelligibility of speech were obtained by different studies of Drullman in 1994: Reducing low-frequency modulations in the temporal envelope of noisy sentences led to degraded performance in phoneme identification (in particular for consonants) and severe reduction in sentence intelligibility. First, this was tested by applying low pass and band pass filters with different cut-off frequencies to modify the amplitude envelope of noisy speech [12]. Follow-up experiments with high pass filters [13] yielded similar results and led to the conclusion that modulation frequencies between 4 and 16 Hz are most important for speech intelligibility. Both of Drullman's works had strong impact on modulation based speech processing in the future [3].

In a similar work, Arai (1996) presented the effect of filtering time trajectories of spectral envelope on speech intelligibility. Low-pass, high-pass and band-pass filters were applied on speech signals which were first processed by Linear Prediction Coding (LPC). The results of perceptual experiments showed that intelligibility was not severely impaired as long as modulation rates between 1 Hz and 16 Hz were preserved [14].

In their extensive research, Atlas and Shamma (2003) concluded that low-frequency modulations of sound are the fundamental carrier of information in speech. They argued that there is considerable evidence, that our perception of sound uses features which are related to underlying signal modulations. Motivated by their finding, a modulation spectral model was proposed [15].

These results and findings regarding the importance of modulations in speech marked the start of modulation based speech processing applications with particular focus on improving existing methods by taking into account beneficial characteristics of temporal modulation.

2.3 Representation of Speech Signals

The goal of a signal representation is to describe particular characteristics of a signal. A representation consists of a model description, model parameters and signal coefficients [16]. We discriminate between parametric and non-parametric representations. If a representation consists of signal coefficients only it is called non-parametric, since no further parameters are used to describe a signal. An example for a non-parametric representation would be the pulse-code modulation in audio signal processing. A model-based representation, such sinusoidal coding, relies on model-parameters and coefficients and is thus referred to as a parametric representation. An appropriate signal representation may improve the accuracy and efficiency of a certain algorithm. Good representations often require relatively few coefficients per unit time for an accurate description of the signal, while still being complete and hence able to describe any signal. This is consistent with the energy of coefficients being concentrated in a small subspace after an energy preserving transform. This property is also referred to as the compactness or sparseness of a representation. Completeness, on the other hand, indicates that a signal can be perfectly reconstructed if the representation is known [17].

Many speech enhancement algorithms are based on an analysis-modification-synthesis (AMS) framework in frequency domain, e.g. obtained by a filter bank approach or Fourier transform. In particular, single-channel speech enhancement is often formulated in the Short-Time Fourier Transform (STFT) domain. We argue that the STFT, the predominant choice in speech enhancement (see e.g. [1] for an overview), while complete, generally does not lead to a sparse signal representation for speech.

In the following, the fundamentals of Fourier analysis will be discussed to gain basic understanding, how it differs from the proposed DS transform domain described in Chapter 3.

2.3.1 Fourier Analysis

The purpose of a spectral transformation, such as the Fourier transform, is to provide an analysis of a signal in terms of its spectral components. The Fourier transform relates a time domain signal $x(t)$ to its frequency domain representation [2]:

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

where t is the time index and $\omega = 2\pi f$ denotes the radian frequency. The resulting spectrum is continuous in both time and frequency. In digital signal processing, however we deal with discrete-time sampled speech. This means, that the signal is only observable at time instances which are multiple integers of the sampling period T. The sampling period corresponds to the rate at which signal realizations are obtained, which is known as the sampling frequency (f_s):

$$f_s = \frac{1}{T}. \tag{2.3}$$

The Discrete-Time Fourier Transform (DTFT) is defined as

$$X(e^{j\Omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\Omega n} \tag{2.4}$$

with n given as the discrete time index and Ω as the normalized radian frequency $\Omega = 2\pi fT$.

In practice, only finite signal segments are available, since block processing is applied in digital signal processing. This is equivalent to applying a window function to the signal, where the length of the window determines the length of the signal segment. In addition, both time and frequency need to be discrete for the implementation of digital signal processing applications. To obtain discrete frequency, the Discrete Fourier Transform (DFT) is used:

$$X(k) = \sum_{n=-\infty}^{\infty} x(n) e^{-j\frac{2\pi kn}{N}}, \tag{2.5}$$

where k is the frequency index and N is the length of the signal segment $x(n)$. The DFT represents the spectrum of a signal at equally spaced points defined by $\Delta\Omega = 2\pi/N$ on the normalized frequency axis. When $x(n)$ is a signal sampled at f_s, the coefficients are spaced by $\Delta f = f_s/N$ [2].

By using a sliding window function over time we are able to analyze spectral signal components at any time instance. Due to redundancy in consecutive values of $X(k)$ and computational complexity, it is common to introduce a decimation factor [18] corresponding to a frame shift, so that $X(k)$ is computed only every Z-th sample. The choice of the decimation factor depends on the desired time-frequency resolution [3]. Taking into account values outside the sliding window as zero, the discrete-time running STFT [2] is defined as

$$X(l,k) = \sum_{n=0}^{N-1} x(n+lZ) w(n) e^{-j\frac{2\pi kn}{N}}, \tag{2.6}$$

where l is the frame index, Z is the frame shift and w is a window function.

The coefficients obtained by any kind of Fourier transform are complex valued, which means they may be written in terms of their real and imaginary parts

$$X(e^{j\Omega}) = \mathrm{Re}\{X(e^{j\Omega})\} + j\ \mathrm{Im}\{X(e^{j\Omega})\}. \tag{2.7}$$

Using polar notation the coefficients are split into magnitude and phase as

$$X(e^{j\Omega}) = |X(e^{j\Omega})| e^{j\phi(e^{j\Omega})}, \tag{2.8}$$

where $|X(e^{j\Omega})|$ and $\phi(e^{j\Omega})$ are called magnitude and phase spectrum, respectively. Other well known properties of the Fourier transform are the *linearity*, i.e. both *homogeneity* and *additivity* are valid, yielding

$$ax_1(n) + bx_2(n) \circ\!\!-\!\!\bullet\, aX_1(e^{j\Omega}) + bX_2(e^{j\Omega}), \tag{2.9}$$

and the *symmetry* given $x(n)$ as a real valued signal

$$X(e^{j\Omega}) = X^*(e^{-j\Omega}), \tag{2.10}$$

where $a, b \in \mathbb{R}$ and * denoting complex conjugation [2]. From the symmetry property we get that N/2 DFT coefficients can be used to describe spectral components up to the Nyquist frequency which corresponds to $\Omega = \pi$ in radians or $f = f_s/2$ in Hz [19, 20].

Discussing the synthesis and further properties of the Fourier transform family would be out of the scope of this thesis. The properties listed above will serve the purpose of drawing a comparison to the newly introduced transform domain, called Double Spectrum (DS). For more information about Fourier transforms, the reader is referred to signal processing textbooks, e.g. [2, 19].

2.3.2 Modulation Domain

In the past, STFT-based speech signal processing proved itself effective for different applications. Recently, the modulation domain has become more popular due to many interesting findings of physiological and psychoacoustical research, as discussed in Section 2.2. The first proposal to use a signal representation which takes into account modulation was perhaps Zadeh (1950) who suggested a two-dimensional bi-frequency spectrum as a combination of acoustic and modulation frequency [21]. In a more recent work, a modulation domain representation was defined by Atlas et al. [15] in 2003 and used for single-channel source separation and audio coding [22]. A simplified structure of their two-dimensional transform, derived from a time domain aliasing cancellation filter bank and a modified Discrete Cosine Transform (DCT), is depicted in Figure 2.3. In their successive work Atlas et al. (2004) defined a similar two-dimensional transform representation using a two-stage STFT procedure: A first base transform was applied to the input signal yielding a time-frequency representation, and a subsequent second transform was performed across time frames resulting in what they referred to as "modulation spectrum". The acoustic frequency commonly corresponds to the frequency axis of the first transform and modulation frequency corresponds to the independent variable of the second transform [23].

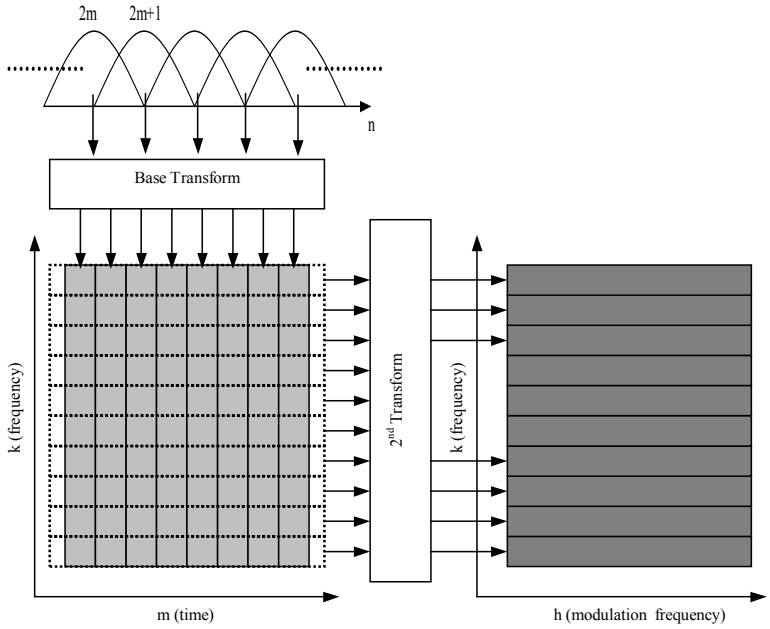


Figure 2.3: Simplified structure of the two-dimensional transform used in an audio coding approach [15]

In 2010 and 2012, Paliwal et al. used the same modulation representation in the STFT-domain to implement popular single-channel speech enhancement methods such as spectral subtraction [24] and Minimum Mean-Square Error (MMSE) Magnitude Estimator in the Short-Time Spectral Modulation (STSM) domain [4, 25]. These two methods, in the remainder referred to as *ModSpecSub* and *MME* respectively, will be used later in this thesis as benchmark methods for the proposed DS-based speech enhancement algorithm. The input signal is first processed using the running STFT as given in (2.6) to obtain the acoustic spectrum $X(l,k)$, where l and k denote acoustic frame index and frequency index, respectively. The acoustic magnitude spectrum $|X(l,k)|$ was then used to derive the modulation spectrum as follows. The time trajectories for each frequency component of $|X(l,k)|$ are processed framewise using a second running STFT

$$\mathcal{X}(\eta,k,m)=\sum_{l=0}^{\mathcal{N}-1}|X(l+\eta\mathcal{Z})|u(l)e^{-j\frac{2\pi lm}{\mathcal{N}}}, \tag{2.11}$$

where η is the modulation frame index, k is the acoustic frequency index, m refers to the modulation frequency index, $\mathcal{N}$ is the modulation frame index in terms of acoustic frames, $\mathcal{Z}$ is the modulation frame shift in terms of acoustic frames and $u(l)$ is the modulation analysis window function. Since the STFT is used to analyze successive spectral frames, the resulting modulation spectrum can be written in polar form as

$$\mathcal{X}(\eta, k, m) = |\mathcal{X}(\eta, k, m)|e^{j\angle\mathcal{X}(\eta,k,m)}, \tag{2.12}$$

where $|\mathcal{X}(\eta, k, m)|$ is the modulation magnitude spectrum and $|\angle\mathcal{X}(\eta, k, m)|$ is the modulation phase spectrum [25]. Unlike the acoustic phase spectrum which is often regarded as unimportant and for speech enhancement, the modulation phase spectrum is said to contain useful information according to Hermansky (1995) [26]. However, in the implementation of *ModSpecSub* and *MME* [4, 25], $|\angle\mathcal{X}(\eta, k, m)|$ was left unchanged for simplicity reasons. The AMS-based framework for speech enhancement in the STSM domain is shown in Figure 2.4.

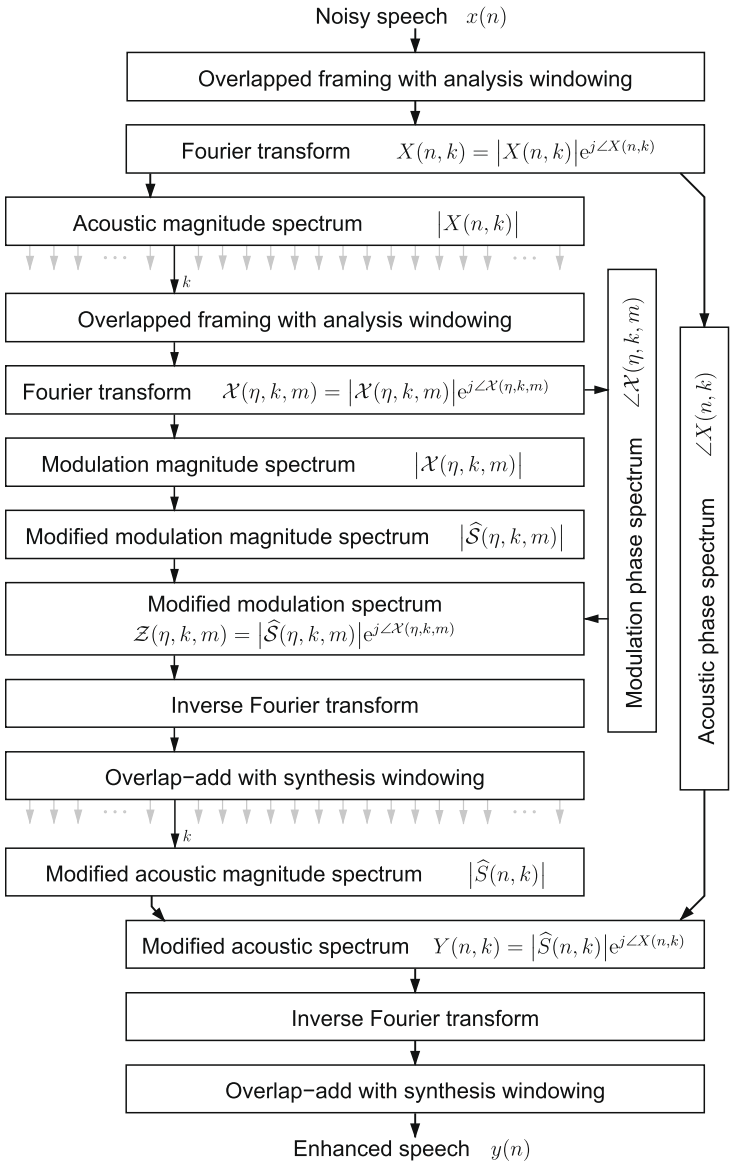


Figure 2.4: Block diagram of the AMS-based framework for speech enhancement in the STSM domain [4]

2.3.3 A Canonical Representation using a Two-Stage Transform

In the previous Sections we presented ideas how to represent speech signals using their spectral information in frequency domain and further their spectro-temporal information in modulation domain. In contrast to the STSM domain, which is based on the STFT as shown in 2.3.2, other modulation domain representations use pitch-synchronous transforms to takes into account an important feature of speech: the pitch.

The motivation for using pitch to define a signal representation is that voiced speech features a strong periodic structure due to the vibration of the vocal cords. During voicing the fundamental frequency f_0 shows only slight deviances, i.e. remains quasi constant for short time intervals. The resulting periodic structure implies a high redundancy in the signal, which can be exploited in speech coding or enhancement techniques [16]. In a work of Nilsson (2007), a "canonical" signal representation of speech was proposed, which has the efficiency, in terms of being compact, similar to that of parametric modeling and additionally has the property of completeness, which guarantees perfect signal reconstruction [17]. This representation considers two prominent features of speech: short-term dependencies due to the resonances of the vocal tract and long-term dependencies associated with the pitch. The resulting canonical representation is a continuation of the ideas presented in previous works of Kleijn (1993, 2000) about sinusoidal coding and waveform interpolation [27, 28].

The system proposed in [17] consists of four processing blocks: Linear Prediction (LP) analysis, constant pitch warping, pitch-synchronous transform, and modulation transform. Figure 2.5 shows a block diagram of the proposed system.

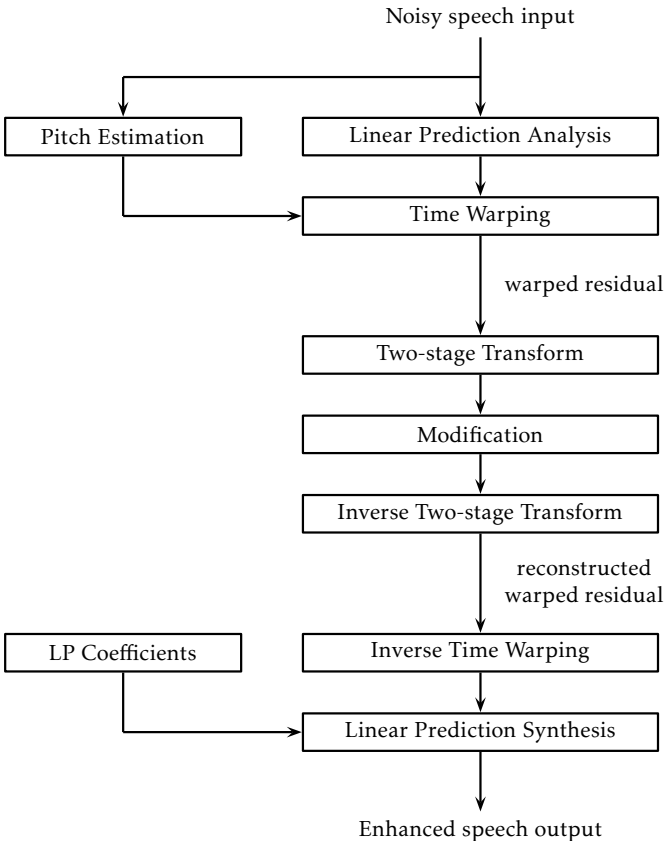


Figure 2.5: Block diagram for a canonical speech representation system [16]. The input signal is processed using LP analysis and a time warping procedure to get a residual signal of constant pitch. This warped residual undergoes a two-stage transform to arrive at a compact signal representation tailored for prosodic modification, coding and other speech related applications.

The LP analysis is used as an autoregressive (AR) model to capture vocal tract related short-term dependencies within the speech signal. This procedure is often used in conventional coding systems such as the adaptive differential pulse-code modulation or the code excited linear prediction [2]. The AR model of a discrete-time signal $s(n)$ is defined as

$$s(n) = \sum_{m=1}^{M} a_m^M s(n-m) + e(n), \tag{2.13}$$

where M is the model order, a_m^M is the set of parameters specifying the AR model and $e(n)$ denotes the prediction error (residual) [16]. The parameters a_m^M are often referred to as LP coefficients and estimated from speech segments (typically 20 ms) through LP analysis [29].

For voiced speech the duration and shape of consecutive pitch cycles change slowly in general, which implies redundancy in long-term dependencies of speech. However, since pitch is varying over time, it is a difficult task to design a signal transform that is capable of concentrating the energy of its coefficients in a small subspace and hence guarantee compactness. For this reason, time warping is proposed, which warps the LP residual $e(n)$ into a signal of constant pitch $e_{wrp}(n)$. This facilitates a compact representation obtained by two subsequent transform stages. The warper performs time scaling, i.e. relating the original time domain t to a new warped time domain τ, in which pitch is constant. The warping function $t(\tau)$ compares signal components to components delayed by one pitch period and seeks the optimal B-spline coefficients that minimize the squared error. This process can also be understood as a form of irregular over-sampling of the original LP residual to make it of constant pitch [16, 17]. From the mapping performed by $t(\tau)$ it is possible to derive an inverse mapping $\tau(t)$, which is needed in the reconstruction stage to recover the original time scaling. A detailed description of the warper is found in [30]. Figure 2.6 shows an example for the time warping procedure to obtain a signal of constant pitch.

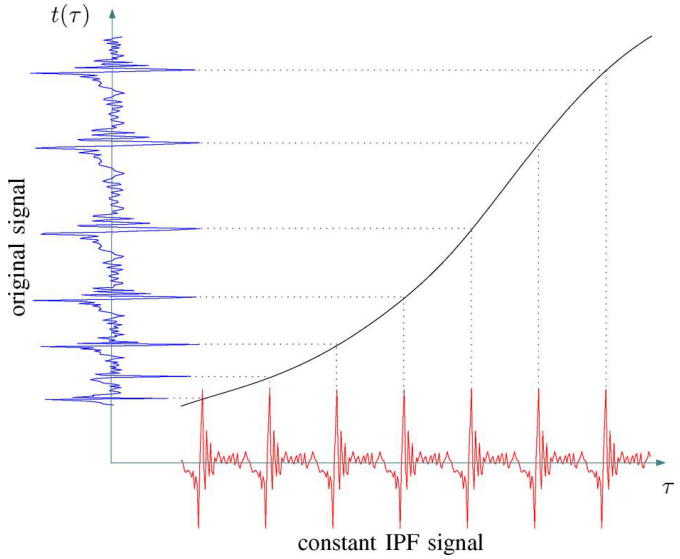


Figure 2.6: Example for the warping function $t(\tau)$ to map the original signal to a signal with constant instantaneous pitch frequency (IPF) [30].

The combination of pitch-synchronous and modulation transform, referred to as a *two-stage transform*, results in lapped frequency transforms, which approximates the Karhunen-Loève Transform (KLT) for stationary signal segments [31]. The KLT maximizes the coding gain, which can be seen as a particular form of energy concentration [16, 17]. In [17], using such a two-stage transform has the purpose of exploiting features of the warped LP residual in order to achieve a highly energy concentrated representation: Considering a speech signal belonging

to a steady voiced sound, $e_{wrp}(n)$ consists of a sequence of similarly shaped pitch cycles. As a result, the *pitch-synchronous transform* produces a sequence of modulated lapped transform (MLT) coefficients which change slowly over time. Applying another transform, referred to as *modulation transform*, on this sequence yields information about how spectral coefficients evolve over time: Coefficients of low modulation bands represent slowly evolving signal components, whereas coefficients of high modulation bands correspond to rapidly changing signal components [17]. An example of a voiced speech segment analyzed by means of the system proposed in [17], consisting of LP analysis, time warping and two-stage transform is presented in Figure 2.7.

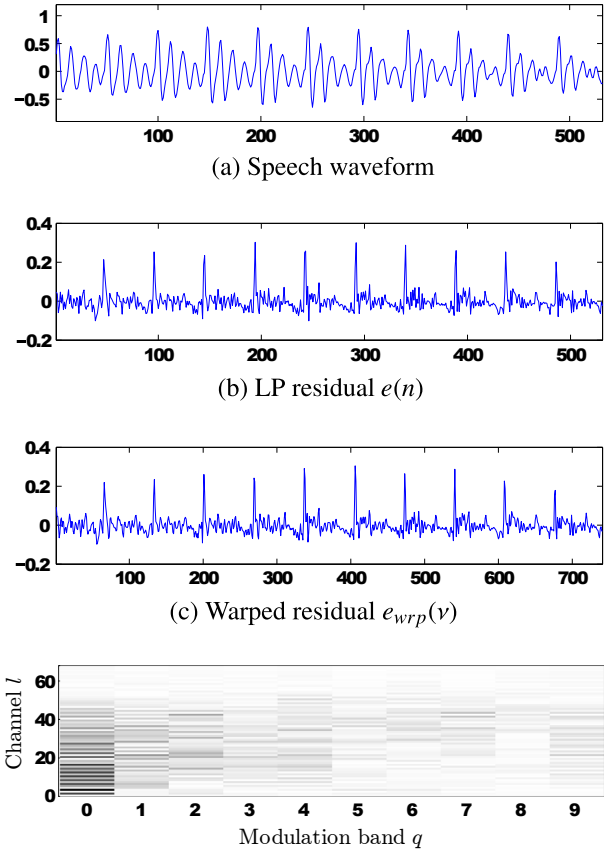


Figure 2.7: Example for constant-pitch warping and subsequent two-stage transform of a voiced speech segment [32]: (a) shows the waveform of the input speech segment, (b) and (c) show the LP residual and its warped outcome, and (d) presents the transform coefficients of the two-stage transform, where q and l denote modulation band and frequency channel, respectively. The pitch period given in samples is $P_0 = 68$ at a sampling rate of $f_s = 8$ kHz.

The selection of modulation window lengths, i.e. the number of successive MLT coefficients, for the modulation transform is based on an energy concentration criterion which assigns short windows to rapidly changing regions and longer windows to steady regions. This is done by a best-basis selection using the basis functions of the transform. The initial length of a window is extended as long as there is an increase in the energy concentration of the transform coefficients [17]. As a by-product of the compact representation thus obtained, the two-stage transform allows the identification of periodic and aperiodic signal components, which in the case of speech signals facilitates the separation of voiced and unvoiced components [16]. In successive works by Huang et al. (2011,2012,2015) the two-stage transform domain was extended to speech enhancement by periodicity enhancement using different coefficient weighting procedures [32–34]. In these works, noise reduction was achieved by modifying the energy balance of the modulation bands in the two-stage transform domain, which restored the harmonicity of noise corrupted

speech.

In general, the idea of using two-dimensional transforms for signal processing applications has been very popular in image processing in the past decades: Several image coding techniques such as *JPEG* compression (established by the Joint Photographic Experts Group) apply the Discrete Cosine Transform (DCT) to the rows and columns of an image to get a compact signal representation in frequency domain [35,36]. On the other hand, the methods presented in [4,17, 22,32] pose examples which prove the effectiveness of two-dimensional transform representations for audio or speech signal processing applications as well.

3

Double Spectrum: Fundamentals

Using the two-stage transform, defined in [17], as a two-dimensional speech representation comprised of both acoustic and modulation frequency, we obtain a novel signal representation which we refer to as Double Spectrum (DS). Figure 3.1 shows the framework for the canonical speech representation system [17] and how it is modified for a DS framework. Our goal is to apply the two-stage transform directly on the noisy signal and use block processing to selectively modify voiced and unvoiced time blocks to achieve noise suppression. For this reason, the proposed DS approach does neither rely on LP analysis, time warping, nor on a best-basis selection, in contrast to methods described in [17, 32–34]. This serves the purpose of overcoming the need of over-sampling and other implementation issues related to the adaptive number of modulation bands, which makes our method simpler and faster.

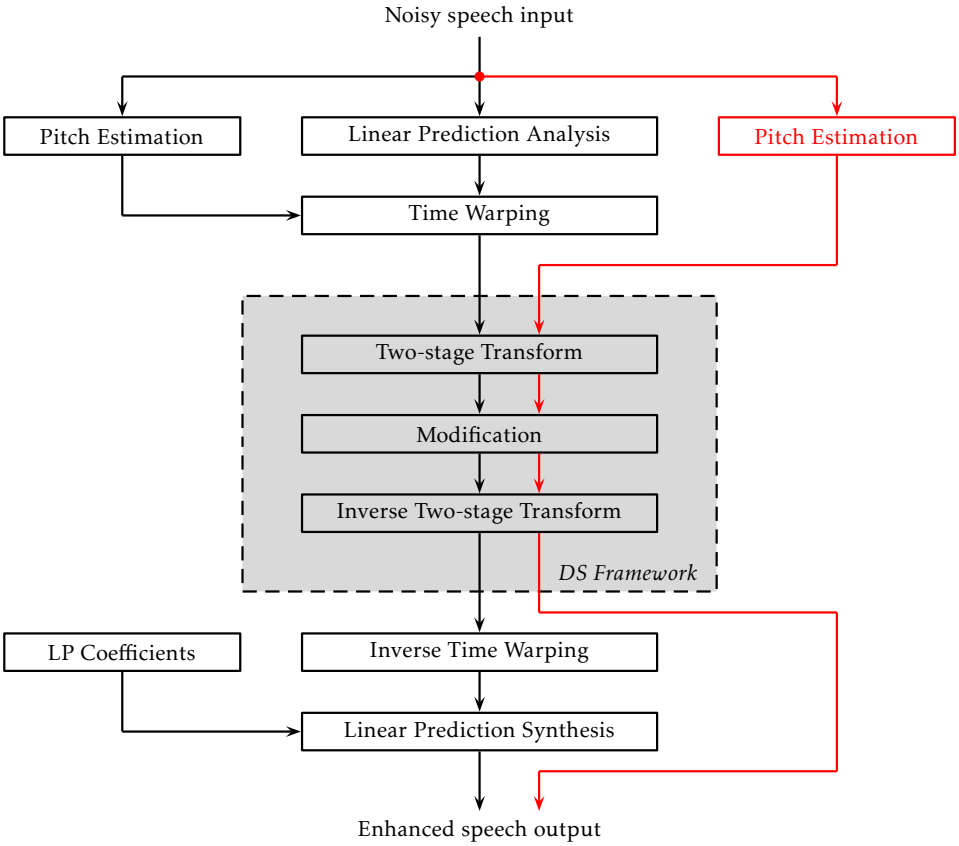


Figure 3.1: Using the block diagram for a canonical speech representation system [16] as the basis of the proposed DS framework: The red line shows the signal processing flow of a DS-based system, bypassing the LP analysis and time warping blocks proposed in [17]. The highlighted grey block shows a simplified version of the DS framework comprised of a two-stage transform, signal modification in the DS domain and an inverse transform.

Before we perform DS analysis in terms of the two-stage transform, a pitch-synchronous time block segmentation (TBS) needs to be applied to the input signal. This is done in order to take into account the time-varying nature of speech: Since pitch is varying over time, and both the pitch-synchronous transform and the modulation transform do not adapt to this property, we introduce block processing under the assumption of quasi-stationarity of speech [2, 19]. The signal segments thus obtained are then analyzed in terms of their modulation-spectral components. In the following Sections, 3.1 and 3.2, the essential steps of DS analysis and synthesis are described in detail.

3.1 Analysis

3.1.1 Time Block Segmentation

First, an estimated pitch trajectory needs to be extracted from the input signal, which is then used to determine the instantaneous pitch period P_0 of the corresponding time block. The pitch-synchronous TBS is thus a function of P_0 and dependent on the f_0 estimation. A detailed discussion on the topic of the latter is given in Section 4.1. The TBS separates the input speech signal into L time blocks of variable length. The length of each time block is an integer multiple of the normalized pitch period $\tilde{P}_0$ (in samples), which is defined as

$$\tilde{P}_0 = \frac{f_s}{f_0} \tag{3.1}$$

and computed from the f_0 estimate of the respective signal segment. Since f_0 varies over time, this means that we ignore its local variation during the segmentation. A time block is further subdivided into $\mathcal{L}$ frames, each of length $\tilde{P}_0$. To avoid discontinuities at the transition of consecutive time blocks overlapping is introduced. The overlap also depends on $\tilde{P}_0$ and will be discussed in more detail in Section 3.4. An example for TBS performed on a clean signal is shown in Figure 3.2.

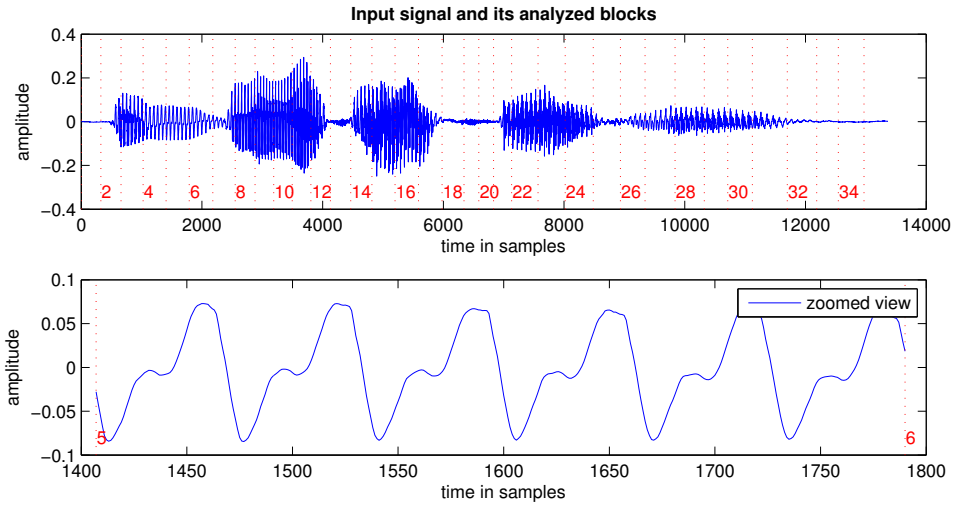


Figure 3.2: Example for pitch-synchronous time block segmentation (TBS) of a clean speech signal: In the upper graph, the dotted red lines mark the beginning of each time block and the numbers denote a time block index. The overlap between two consecutive time blocks is chosen as $\tilde{P}_0$. The lower graph presents a zoomed view of block 5, which shows the periodic waveform of a voiced sound.

3.1.2 Pitch-synchronous Transform

Each time block obtained from the TBS is analyzed in terms of the two-stage transform. For the pitch-synchronous transform we select an MLT which facilitates a critically sampled uniform filter bank with coefficients localized in the time-frequency plane. The MLT is implemented using a DCT-IV in combination with a square-root Hann window of length $2\tilde{P}_0$ and 50% overlap [17]. The usage of a square-root Hann window at analysis and synthesis stage as a matched filter satisfies the power complementarity constraint needed for perfect reconstruction, namely that a periodic extension of squared windows has to be constant [16]. The square-root Hann window $w(n)$ is defined as

$$w(n) = \begin{cases} \sqrt{\frac{1}{2}\left(1 - \cos\left(\frac{2n\pi}{N}\right)\right)} & \text{if } n = 0, ..., N-1 \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

Let $\nu = 0, 1, ..., 2\tilde{P}_0 - 1$ be a time index and let $x^{(\ell)}$ be the ℓ'th pitch-synchronous time frame of a time block, i.e. $x^{(\ell)}(\nu) = x(\ell\tilde{P}_0 + \nu)$. The first-stage transform coefficients $f(\ell, k)$ are then obtained as

$$f(\ell, k) = \sum_{\nu=0}^{2\tilde{P}_0-1} x^{(\ell)}(\nu) w(\nu) \sqrt{\frac{2}{\tilde{P}_0}} \cos\left(\frac{(2k+1)(2\nu - \tilde{P}_0 + 1)\pi}{4\tilde{P}_0}\right), \tag{3.3}$$

where $\ell = 0, 1, \ldots, \mathcal{L} - 1$ and $k = 0, 1, \ldots, \tilde{P}_0 - 1$ denote *frame* index and *frequency channel* index, respectively [32]. The output of the first transform is a sequence of MLT coefficients that slowly evolve over time for periodic signal segments, but rapidly for aperiodic segments. Due to the pitch-synchronous nature of time frames, the cardinality of frequency channels is $K = \tilde{P}_0$.

3.1.3 Modulation Transform

In the modulation transform a number of $\mathcal{L}$ consecutive MLT coefficient frames obtained by the pitch-synchronous transform are merged into one segment. For each frequency channel $k = 0, 1, ..., K - 1$ the modulation transform is performed across this segment using a DCT-II. To facilitate the implementation of the modulation transform as a critically sampled filter, we combine the DCT-II with a rectangular window [17] of length Q. The modulation coefficients $g(q, k)$ thus obtained are then given by

$$g(q, k) = \sum_{\ell=0}^{Q-1} f(\ell, k) v(q) c(q) \sqrt{\frac{2}{Q}} \cos\left(\frac{(2k+1)q\pi}{2Q}\right), \tag{3.4}$$

where $q = 0, 1, \ldots, Q - 1$ is the *modulation band* index, $v(q)$ denotes the rectangular window, $c(0) = 1/\sqrt{2}$ and $c(q) = 1$ for $q \neq 0$. Note that the number of modulation bands Q is determined by the number of pitch-synchronous frames $\mathcal{L}$ as $Q = \mathcal{L}$.

The two-stage transform procedure is performed for a time signal consisting of L time blocks with the corresponding coefficients $g^{(l)}(q, k)$, where $l = 0, 1, ..., L-1$ is the *time block* index. The definition for *Double Spectrum* is now given by $DS(q, k)$, which is equivalent to $g(q, k)$ interpreted as a matrix with K frequency channels as rows and Q modulation bands as columns. Figure 3.3 schematically visualizes a speech signal in terms of a sequence of Double Spectra, showing $DS^{(l)}(q, k)$ for a set of L time blocks $l = 0, ..., L - 1$.

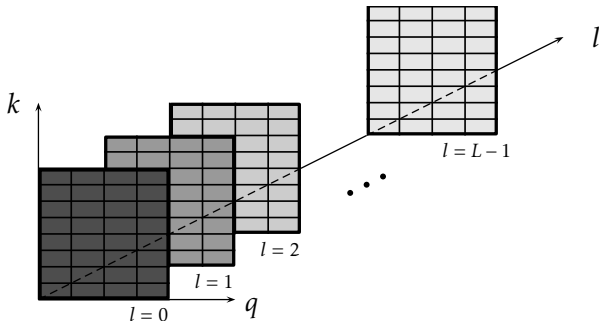


Figure 3.3: Illustration of a speech signal depicted as a sequence of Double Spectra $DS^{(l)}(q,k)$ shown for time blocks $l = 0, 1, ..., L-1$. Each time block is a signal segment comprised of $\mathcal{L}$ frames which undergo the two-stage transform and determine the number of modulation bands Q of the DS.

3.2 Synthesis

In Section 3.1 we described the DS analysis procedure which utilized the two-stage transform to obtain coefficients located in an "frequency-modulation" plane. In order to recover the time-domain signal from the DS representation, we need to apply inverse of the two-stage transform. Both the pitch-synchronous and the modulation transform have orthogonal bases, i.e. their analysis and synthesis frames are identical, except for a factor [28]. For this reason, their frame functions (bases) are called *tight* and their inverses are trivial [17]. The MLT coefficients of the pitch-synchronous transform are obtained by the expansion

$$f_{syn}(\ell,k) = \sum_{q=0}^{Q-1} g(q,k)v(q)c(q)\sqrt{\frac{2}{Q}}\cos\left(\frac{(2k+1)q\pi}{2Q}\right), \tag{3.5}$$

which is known as the inverse DCT operation. Similarly, the time-domain signal segment for the ℓ-th time frame is recovered by

$$x_{syn}^{(\ell)}(\nu) = \sum_{k=0}^{2\tilde{P}_0-1} f(\ell,k)w(\nu)\sqrt{\frac{2}{\tilde{P}_0}}\cos\left(\frac{(2k+1)(2\nu-\tilde{P}_0+1)\pi}{4\tilde{P}_0}\right). \tag{3.6}$$

Taking into account L time blocks, the signal segment corresponding to the l-th time block $x_{l,syn}(n)$ is then obtained by an overlap-and-add (OLA) procedure applied to the $\mathcal{L}$ frames as

$$x_{l,syn}(n) = \sum_{\ell=0}^{\mathcal{L}-1} x_{l,syn}^{(\ell)}(\nu). \tag{3.7}$$

In practice, a signal consists of L time blocks which are overlapped to prevent discontinuities at the segment boundaries. A detailed discussion about overlap is given in Section 3.5.4.

3.3 Double Spectra of Characteristic Signal Types

In this Section, examples for DS representations of characteristic signal types are discussed. Different patterns in DS can be observed by plotting the energy of each coefficient. For a meaningful graphical representation, it is convenient to map the energy, given by $E = |DS(q,k)|^2$, to the decibel (dB) domain in order to compress the broad dynamic range of the energy as

$$E_{dB} = 10 \cdot \log\left(|DS(q,k)|^2\right), \tag{3.8}$$

where log(.) denotes the decadic logarithm. [32]. In the following, four different types of signals are investigated in terms of their pitch-synchronous and modulation transform: a sinusoid, a clean voiced signal, a noisy voiced signal and an unvoiced signal.

3.3.1 Sinusoid

For this example a sinusoid with $f_0 = 200$ Hz is used, resulting in a normalized pitch period of $\tilde{P}_0 = 40$. Figure 3.4 shows the time-domain signal and its pitch-synchronous time blocks. The

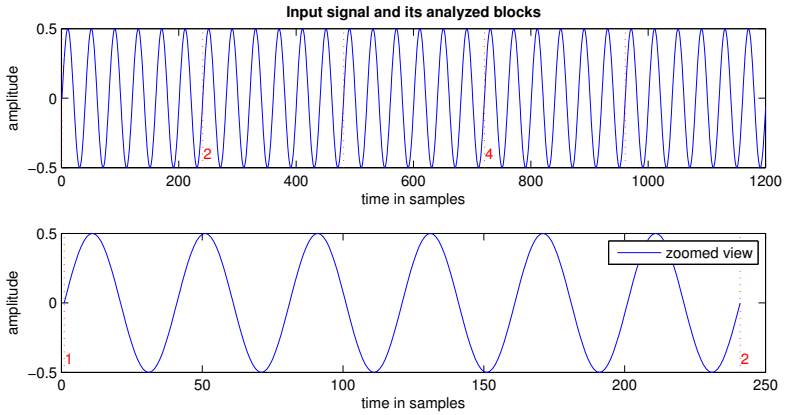


Figure 3.4: Example for TBS of a sinusoidal signal and zoomed view on a block of six pitch cycles, $f_0 = 200$ Hz and $\tilde{P}_0 = 40$, at a sampling frequency of $f_s = 8$ kHz.

pitch-synchronous transform yields a uniform filter bank and the modulation transform the DS representation as shown in Figures 3.5 and 3.6, respectively. It can be observed that a sinusoid, as the prime example of a periodic signal, yields DS coefficients concentrated only in the first modulation band $q = 0$. The energy spread across frequency k can be explained by the impact of discrete-time sampling and window choice, which introduces the *spectral leakage* effect [2,19].

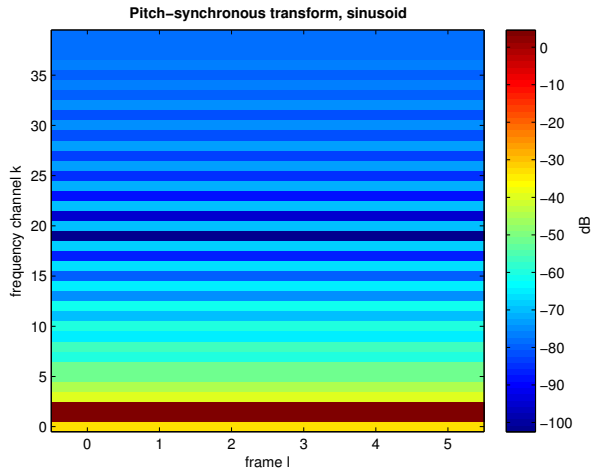


Figure 3.5: Uniform filterbank obtained by pitch-synchronous transform (DCT-IV). For a sinusoid the coefficients are constant over consecutive pitch cycles.

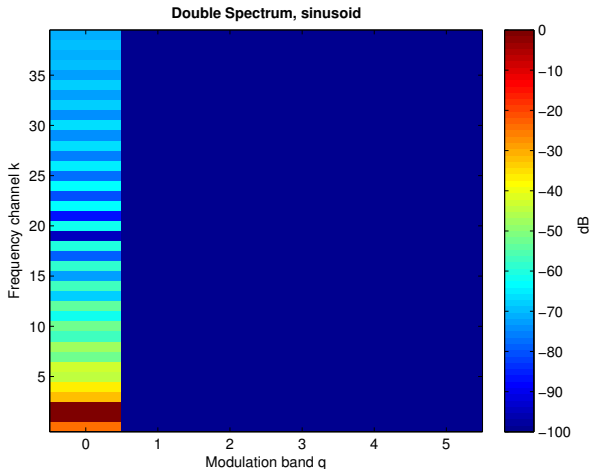


Figure 3.6: Double Spectrum obtained by modulation transform (DCT-II). For a sinusoid the DS coefficients are concentrated only in the first modulation band.

3.3.2 Clean Voiced Signal

For an example of DS analysis performed on a clean voiced signal a speech segment of an utterance by a male speaker was used. Figure 3.7 shows the signal waveform, the TBS and a zoomed view on the voiced segment.

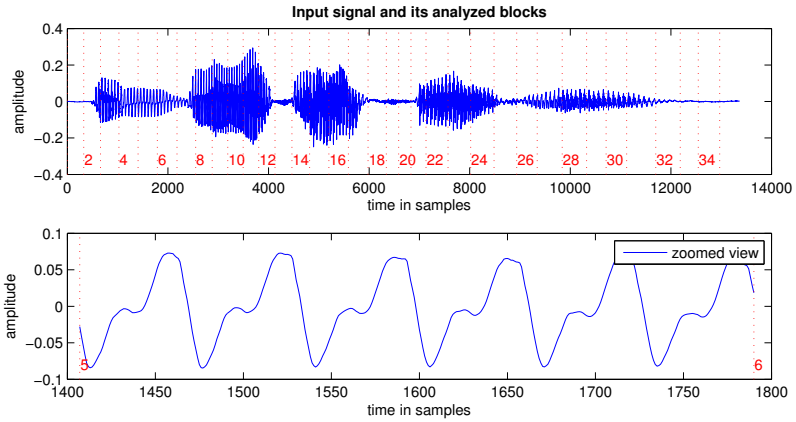


Figure 3.7: Example for TBS of a clean signal and zoomed view on a block showing a voiced segment, $f_0 \cong 123$ Hz and $\tilde{P}_0 = 65$.

In Figures 3.8 and 3.9, the corresponding pitch-synchronous transform and modulation transform are depicted, respectively.

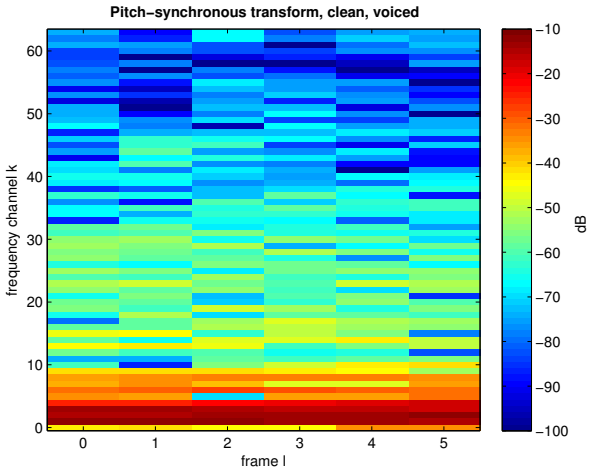


Figure 3.8: Uniform filterbank obtained by pitch-synchronous transform (DCT-IV). For a clean voiced signal segment the coefficients change slowly over time.

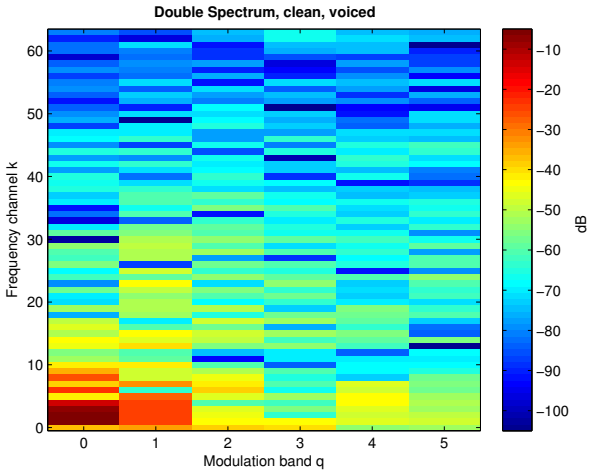


Figure 3.9: Double Spectrum obtained by modulation transform (DCT-II). For a clean voiced signal segment a strong portion of the energy is concentrated in low modulation bands.

3.3.3 Noisy Voiced Signal

To demonstrate the impact of noise on the DS representation, the signal used in the previous example of Section 3.3.2 is now corrupted by additive white Gaussian noise (AWGN) at a global input signal-to-noise ratio (SNR) of 5 dB. The TBS was carried out using the f_0 information of the clean signal, which we refer to as f_0-oracle scenario throughout the remainder of this thesis.

The reason for using f_0-oracle is to have a fair comparison between the same time block with and without the influence of noise. The TBS of the noisy voiced signal is shown in Figure 3.10

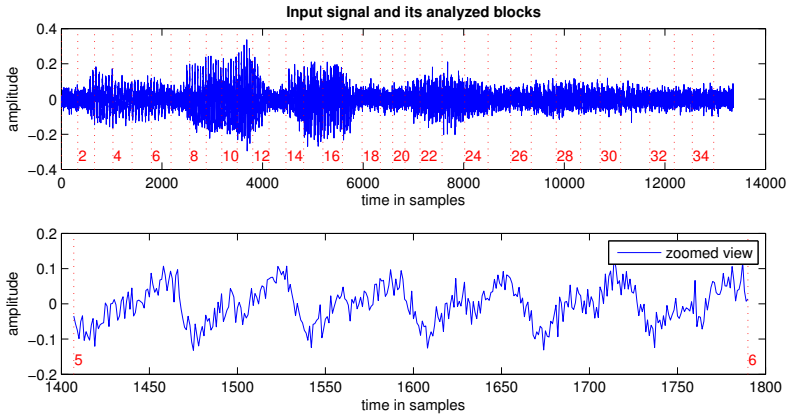


Figure 3.10: Example for TBS of a noisy signal and zoomed view on the same time block used in Section 3.3.2, $f_0 \cong 123$ Hz and $\tilde{P}_0 = 65$.

The impact of noise added to a clean signal can be observed in the DS domain, as shown in Figure 3.12. Since noise is in general of aperiodic nature, this has the effect of contributing to energy distributed over higher modulation bands, i.e. $q \neq 0$. A specific distribution over modulation bands depends on the noise type: Broadband noise such as *white* noise is approximately uniformly distributed over q, whereas some narrow band, non-stationary noise types, such as *babble* noise, may feature periodic components and thus do not necessarily show a uniform distribution across the modulation bands.

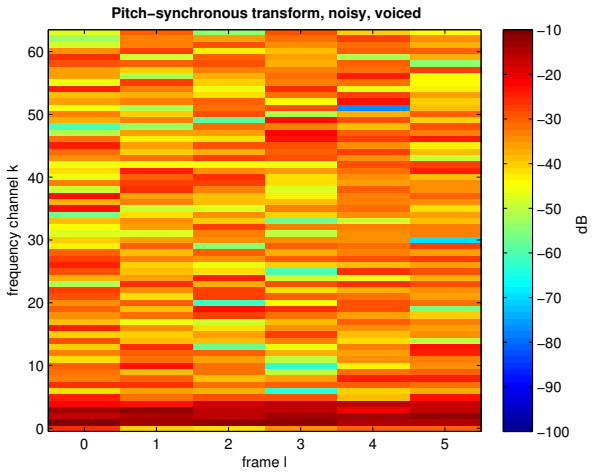


Figure 3.11: Uniform filterbank obtained by pitch-synchronous transform (DCT-IV). Due to additive noise the coefficients show stronger variation over consecutive pitch cycles than in the clean case (Figure 3.8).

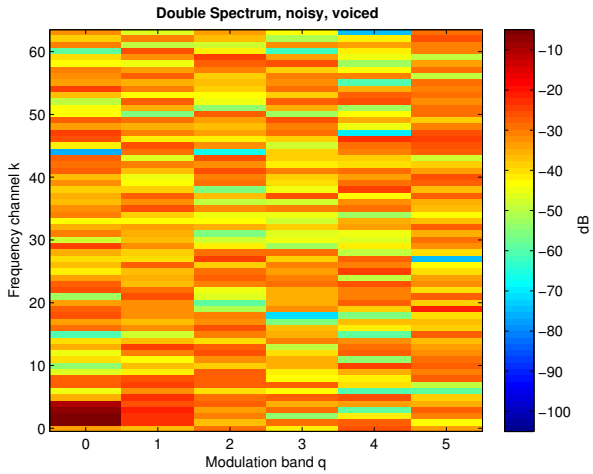


Figure 3.12: Double Spectrum obtained by modulation transform (DCT-II). Since TBS detected a periodic signal segment, the energy is mainly concentrated in low modulation bands, however we observe increased energy for $q \neq 0$.

3.3.4 Unvoiced Signal

The unvoiced components of a speech signal, e.g. fricatives such as /f/, show aperiodic behavior and are thus similar to noise in the DS domain. In Figure 3.13 the TBS including a zoomed

view on a time block representing unvoiced speech is presented. Note that since in the unvoiced case we do not deal with periodic signals, there is no unique P_0 value.

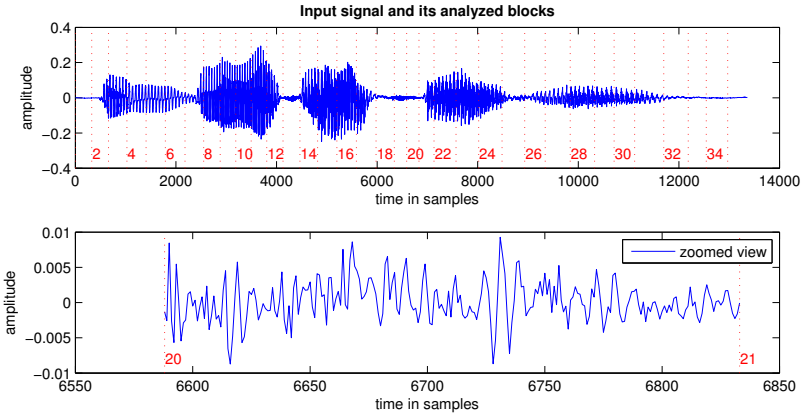


Figure 3.13: Example for TBS of a clean signal and zoomed view on a block showing an unvoiced segment

The two-stage transform of the unvoiced signal segment is shown in Figures 3.14 and 3.15. The DS representation has a noise-like structure with coefficients distributed randomly across q.

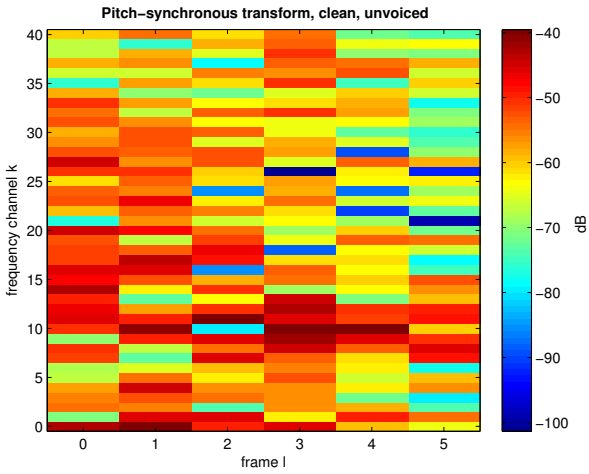


Figure 3.14: Uniform filterbank obtained by pitch-synchronous transform (DCT-IV). Unvoiced signals feature coefficients which tend to change rapidly across time frames.

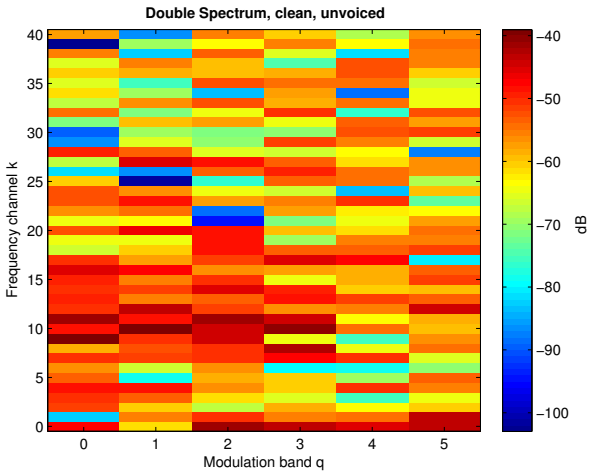


Figure 3.15: Double Spectrum obtained by modulation transform (DCT-II). In the unvoiced case there is no distinct energy concentration in the DS domain.

3.4 Useful Properties in the Double Spectrum Domain

In the previous Section, characteristic signal types were investigated in terms of their DS. In this Section, we want to take a closer look at the DS representation itself and point out its beneficial properties. In several experiments we found four properties in the DS domain, which may be exploited for speech enhancement applications: sparsity, linearity, real-valued coefficients and facilitating a harmonic filter bank.

3.4.1 Property I: Sparsity

For a periodic signal segment $DS(q,k)$ yields a high energy concentration in the low modulation bands. In particular, the first modulation band $q = 0$ represents the periodic component of a signal, whereas the other modulation bands describe the aperiodic parts. This property can be easily explained by assuming a strictly periodic time signal, e.g. a sinusoidal waveform, as discussed in Section 3.3.1. Applying the pitch-synchronous transform yields MLT coefficients which are identical for consecutive frames. The subsequent modulation transform is hence applied to a constant data sequence, yielding only one non-zero coefficient for $q = 0$, which can be understood as the DC component of the DCT-II transform. This property may be exploited for a periodic-aperiodic decomposition, which in terms of speech signals denotes a separation of voiced-unvoiced components, or for restoring the harmonicity of noise corrupted speech by finding an appropriate balance between low and high modulation bands [32–34], which may be relevant in speech enhancement applications.

3.4.2 Property II: Linearity

In the time domain, the noisy signal $y(n)$ is a superposition of the clean signal $x(n)$ and the noise signal $d(n)$. In the DS domain this superposition is preserved, since DS is a *linear* operator. A system is called linear if it has the *additivity* and the *homogeneity* property [20], both of which are valid for the DS transform. The additivity in DS is given as

$$y(n) = x(n) + d(n) \quad \circ\!\!-\!\!\bullet \quad DS_y(q,k) = DS_x(q,k) + DS_d(q,k). \tag{3.9}$$

Homogeneity is ensured by

$$s(n) = ax(n) \quad \circ\!\!-\!\!\bullet \quad DS_s(q,k) = aDS_x(q,k), \tag{3.10}$$

where $s(n)$ is $x(n)$ scaled by a constant factor $a \in \mathbb{R}$. Using equations (3.9) and (3.10) the linearity property can be expressed as

$$y(n) = a(x(n) + d(n)) \quad \circ\!\!-\!\!\bullet \quad DS_y(q,k) = aDS_x(q,k) + aDS_d(q,k). \tag{3.11}$$

Figure 3.16 shows an example for $DS_x(q,k)$, $DS_d(q,k)$ and $DS_y(q,k)$ of the same voiced speech segment to illustrate *additivity*.

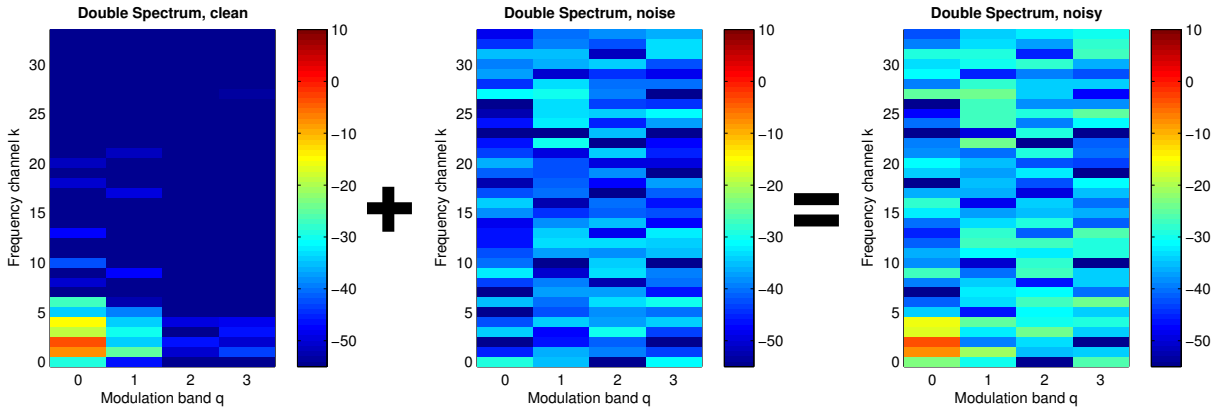


Figure 3.16: Additivity of DS as shown in (3.9)*: (Left) clean, (Middle) noise and (Right) noisy DS.*

This property may be exploited to derive a noise suppression rule similar to spectral subtraction

[24], defined in the DS domain:

$$\hat{x}(n) = y(n) - \hat{d}(n) \;\; \circ\!\!-\!\!\bullet \;\; \widehat{DS}_x(q,k) = DS_y(q,k) - \widehat{DS}_d(q,k), \tag{3.12}$$

where $\widehat{DS}_x$ and $\widehat{DS}_d$ denote clean target and noise estimate in DS, respectively, and $\hat{x}(n)$ and $\hat{d}(n)$ relate to the corresponding time-domain signals.

3.4.3 Property III: Real-Valued Coefficients

Unlike the complex-valued coefficients obtained by Fourier analysis, the coefficients of $DS(q,k)$ are real-valued. For this reason, there is no need to decompose the coefficients into their magnitude and phase components for further processing. However, since the coefficients of two-dimensional cosine transforms are symmetrically distributed around zero [37–40], a decomposition into magnitude and sign information can be performed as

$$DS(q,k) = |DS(q,k)| \; sgn(DS(q,k)). \tag{3.13}$$

Such a decomposition may be advantageous for a noise suppression rule which makes use of a magnitude estimator in DS. We speculate, that both the magnitude $|DS(q,k)|$ and the sign $sgn(DS(q,k))$ may contain information about the phase, similar to the complex exponential well-known from STFT-based signal processing.

3.4.4 Property IV: Harmonic Filter Bank

Another property of the DS representation relates to the pitch-synchronous transform. Since the analysis window length is always $2\tilde{P}_0$, the number of frequency channels is

$$K = \tilde{P}_0. \tag{3.14}$$

Note that since we have a critically sampled filter bank, channel K corresponds to the Nyquist frequency $f_s/2$, which denotes the critical frequency, that guarantees an aliasing-free signal representation following the Nyquist sampling criterion [19, 20]. Given the normalized pitch period

$$\tilde{P}_0 = \frac{f_s}{f_0}, \tag{3.15}$$

and using Equation (3.14), the frequency resolution of the uniform filter bank obtained by the pitch-synchronous transform yields

$$\Delta f = \frac{f_s/2}{K} = \frac{f_0}{2}. \tag{3.16}$$

Similarly, we have

$$\frac{f_s/2}{f_0} = \frac{K}{k_{f_0}}, \tag{3.17}$$

where k_{f_0} relates to the frequency channel corresponding to f_0. By inserting (3.14) and (3.15) we can reformulate (3.17) to find k_{f_0} as

$$k_{f_0} = 2. \tag{3.18}$$

This implies that DS shows a harmonic structure across frequency channels, with $k = 0$ as the DC-component, $k = 1$ corresponding to $f_0/2$ and $k = 2$ corresponding to f_0. This property facilitates comb filtering and may be exploited for harmonic "tunneling", where a running estimate of the noise is obtained by sampling the noise spectrum in the gaps between harmonic spectral peaks [41].

3.5 Challenges and Implementation Aspects

In this Section we present challenges and implementation issues that arise when dealing with signal processing in the DS domain.

3.5.1 Dependency on the Fundamental Frequency

Due to the pitch-synchronous framework consisting of TBS and pitch-synchronous transform, the DS representation relies on robust f_0 estimation. In a heavy noise scenario it may be difficult to determine the exact pitch values which are needed for computing the length of pitch-synchronous frames. Hence, the compactness property of DS, namely that the energy of periodic signal components is rendered into a small subspace at low modulation bands, is easily corrupted by erroneous TBS. For a better understanding, Figure 3.18 shows the impact of inaccurate f_0 estimation and thus suboptimal TBS on the example of a sinusoid, as used in Section 3.3.1. As can be seen, the energy that was once concentrated in the first modulation

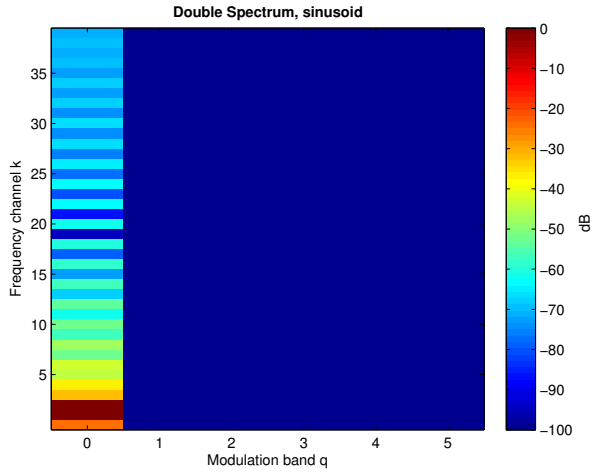


Figure 3.17: Double Spectrum of a sinusoid with correct TBS. The energy is well concentrated in the first modulation band.

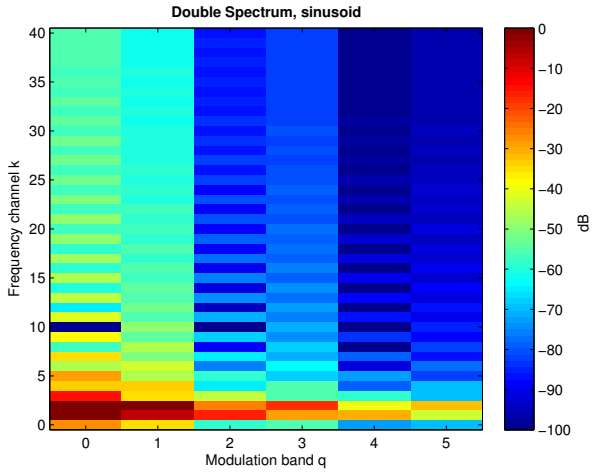


Figure 3.18: Double Spectrum of a sinusoid with incorrect TBS (error of one sample). Energy leakage across q occurs.

band gets leaked into higher modulation bands. This is actually not a big issue as long as there is no modification performed in the DS domain. However, once there is some kind of modification, e.g. a weighting scheme that balances the energy distribution in DS, relevant signal information could be lost. Therefore, robust f_0 knowledge is in particular needed in a DS-based speech enhancement method that tries to enhance the harmonicity of noise corrupted speech. In Chapter 4 we will address the problem of robust pitch estimation by deriving an f_0 estimator that operates in the DS domain, and compare its performance to other f_0 estimators.

3.5.2 Number of Modulation Bands

In [17] and [32] the number of modulation bands is determined such that the energy concentration in the first modulation band is maximized for each time block. In this way, successive blocks of voiced or unvoiced speech, respectively, are grouped into the same time block. This may provide a similar time-frequency resolution as perceived by the human auditory system [28]. In our proposed system suited for speech enhancement however, we use a "static" TBS resulting in a constant number of frames per time block and thus a fixed number of modulation bands Q per DS. To find a suitable choice of Q, we assume that speech signals can be considered to be stationary for short time intervals between 10 and 30 ms [2, 19]. Typical mean values for f_0 are reported as $\bar{f}_{0,\male} = 120$ Hz for male and $\bar{f}_{0,\female} = 210$ Hz for female speakers [42]. With this knowledge we are able to compute a reasonable number of pitch cycles to be used within a time block. Using Equation (2.1), the corresponding mean pitch periods yield

$$\bar{P}_{0,\male} = \frac{1}{120 \text{ Hz}} \cong 8.3 \text{ ms}, \tag{3.19}$$

$$\bar{P}_{0,\female} = \frac{1}{210 \text{ Hz}} \cong 4.8 \text{ ms}. \tag{3.20}$$

Assuming stationary voiced speech segments of 25 ms, we compute the number of pitch cycles $\mathcal{L}$ (frames) fitting into one segment as

$$\mathcal{L}_{\male} = \frac{25 \text{ ms}}{8.3 \text{ ms}} \approx 3, \tag{3.21}$$

$$\mathcal{L}_{\female} = \frac{25 \text{ ms}}{4.8 \text{ ms}} \approx 5. \tag{3.22}$$

Since we know that $\mathcal{L} = Q$, we go for a trade-off between the male and female case and set the number of modulation bands to $Q = 4$ for our further experiments. A recommendation for Q_{min} and Q_{max} for speech processing applications can be given assuming typical range of f_0 in human speech as $f_0 \in [80, 280]$ Hz [19]. The corresponding range of the pitch period is then $P_0 \in [3.6, 12.5]$ ms. Taking into account stationarity during 10 and 30 ms, the lower and the upper bound for the number of modulation bands is

$$Q_{min} = \frac{10 \text{ ms}}{12.5 \text{ ms}} \approx 1, \tag{3.23}$$

$$Q_{max} = \frac{30 \text{ ms}}{3.6 \text{ ms}} \approx 8. \tag{3.24}$$

Using $Q_{min} = 1$ would not be a reasonable choice because in this case the possibility of an energy weighting among modulation bands is not given. In fact, using only one modulation band is equivalent to using only the pitch-synchronous transform to obtain a uniform filter bank. This is essentially a spectral analysis based on pitch-synchronous frames.

3.5.3 Modulation Frequency of a Modulation Band

In the previous Section we discussed the number of modulation bands. We now want to investigate how to determine the corresponding modulation frequency f_q of a certain modulation band q. Similar to the pitch-synchronous transform, the modulation transform yields a uniform filter bank resulting in discrete modulation bands, which describe how the MLT coefficients evolve over time. Due to an overlap of 50% between frame ℓ and $\ell+1$ according to [17], we obtain a set of MLT coefficients every $\tilde{P}_0$ samples. This inherently generates a sparse signal vector that only contains information at integer multiples of $\tilde{P}_0$. Taking into account the sampling period of the time domain signal $T = 1/f_s$, we can now define the sampling period T_q of the MLT coefficients

as

$$T_q = \tilde{P}_0 T. \tag{3.25}$$

The corresponding sampling frequency of the modulation transform is

$$f_{s,q} = \frac{f_s}{\tilde{P}_0}. \tag{3.26}$$

Using the definition of $\tilde{P}_0 = f_s/f_0$ we get

$$f_{s,q} = f_0. \tag{3.27}$$

This means that the pitch defines the sampling frequency of the MLT coefficient series and hence determines the modulation frequency f_q of a respective modulation band. In this case, the Nyquist frequency for critical sampling is $f_0/2$ corresponding to the highest modulation frequency resolvable without aliasing. Considering a critically uniform filter bank of Q modulation bands obtained by the DCT-II, the model order Q determines the modulation frequency resolution

$$\Delta q = \frac{f_0/2}{Q}. \tag{3.28}$$

We can now express the modulation frequency f_q of a certain modulation band q as

$$f_q = q\,\Delta q. \tag{3.29}$$

Note that for every DS the modulation bands may represent different modulation frequencies, since f_q depends on f_0. This in turn means that f_q depends on $\tilde{P}_0$ and hence on the number of frequency channels K. It follows that for every DS there is a trade-off between acoustic frequency and modulation frequency resolution, similar to the well-known time-frequency uncertainty principle $\Delta f \Delta t = 1$ [43].

As an example for computing modulation frequencies, we consider the DS of a voiced signal segment of a female speaker shown in 3.16, where we have $\tilde{P}_0 = 34$ corresponding to $f_0 = 235$ Hz, and $Q = 4$ modulation bands. We already know that $q = 0$ stands for the DC component, i.e. $f_q = 0$ Hz. The modulation frequency resolution yields

$$\Delta q = \frac{235 \text{ Hz}}{2 \cdot 4} = 29.375 \text{ Hz}. \tag{3.30}$$

Thus, for $q = 1$ we have a modulation frequency of $f_q = 29.375$ Hz. If we consider a male speaker with $f_0 = 120$ Hz we get $f_q = 15$ Hz. From the findings of several modulation studies summarized in Section 2.2, we know that modulation frequencies up to 16 Hz are important for the intelligibility of speech. As a consequence, we regard the first two modulation bands as important for speech enhancement, assuming $Q = 4$ following Section 3.5.2.

3.5.4 Window Overlap of Time Blocks

It is known that for perfect reconstruction, the sum of the analysis windows shifted by the increments defined in $\mathcal{Z}$ (Equation (2.11)) needs to add up to a constant value [3]:

$$C = \sum_{l=-\infty}^{\infty} w(n - l\mathcal{Z}). \tag{3.31}$$

In the DS framework the overlap within a time block, i.e. the overlap between frames ℓ and $\ell+1$ is set to 50% according to [17]. In this case the constraint given in Equation (3.31) is met, since each frame has a length of $2\tilde{P}_0$ with an overlap of $\tilde{P}_0$. For consecutive time blocks however, the overlap should not necessarily be 50%, given that $\tilde{P}_0$ of the old and the new time block may not be equal. Therefore, discontinuities in the envelope of the summed window amplitude may occur, which results in artifacts in the reconstructed signal. An example for such an overlap mismatch is depicted in Figure 3.19. To overcome this problem, we propose an optimization

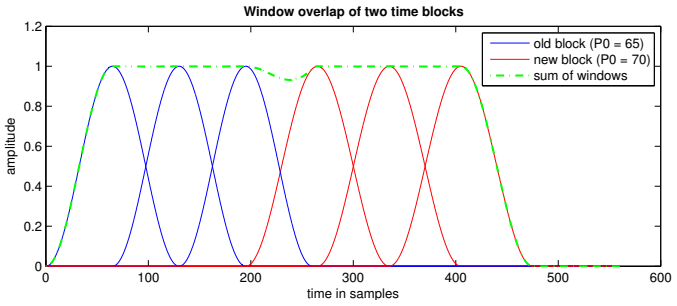


Figure 3.19: Overlap mismatch occuring due to time blocks with frames of different length. The dash-dotted green line represents the summed amplitude of overlapped windows.

criterion, which minimizes the impact of the overlap mismatch and thus reduces introduced artifacts. Within a time block the values of summed windows S_w at the intersections of two frames are constant, since the overlap is $\tilde{P}_0$. At time block transitions the sum of two windows S_t is different, as can be seen in Figure 3.19. The goal is to find the time shift τ between the old and the new time block that minimizes the error $e(\tau)$ given by the distance between S_w and S_t in a least squares sense:

$$\tau_{opt} = \underset{\tau}{\operatorname{argmin}} \, |e(\tau)|^2 = \underset{\tau}{\operatorname{argmin}} \, |S_w - S_t(\tau)|^2. \tag{3.32}$$

This problem yields a simple closed form solution:

$$\tau_{opt} = \frac{\tilde{P}_{0,\text{old}} - \tilde{P}_{0,\text{new}}}{2}. \tag{3.33}$$

The optimal overlap is then computed as

$$OVLP_{opt} = \tilde{P}_{0,\text{old}} - \tau_{opt} = \frac{\tilde{P}_{0,\text{old}} + \tilde{P}_{0,\text{new}}}{2}. \tag{3.34}$$

We see that taking the mean value of the old and the new $\tilde{P}_0$ resolves the problem. The optimization of the overlap mismatch shown in Figure 3.19 is illustrated in Figure 3.20.

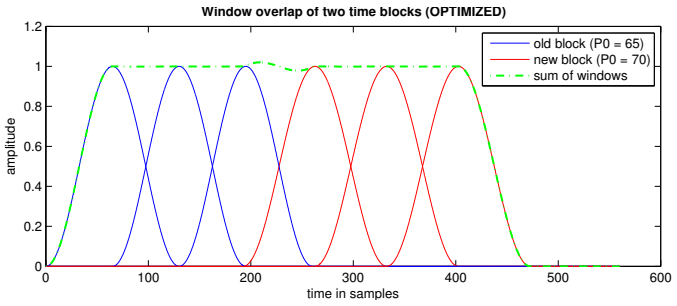


Figure 3.20: Optimized overlap between two time blocks with different frame lengths. The summed window amplitude still shows a discontinuity, but resulting in less artifacts in the DS synthesis stage.

From an implementation point of view this optimization procedure is computationally costly. In the first step, the normalized pitch period of the old time block $\tilde{P}_{0,\text{old}}$ is used as overlap, which gives the non-ideal starting point for the new time block. The instantaneous pitch period of the new time block $\tilde{P}_{0,\text{new}}$ is then estimated at this point and may differ from $\tilde{P}_{0,\text{old}}$, resulting in an overlap mismatch. By applying the optimization we find a better starting point of the new time block, which is optimal in terms of the least squares criterion given in Equation (3.32). Once this starting point is found, a new estimation of the instantaneous pitch is needed to find the most accurate value for $\tilde{P}_0$.

This overlap enhancement procedure is in particular beneficial when used in combination with a time block overlap of $\tilde{P}_0$. Since time blocks are generally longer than 50 ms, the pitch is likely to change after this period of time. Perceptually, we observed slight improvement for the analysis-synthesis of clean signals using this modification. Artifacts at time block transitions of the reconstructed signal could be effectively reduced to a minimum. In the presence of noise added to the signal, however, there is almost no perceptual or measurable improvement. For the sake of simplicity and speed this overlap optimization procedure will not be used in our proposed speech enhancement framework.

4

Pitch and Speech Presence Probability Estimation

Pitch is a key feature in many speech processing applications. For this reason, the estimation of the fundamental frequency f_0 is an essential component in many systems. For example in [17], we arrive at a sparse signal representation by using pitch-synchronous frames for the two-stage transform procedure. However, the effectiveness of this representation strongly depends on the estimated pitch, as already pointed out in Section 3.5.1. In the following Section, we present a method to estimate pitch in the DS domain and compare its performance to other f_0 estimators, which are robust to the impact of noise.

4.1 Pitch Estimation using Double Spectrum

Typically, pitch estimation algorithms either use a stochastic model for certain parameters of a noisy speech signal (*parametric*) or exploit the harmonic structure in frequency domain, the periodicity in time domain or a mixture of both (*non-parametric*) [44]. From Section 3.4.1 we already know, that information about periodicity is stored within the low modulation bands of the DS representation. In particular, the energy of the first modulation band, i.e. $q = 0$, relates to the strictly periodic component of the analyzed signal segment. This fact leads to the idea of defining a measure for periodicity in the DS domain. Similar to [32,33], we introduce a measure that compares the summed energy of the first modulation band E_1 to the total energy $E_{1:Q}$, called the Modulation Band Ratio (MBR). We define the MBR as

$$\mathrm{MBR}(K) = \frac{E_1}{E_{1:Q}} = \frac{E_1}{E_1 + E_{2:Q}} \ , \tag{4.1}$$

where

$$E_1 = \sum_{k=0}^{K-1} |DS(0,k)|^2, \tag{4.2}$$

$$E_{1:Q} = \sum_{q=0}^{Q-1} \sum_{k=0}^{K-1} |DS(q,k)|^2, \tag{4.3}$$

and $E_{2:Q}$ is the summed energy from the second to the Q-th modulation band. For periodic frames (voiced speech) the MBR reaches values close to 1:

$$\mathrm{MBR}_V(K) = \frac{E_1}{E_1 + E_{2:Q}} \approx \frac{E_1}{E_1} = 1, \tag{4.4}$$

while for non-periodic frames (unvoiced speech) the mean MBR is close to zero. Assuming approximately uniformly distributed energy across q we get

$$\mathrm{MBR}_{UV}(K) = \frac{E_1}{E_1 + E_{2:Q}} \approx \frac{E_1}{Q \cdot E_1} = \frac{1}{Q} \,. \tag{4.5}$$

This leads us to the idea of a DS-based f_0 estimator: Given a speech signal and an arbitrary starting point of a time block, we argue that there has to be a frame length $\tilde{P}_0$, which maximizes the energy in the first modulation band of the corresponding DS. Figure 4.1 illustrates the underlying principle of the proposed f_0 estimator in time domain. Since the DS itself needs $\tilde{P}_0$,

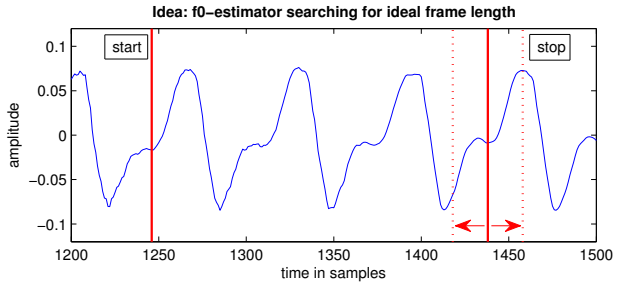


Figure 4.1: Principle for an f_0 estimator searching for the optimal frame length.

this poses a chicken-and-egg problem. To resolve this issue, we apply an iterative search to find the optimal value for $\tilde{P}_0$, which in the DS domain relates to the number of frequency channels K. The optimal frequency index K^* that maximizes the MBR is found by

$$K^* = \underset{K}{\operatorname{argmax}}\ \mathrm{MBR}(K). \tag{4.6}$$

Using the relation between f_0 and frequency channels k given in Equation (3.17), the fundamental frequency estimate is

$$f_0^* = \frac{f_s}{K^*}. \tag{4.7}$$

4.1.1 Implementation of the Double Spectrum Pitch Estimator

The implementation of an f_0 estimator can either be offline, i.e. the signal which is to be analyzed in terms of its pitch trajectory is available as a whole, or online, which means the signal is processed block by block in a serial fashion. For offline algorithms, it is typical to apply some sort of pre- or post-processing to the signal, e.g. filters, spectral normalization [44] or temporal smoothing. In the offline version of our proposed f_0 estimator, which we refer to as Double Spectrum Pitch Estimator (DS-f_0), we use no preprocessing, however we apply a simple first order lowpass filter to smooth the output trajectory of f_0 values. The algorithm uses block processing, with a step size of 5 ms for analysis blocks. The block length is determined by the number of modulation bands, which is set to $Q = 4$ following our reasoning from Section 3.5.2. The search room for f_0 values is chosen as $f_0 \in [80, 320]$ Hz similar to the typical f_0 range in human speech [19]. The lower and the upper bound of the corresponding $\tilde{P}_0$ search room $[\tilde{P}_{0,min}, \tilde{P}_{0,max}]$ are defined as

$$\tilde{P}_{0,min} = \left\lfloor \frac{f_s}{f_{0,max}} \right\rceil, \tag{4.8}$$

$$\tilde{P}_{0,max} = \left\lfloor \frac{f_s}{f_{0,min}} \right\rceil, \tag{4.9}$$

where $\lfloor . \rceil$ denotes the nearest integer rounding operator. For $f_s = 8$ kHz, $f_0 \in [80, 320]$ relates to a search range of $\tilde{P}_0 \in [25, 100]$ samples for the optimal frame length. This is equivalent to finding the number of $K \in [25, 100]$ frequency channels, that yields the highest MBR value.

To demonstrate the potential of DS-f_0, we compare our method versus two benchmark algorithms operating offline as well: PEFAC, a pitch estimation algorithm robust to high levels of noise [44] and IPF, an algorithm for the estimation of the instantaneous pitch frequency of speech using time warping and a B-spline expansion [30]. In our experiment we chose sample sentences of a male and a female speaker taken from GRID corpus [45], and observed the pitch trajectories over time for each algorithm. Since this experiment should serve as a proof of concept for DS-f_0, we skipped further evaluation measures. In the absence of the true pitch, we use PEFAC applied on the clean signal as a reference for correct pitch (oracle). Note that all trajectories are time aligned using cubic interpolation to match their lengths in the plot. Figures 4.2 and 4.3 show the f_0 trajectories of the male speech signal for clean and noisy (white, 0 dB) scenarios, respectively. It can be observed that the red line representing DS-f_0 follows the oracle trajectory

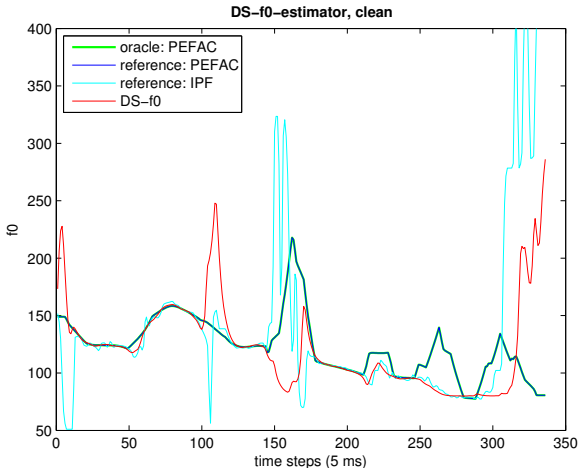


Figure 4.2: Pitch estimation performed on a clean sentence of a male speaker.

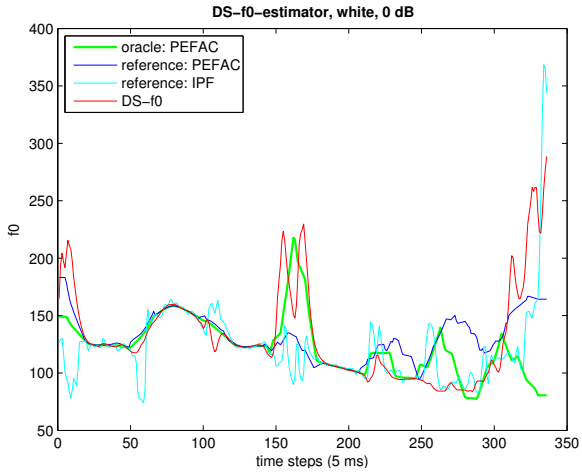


Figure 4.3: Pitch estimation performed on a noisy signal (female, white, SNR = 0 dB).

well. During regions of unvoiced speech f_0 estimators tend to show fluctuating behavior, since an unvoiced signal does not contain periodic information. The results for the clean and noisy signal of the female speaker are presented in Figures 4.4 and 4.5. Pitch estimators are typically

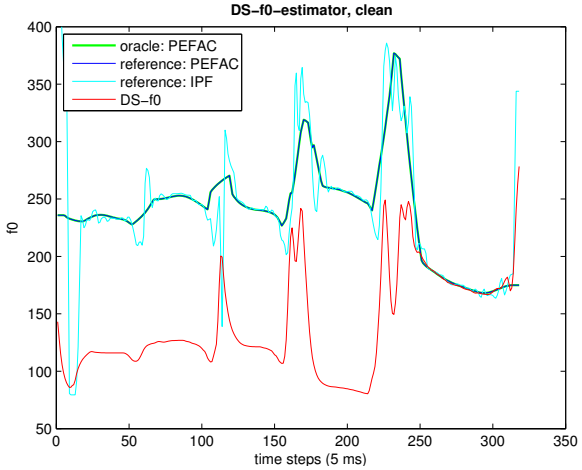


Figure 4.4: Pitch estimation performed on a clean sentence of a female speaker.

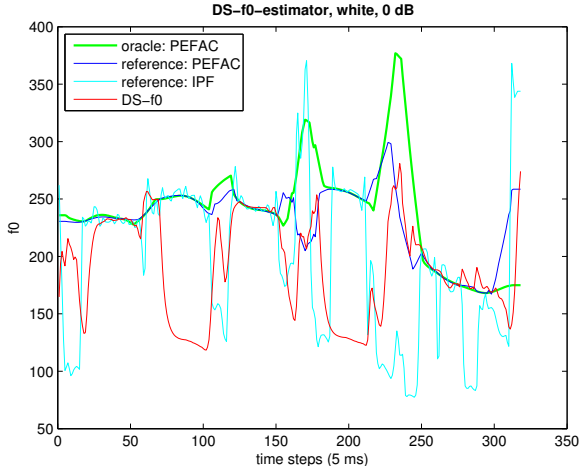


Figure 4.5: Pitch estimation performed on a noisy signal (female, white, SNR = 0 dB).

prone to octave errors, which means the erroneous detection of f_0 values that are multiplied or divided by two (or any simple rational number) [44]. As can be seen in Figures 4.4 and 4.5, DS-f_0 introduces a halving problem, i.e. the subharmonic was detected as pitch. The reason for this may be that the DS representation yields higher MBR values if a lower f_0 is chosen. In other words, if the subharmonic of f_0 is detected, there is one more harmonic in the first modulation band containing high energy: the harmonic corresponding to the actual pitch. To overcome the halving problem, many pitch estimation algorithms apply temporal continuity constraints to ensure a smooth f_0 trajectory or identify multiple pitch candidates to solve the problem using a stochastic framework. In DS-f_0, neither of these two methods is employed, which leaves room for improvement. If we modify the f_0 search and set the lower bound $f_{0,min} = 150$ Hz, the pitch track is corrected and the performance of our pitch estimator is satisfying. The corrected f_0 track using a search room of $f_0 \in [150, 320]$ Hz is presented in Figure 4.6.

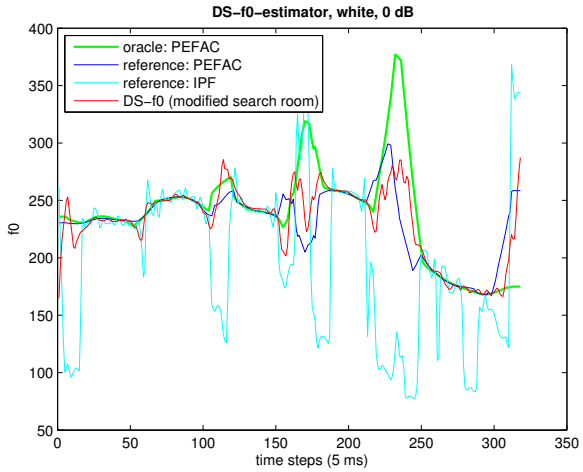


Figure 4.6: Pitch estimation performed on a noisy signal (female, white, SNR = 0 dB). An f_0 search room of $[150, 320]$ Hz resolves the halving problem seen in Figures 4.4 and 4.5.

4.1.2 Using Pitch Estimators for Time Block Segmentation

The TBS needed for DS analysis requires the estimated pitch track to compute the normalized pitch period $\tilde{P}_0$. Finding an accurate match of the true $\tilde{P}_0$ is a difficult task: On the one hand, the impact of noise on the signal will degrade the accuracy of the estimated pitch. On the other hand, instantaneous f_0 values are computed at discrete time instances, e.g. in steps of 5 ms, which requires interpolation to align the pitch trajectory with the speech signal. In addition, computing $\tilde{P}_0$ from f_0 introduces rounding to the nearest integer value, since we operate in the discrete-time domain. Both of the aforementioned issues limit the precision of TBS and result in leakage of energy in DS. As a consequence, a speech enhancement algorithm based on DS will suffer from speech distortion once the compactness gets affected. To address these issues and the dependency on offline f_0 estimation algorithms, such as PEFAC, we propose an online implementation of DS-f_0, that iteratively seeks the optimal frame length $\tilde{P}_0$ in a forward fashion. Similar to the principle shown in Figure 4.1, this online version of DS-f_0 needs a starting index and searches for the frame length, i.e. $\tilde{P}_0$, that maximizes the MBR, as described in the previous Section. This procedure has the advantage of finding the time block with the highest energy concentration in the first modulation band, i.e. the most compact representation achievable in terms of MBR. We argue that even though our DS-f_0 method may not produce the most accurate outcome for pitch tracking, it is optimal in a sense to find the time blocks that contain the most information about periodicity in DS. This may be advantageous considering a speech enhancement method which relies on improving periodicity, similar to [33]. A comparison of

TBS relying on estimated pitch by PEFAC and TBS using the DS-f_0 forward method will be conducted in the following Section and later on in Section 6.2.1 in terms of the performance of our speech enhancement method.

4.2 Speech Presence Probability Estimation

Taking into account that a speech signal is a sequence of words and pauses, many common speech enhancement systems use information about the speech presence probability (SPP) to improve the performance of speech target or noise estimates. The SPP is estimated from the noisy observation and can be used to compute the speech presence uncertainty (SPU) and to define a binary hypothesis test, where $\mathcal{H}_1(\ell)$ indicates that speech is present at time frame ℓ, and $\mathcal{H}_0(\ell)$ indicates speech absence [1]. An early example for a method which used the SPU was the MMSE Short-Time Spectral Amplitude (STSA) Estimator Under Signal Presence Uncertainty [46]. In the design of our filter method we take into account the SPP to selectively modify regions of speech presence or absence.

In Section 4.1 we defined the MBR as a measure for periodicity and used it for f_0 estimation. When dealing with speech signals, the MBR can also be used as a degree of voicing based on the energy concentration in the low modulation bands [32, 33]. As already shown in Section 3.3 for voiced speech, the energy is mostly concentrated in the first modulation band $q = 0$, while for unvoiced speech this is not true. The MBR is therefore capable of discriminating voiced and unvoiced speech. In theory, MBR yields values close to 1 for voiced and close to 0 for unvoiced frames, hence is a good measure for SPP itself. In this Section, we perform a detailed analysis of the energy concentration in terms of histograms to justify the use of MBR as a means of estimating the SPP in the DS domain.

4.2.1 Energy Concentration Analysis in Double Spectrum using Histograms

For our experiment we extracted DS data from 30 phonetically-balanced sentences taken from NOIZEUS speech corpus [47] and selectively labeled Double Spectra as voiced or unvoiced. The speech files were downsampled from the original sampling rate of 25 kHz to 8 kHz to simulate telephony speech. The number of extracted Double Spectra $\mathcal{N}$ was in the range of 4000 to 7000 (depending on the parameter setup) for the voiced and the unvoiced class each. The number of modulation bands in every DS was set to $Q = 4$. In the absence of true voiced-unvoiced (VUV) labeling, e.g. done by manual annotation, we used the voicing probability output of the PEFAC algorithm [44] as a decision rule to create the voiced and the unvoiced dataset. Since the SPP provided by PEFAC does not represent the ground truth of VUV classification, the results obtained should be viewed with skepticism. For the sake of a reliable classification, a DS was labeled *voiced* if the voicing probability output yielded values $p_v > 0.9$, and *unvoiced* for $p_v < 0.1$. For the TBS two different setups were used:

a) TBS carried out by computing $\tilde{P}_0$ from f_0 values estimated by PEFAC,

b) TBS by iterative search of $\tilde{P}_0$ using the DS-f_0 forward algorithm.

After the voiced and unvoiced data was extracted and accordingly labeled, we computed the MBR of every DS and performed a statistical evaluation in terms of histograms. Since in general the number of voiced and unvoiced data samples is different, we use the normalized histogram, i.e. the histogram divided by the number of data samples, in order to compare the voiced and the unvoiced class in one single plot. Additionally, the y-axis of the histograms is scaled by a factor 100 to view the percentage of samples in every bin. The number of bins B is set so as to approximate the rule of thumb $B \approx \sqrt{\mathcal{N}}$ for voiced and unvoiced data, respectively. Figure 4.7

shows the normalized histograms of the MBR for voiced and unvoiced DS data obtained from clean speech signals using setup a) for TBS (PEFAC). In Figure 4.8 the histograms are plotted in a noisy scenario with DS obtained from speech signals corrupted by AWGN with an SNR of 5 dB. In the clean case, we observe that for the unvoiced class MBR values are concentrated

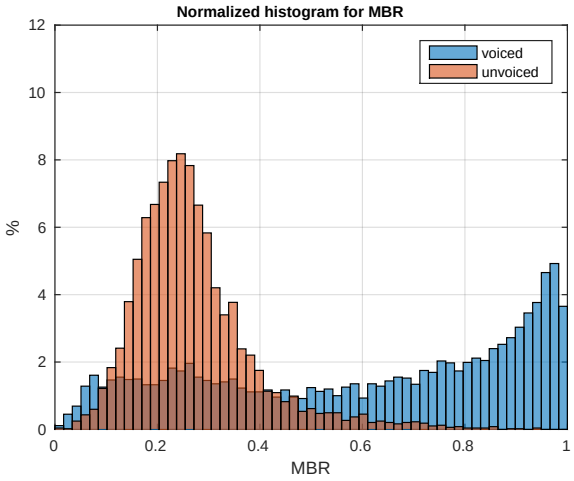


Figure 4.7: Histogram of MBR for voiced and unvoiced clean DS data, TBS based on f_0 estimation by PEFAC.

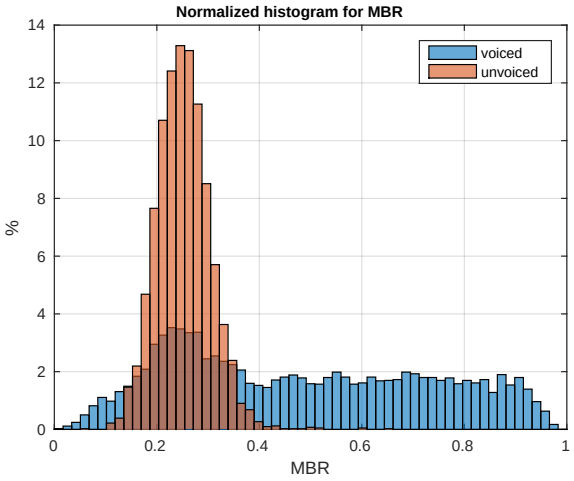


Figure 4.8: Histogram of MBR for voiced and unvoiced DS data, white noise, 5 dB, TBS based on f_0 estimation by PEFAC.

around 0.25 which matches our expectation following Equation (4.5) that $MBR_{UV} \approx 1/Q$. The fact that for voiced data the MBR is not close to 1 is an indicator that either too many DS contributing to the unvoiced class were selected by using PEFAC's voicing probability for classification or that time blocks were segmented erroneously, which lead to leakage of energy in DS. For the noisy scenario one can see that $MBR_V \approx 1$ given in Equation (4.4) no longer holds. However, it seems that due to the impact of white noise, the unvoiced class is even stronger concentrated around an MBR of 0.25. As a measure of classification performance we computed the error probability P_e. Let R be a realization of MBR and $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the hypothesis that a unvoiced and voiced signal segment was present, respectively. The error probability P_e is then defined as

$$P_e = P_0 P_{e,0} + P_1 P_{e,1}, \tag{4.10}$$

where $P_0 = P(\mathcal{H}_0)$, $P_1 = P(\mathcal{H}_1)$ are the prior probabilities of unvoiced speech and voiced speech, respectively, and

$$P_{e,0} = P(R > \mathcal{B}|\mathcal{H}_0), \tag{4.11}$$
$$P_{e,1} = P(R < \mathcal{B}|\mathcal{H}_1), \tag{4.12}$$

with $\mathcal{B}$ as a decision boundary set between the two classes [48]. In the literature a priori probabilities for speech presence or voicing typically range between 0.5 and 0.8 [2]. In our experiment, we select the prior probability for voiced speech as $P_1 = 0.6$. In both clean and noisy case, the optimal decision boundary is $\mathcal{B} = 0.35$, as can be observed in Figures 4.7 and 4.8. The resulting error probabilities for VUV classification using MBR are $P_e = 0.24$ and $P_e = 0.26$ for clean and noisy, respectively.

We now repeat the histogram experiment with TBS setup b) by iterative search of $\tilde{P}_0$ using the DS-f_0 forward algorithm. The histograms of MBR values for voiced and unvoiced DS data are shown in Figures 4.9 and 4.10 for both clean and white noise scenario at 5 dB, respectively. Compared to histograms created by setup a) using PEFAC, the results obtained by setup b)

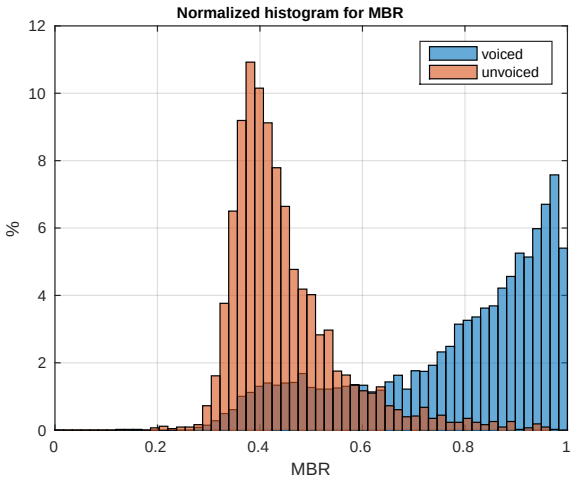


Figure 4.9: Histogram of MBR for voiced and unvoiced clean DS data, TBS based on f_0 estimation by DS-f_0.

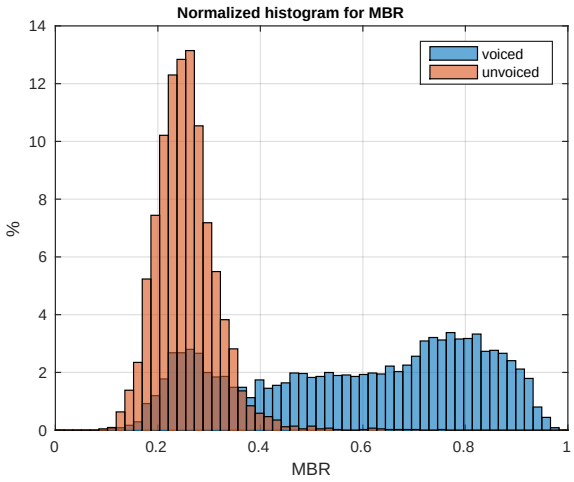


Figure 4.10: Histogram of MBR for voiced and unvoiced DS data, white noise, 5 dB, TBS based on f_0 estimation by DS-f_0.

using DS-f_0 for TBS show a better energy concentration of the voiced class. For clean DS data, as presented in Figure 4.9, we achieve a better VUV separation, however MBR values of the unvoiced class are not concentrated around $1/Q = 0.25$, but at approx. 0.4. This can be explained by the fact that DS-f_0 optimizes the length of each time block in order to achieve the highest MBR value possible. For this reason, clean time blocks were automatically segmented in a way to preserve periodic components of regions labeled as unvoiced by p_v computed by PEFAC. In the white noise scenario (Figure 4.10), MBR values of unvoiced data are again concentrated at 0.25, because DS-f_0 could not extract as much periodic information from data labeled unvoiced due to superimposed noise. In contrast to Figure 4.8, a greater portion of MBR values belonging to the voiced class is concentrated near 1. The error probability for setup b) is $P_e = 0.16$ for both clean and noisy scenario. This concludes that using DS-f_0 for TBS is more effective than relying on f_0 estimated by PEFAC in terms of maintaining high energy concentration.

Similar experiments were carried out by Huang et. al (2011) who used the proposed framework of Nilsson (2007), described in Section 2.3.3 and in Figure 2.5. In their work they quantitatively analyzed the energy concentration of VUV signal segments using three different measures. One of which was called the *percentage energy* of the first modulation band, which is defined in the same way as our proposed MBR. It was confirmed that energy concentration trends of the VUV classes are highly discriminated even in heavy noise scenarios by using periodicity measures based on the energy of the first modulation band in the two-stage transform domain [33].

4.2.2 Modulation Band Ratio as a Speech Presence Probability Measure

We have seen that by observing the energy distribution in DS it is possible to identify voiced and unvoiced sounds using MBR as a soft decision measure. In a speech enhancement scenario this property can be exploited in order to restore the harmonic structure of speech at voiced frames and to suppress noise during unvoiced regions. We now want to investigate the performance of MBR as measure for SPP. For this purpose, we run an experiment where we monitor the MBR computed from consecutive time blocks over time. The resulting MBR trajectory is then compared to the voicing probability output p_v of PEFAC [44].

Similar to the experiment on f_0 estimation of Section 4.1.1, we chose the same female speaker as a test signal for SPP estimation. In our implementation we empirically modified the MBR to only use frequency channels up to the 6-th harmonic, i.e. $k \in [0, 12]$, since this yielded more

robust results for SPP estimation in a noisy environment. We write the modified MBR as

$$\mathrm{MBR}_{6th} = \frac{\sum_{k=0}^{12} |DS(0,k)|^2}{\sum_{q=0}^{Q-1} \sum_{k=0}^{12} |DS(q,k)|^2} \ . \tag{4.13}$$

The modified MBR is computed from the DS of each time block. The TBS created time blocks with an overlap of $(Q-1)\tilde{P}_0$, which corresponds to a step size of $\tilde{P}_0$. In the following we call the SPP estimate computed from MBR_{6th} Double Spectrum Speech Presence Probability Estimator (DS-SPP). The resulting SPP estimates obtained from DS-SPP and PEFAC are aligned to the speech signal by cubic interpolation. Note that for the sake of a convenient visual comparison, the clean signal is plotted over time in all scenarios (clean and noisy). Figure 4.11 shows the SPP estimation results for clean speech. Similar to the experimental histogram results shown in

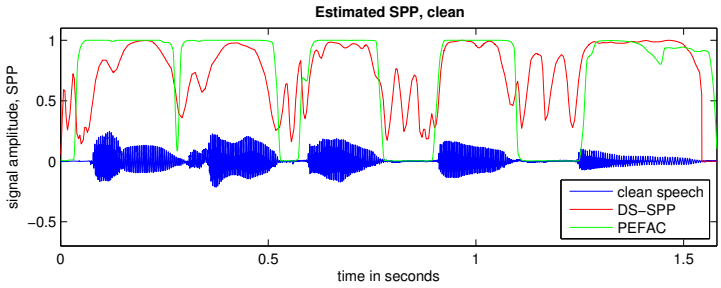


Figure 4.11: Estimated SPP trajectories for a clean signal. DS-SPP (red) shows unwanted fluctuations during regions of speech absence. The benchmark p_v by PEFAC (green) works well.

Figure 4.9, the DS-f_0 causes the MBR to reach values up to 0.8 even during speech absence and unvoiced regions when analyzing the clean speech signal. We now want to take a look at the SPP estimates under the influence of white and babble noise as an example for non-stationary noise. For babble noise, we used the audio file provided in the NOISEX-92 database [49]. Figures 4.12 and 4.13 present the SPP trajectories for white and babble noise at 5 dB SNR, respectively.

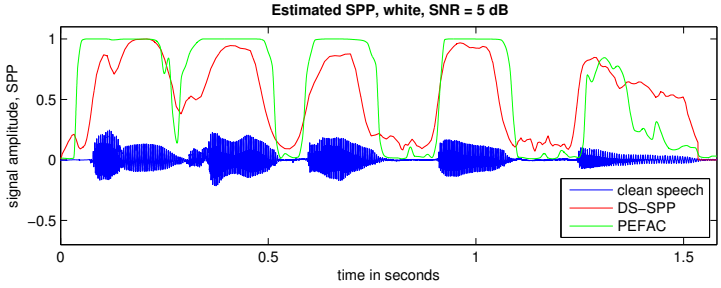


Figure 4.12: Estimated SPP trajectories for a noisy signal, white, SNR = 5 dB. DS-SPP performs similar to PEFAC and shows desired behavior during regions of speech absence.

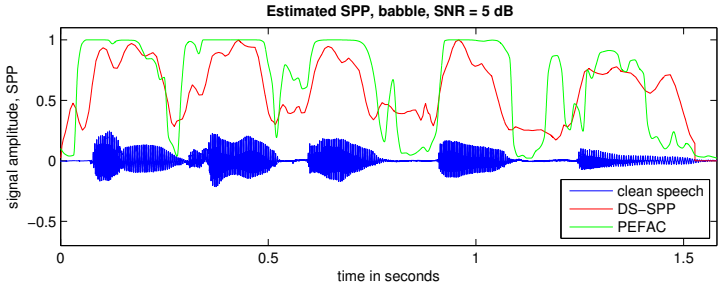


Figure 4.13: Estimated SPP trajectories for a noisy signal, babble, SNR = 5 dB. In contrast to PEFAC, DS-SPP does not show fluctuations in speech absence and thus works more robustly.

It can be observed that DS-SPP works robust in the noisy scenario. Although not reaching values near 0, the MBR tends to be a good measure for estimating SPP even in presence of non-stationary noise (Figure 4.13). In contrast to p_v of PEFAC, DS-SPP does not yield estimates which tend to binarize between values of 0 and 1, but provides smoother results over time.

The experimental results presented in this Section should serve as a proof of concept for a DS-based SPP estimation method. The design of a robust SPP estimator would require a more sophisticated approach on how to incorporate the MBR effectively. For example, the errors occurring during unvoiced regions of the clean signal (Figure 4.11) indicate that relying solely on a periodicity measure for SPP estimation is not satisfactory. However, for our experiments the MBR suffices as an approximation of the instantaneous SPP. Thus, we incorporate the MBR as an SPP estimate computed in the DS domain in our proposed DS-based speech enhancement system.

5

Speech Enhancement using Double Spectrum

Speech enhancement algorithms are designed to improve perceptual aspects of speech signals which have been corrupted by noise and acoustic disturbances. Typically, the main focus lies on improving the perceived quality and the intelligibility of degraded speech. This is achieved by reducing the impact of noise on the speech signal, which in the literature is often referred to as noise suppression. Noise can be of various nature, since it can differ in temporal and spectral characteristics, as well as in the noise level. We encounter noise in everyday life, for example in the street (traffic noise, construction work), in the car (engine noise), at work (busy office environment), at restaurants (people talking), in other public places and maybe also at home. Depending on the noise we can differentiate between stationary, i.e. temporally slowly changing, and non-stationary, i.e. rapidly varying, noise types. Obviously, the latter poses a bigger challenge in noise reduction applications than suppressing stationary noise. Examples for quasi-stationary and non-stationary noise are car and babble noise, respectively [19].

In single-channel speech enhancement we deal with systems consisting of only one microphone channel. Therefore, we consider an additive signal model as

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

where $y(n)$ is the observed noisy signal comprised of the unknown clean signal $x(n)$ and the additive noise $d(n)$, with n denoting the discrete-time index. Most commonly, these signals are considered as realizations of stochastic processes assuming the clean target and the noise process as statistically independent. The problem formulation of conventional speech enhancement methods is as follows: Given a realization $y(n)$ of the noisy process, find an estimate of the clean target realization $\hat{x}(n)$. What separates different speech enhancement methods from each other is

- the domain in which the clean target is estimated (e.g. time, frequency, subspace etc.),
- the prior assumptions made about the clean target and the noise process,
- the way the quality of the estimate $\hat{x}(n)$ is assessed [1].

In the literature many speech enhancement algorithms can be found which use an analysis-modification-synthesis (AMS) framework based on short-time processing in a specific domain, e.g. in the STFT domain [3]. In more detail, the modification step includes the detection of certain features helpful for a noise suppression task (e.g. SPP), and the estimation of noise and target signal. Figure 5.1 illustrates an AMS framework designed for speech enhancement.

In general, speech enhancement methods reported in the past are categorized as follows [3, 19]:

a) Spectral subtraction [24],

b) Wiener filtering [50],

c) Statistical-model-based methods (e.g. MMSE [46]),

d) Subspace algorithms [51].

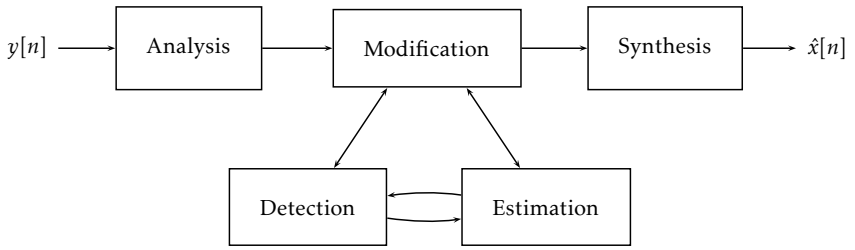


Figure 5.1: Extended AMS framework used in conventional speech enhancement applications.

In this thesis we investigate *three* approaches for speech enhancement in the DS domain. The noise suppression rules are either based on spectral subtraction or derived from the Wiener solution. In the following Sections we present the motivation and the derivation of each algorithm.

5.1 Periodicity Enhancement using Coefficient Weighting

Speech periodicity enhancement (PE) by coefficient weighting in the two-stage transform domain was proposed in works of Huang et. al [32–34]. In [33] a *fixed* and an *adaptive* weighting scheme were presented. Both methods showed promising results in terms of emphasizing the periodic component of voiced speech. However, only the adaptive scheme could effectively improve harmonicity of speech while retaining the aperiodic part of unvoiced speech components. In their implementation of [33] the framework of a canonical speech representation [17], as described in Section 2.3.3, was used. The two-stage transform (see Section 3.1) was applied to the warped LP residual $e_{wrp}(n)$ of the noisy speech signal $y(n)$. The f_0 estimation was performed by the Instantaneous Pitch Frequency (IPF) algorithm [30], which was briefly outlined in Section 2.3.3 and tested in Section 4.1.1. The segmentation of $e_{wrp}(n)$ into frames was determined based on an energy concentration criterion [16, 17].

5.1.1 Fixed Weighting

Periodicity enhancement can be achieved by adjusting the energy balance of the modulation bands in the transform domain. The fixed weighting method uses relatively heavier weights applied to lower modulation bands and lighter weigths to higher bands, while the weights along frequency channels were kept constant. The weighting function W_q is defined as

$$W_q = \max\left(1 - \frac{q}{3}, 0\right), \tag{5.2}$$

i.e. $W_0 = 1$, $W_1 = 2/3$, $W_2 = 1/3$, and $W_q = 0$ for $q \geq 3$ [32,33]. The proposed weighting scheme is now adapted to our DS framework (Figure 3.1), where we omit the LP analysis and constant pitch warping step. Figure 5.2 shows a block diagram of a speech enhancement system using fixed weights as suggested in [33], implemented in the DS domain. Note that in Figure 5.2 the superscript (l) denotes the index for the l-th time block, $\hat{f}_0$ describes a pitch estimator and OLA is the overlap-and-add procedure applied to reconstruct signal segments at the synthesis stage. Speech enhancement in terms of PE is achieved by using the weighting function W_q directly as a gain function $G(q,k)$ applied to the noisy DS as

$$\widehat{DS}_x(q,k) = G(q,k)DS_y(q,k), \tag{5.3}$$

where $G(q,k) = W_q$ in the case of fixed weighting.

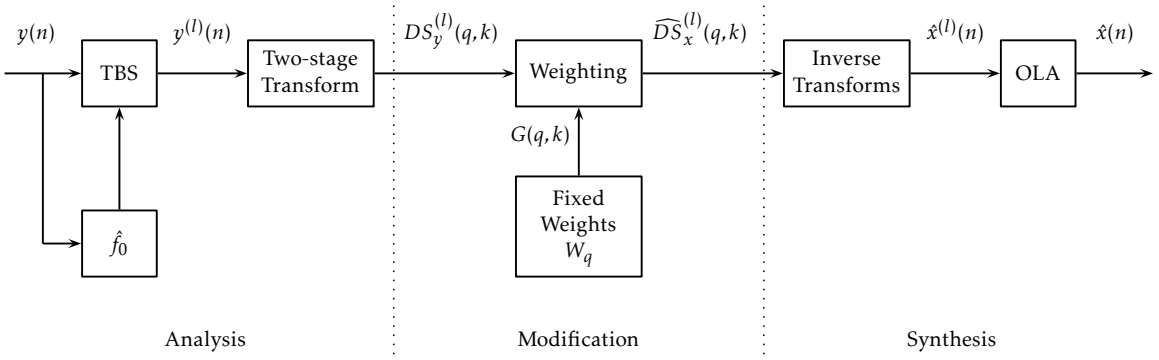


Figure 5.2: Framework for using PE with a fixed weighting scheme according to [33] in the DS domain.

A problem associated with this method is the artificial harmonization of speech during speech pauses and the strong attenuation of unvoiced components and VUV transitions. This issue may be resolved with the help of a voicing measure that indicates when to restore periodicity and when to suppress noise. This leads to an adaptive weighting approach.

5.1.2 Adaptive Weighting

The motivation for an adaptive weighting scheme in the transform domain is to selectively modify voiced and unvoiced regions. Using a periodicity measure for VUV soft-decision we are able to dynamically adjust the weights W_q as follows: For voiced segments, the weights are set so as to restore periodic components and as well as the harmonic structure of noisy speech. For unvoiced segments the weights are designed in a way to preserve aperiodic components and to reduce noise during speech absence. The measure used as a degree of voicing is the normalized energy of the first modulation band with respect to the mean energy of the signal segment [33]:

$$\tilde{E}_1 = \frac{E_1}{\zeta^2}, \tag{5.4}$$

where E_1 is the energy of the first modulation band and ζ is the root-mean-square (RMS) value of the signal segment. Since it is desired to have intermediate weights for segments which are not clearly classified as voiced or unvoiced, $\tilde{E}_1$ is used to derive a smooth weighting function as

$$W_q = \max\left(s_1(\tilde{E}_1) + s_2(\tilde{E}_1) \cdot q, 0\right). \tag{5.5}$$

The functions $s_1(\tilde{E}_1)$ and $s_2(\tilde{E}_1)$ control the balance between PE and signal preservation. In [32, 33] these functions are implemented based on a sigmoid function:

$$s_1(\tilde{E}_1) = A + \frac{1-A}{1+e^{-\alpha(\tilde{E}_1-\beta)}}, \tag{5.6}$$

$$s_2(\tilde{E}_1) = -\frac{1}{3} + \frac{1}{3} \cdot \frac{1}{1+e^{\alpha(\tilde{E}_1-\beta)}}, \tag{5.7}$$

where A defines the desired attenuation level for unvoiced segments, and α and β control the slope and threshold between PE and noise reduction, respectively. In the implementation of [33], the values of A, α and β were set empirically as $A = 0.5$, $\alpha = 0.1$ and $\beta = 73$.

The weighting rule defined in Equation (5.5) can be used for a gain function applied to the noisy signal in the DS domain. Similar to the fixed weighting scheme, we set $G(q,k) = W_q$. A block diagram of the adaptive weighting scheme using the DS framework is presented in Figure 5.3.

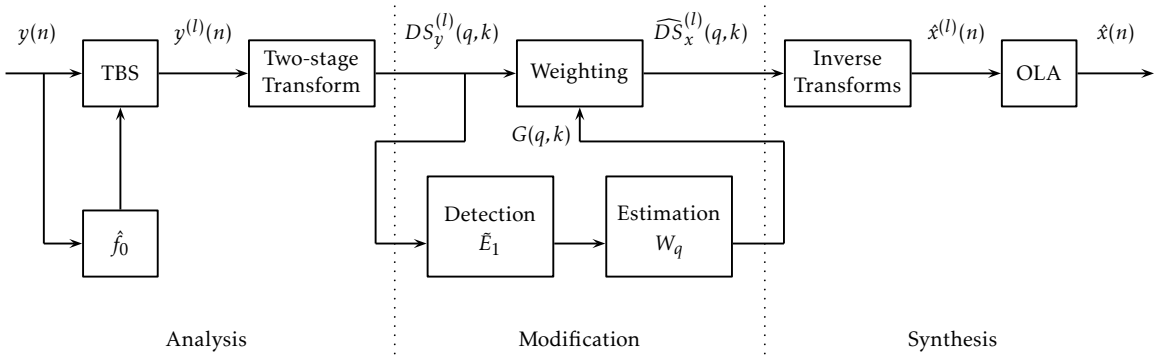


Figure 5.3: Framework for using PE with an adaptive weighting scheme according to [33] in the DS domain. In the modification stage the normalized energy of the first modulation band $\tilde{E}_1$ is evaluated and used for the design of the weighting function W_q.

In their studies [32, 33], Huang et. al reported the effectiveness of adaptive weights which are dynamically adjusted according to the energy concentration in the transform domain. However, the weighting function was only defined along modulation bands q but independent of the frequency channels k. We argue that making use of both dimensions as a joint weighting is the key to develop an even more effective noise suppression rule.

5.2 Noise Suppression by Adaptive Double Spectrum Weighting

Similar to the weighting schemes of Huang et. al [33], described in Section 5.1, the method we propose in this Section is designed to find a speech target estimate by applying a weighting to the noisy DS. In contrast to [33], our method facilitates a joint weighting along the k and the q-axis, to make use of joint frequency and modulation frequency dimensions. This method, called Adaptive Double Spectrum Weighting (ADSW), is submitted to Interspeech 2016, which is currently under review [52]. The submitted version of the paper is found in the Appendix. The starting point for our derivation of ADSW is again

$$\widehat{DS}_x(q,k) = G(q,k)DS_y(q,k), \tag{5.8}$$

where $G(q,k)$ is in this case a cascade of two weighting schemes: $W_e(q,k)$ to dampen noise-dominant coefficients, and $W_q(q,k)$ to enhance the harmonicity of speech similar to [33]. Each scheme is described in the following.

5.2.1 Energy Based Coefficient Weighting W_e

The first weighting used in ADSW is an energy based coefficient weighting $W_e(q,k)$ which compares the energy of each DS coefficient with respect to the mean energy of $DS_y(q,k)$, resulting in the relative energy $E_{rel}(q,k)$ defined as

$$E_{rel}(q,k) = KQ\ \frac{|DS(q,k)|^2}{E_{1:Q}}. \tag{5.9}$$

Since E_{rel} shows a broad dynamic range, we apply the decadic logarithm as a non-linear mapping function. From our observations we draw the conclusion that coefficients with low energy in a range of approximately $E_{rel} \in [0,1]$ contribute to noise, whereas high energy coefficients, i.e. $E_{rel} \in [10,\infty]$ relate to the speech class. However, this depends on the SNR level and the performance of TBS which is responsible for compact rendering of DS coefficients in a small

subspace. The better the TBS in the DS analysis stage, the stronger the energy is concentrated in a certain coefficient $DS(q,k)$. Additionally, we constrain the weights to a lower bound of 0 by adding 1 to E_{rel} and get

$$W_e(q,k) = \log(E_{rel}(q,k)+1)\,. \tag{5.10}$$

The purpose of W_e is to deemphasize coefficients that are likely to carry noise information ($W_e < 1$) and to give more weight to coefficients which could be attributed to the speech target ($W_e > 1$). Figure 5.4 shows an exemplary representation of $E_{rel}(q,k)$ and Figure 5.5 plots the curve of the energy based weighting function W_q. Note that our choice for a logarithmic mapping

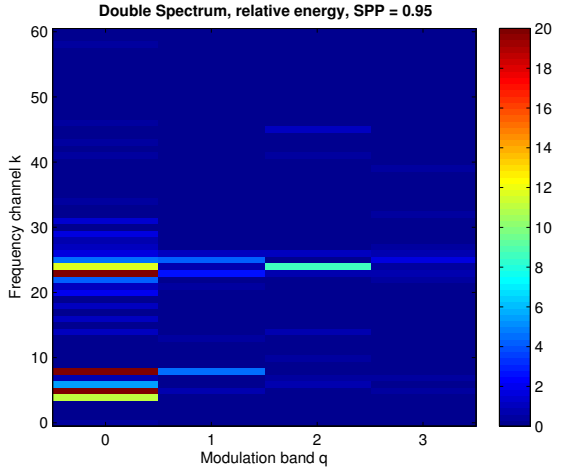


Figure 5.4: Relative energy $E_{rel}(q,k)$ of a DS taken from a noisy speech segment, SPP = 0.95. The energy is concentrated in only few coefficients with $E_{rel} \gg 1$.

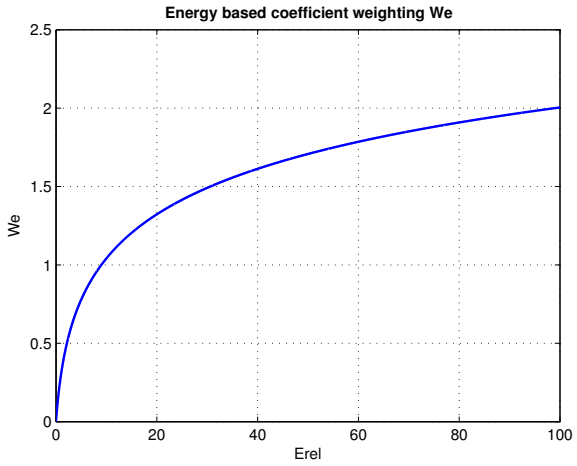


Figure 5.5: Logarithmic mapping function of W_e in dependence of the relative DS energy E_{rel}.

function, which serves as coefficient compression, is empirically chosen and not ideal any case. In general, compression functions have a great influence on the enhanced signal output, and in particular on the speech quality [53]. Since base-ten logarithmic compression is one of the most popular types used in speech processing [3, 54], we chose the design of the energy based weighting function accordingly.

5.2.2 Harmonicity Enhancement by Modulation Band Weighting W_q

As the second weighting, we propose $W_q(q,k)$ to enhance the harmonicity and periodicity of noisy speech, similar to the PE methods [32, 33] described in Section 5.1. To this end, we need a periodicity indicator for each frequency channel. Similar to Equation (4.1), we consider the MBR of the respective frequency band, MBR_k given by

$$\mathrm{MBR}_k = \frac{|DS(0,k)|^2}{\sum_{q=0}^{Q-1}|DS(q,k)|^2}. \tag{5.11}$$

In contrast to the fixed and adaptive weighting methods of [33], we propose an exponentially decaying modulation weighting, which turned out to be a better choice in DS. Therefore, we use

$$W_q(q,k) = e^{-\mathrm{MBR}_k q}, \tag{5.12}$$

where MBR_k serves as the decay factor of the exponential weighting. An example for the weighting matrix W_q is presented in Figure 5.6 by means of a DS image. Figure 5.7 exemplifies

the exponential decaying characteristic in $W_q(q,k)$ for different frequency channels k and across all modulation bands q.

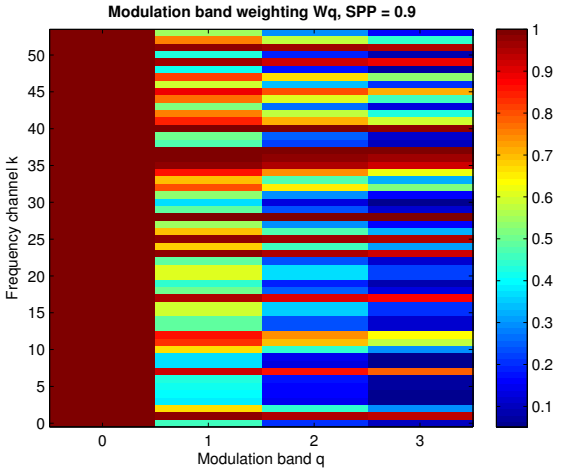


Figure 5.6: Example of modulation band weighting W_q for a voiced frame (SPP = 0.9). Frequencies yielding higher values for MBR_k show a strong decay over q.

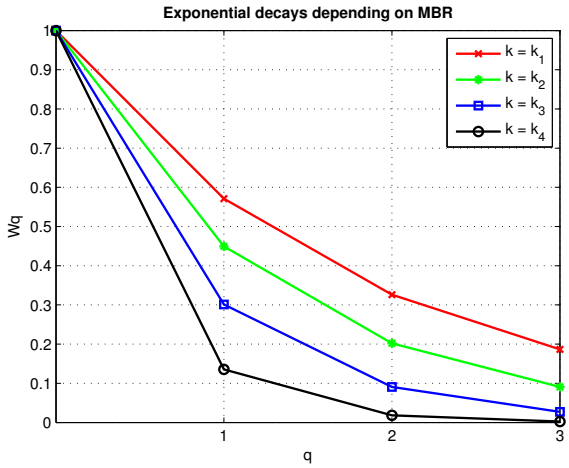


Figure 5.7: W_q as an exponentially decaying function shown for different frequency channels, e.g. $k_1 = 2000$ Hz, $k_1 = 700$ Hz, $k_1 = 500$ Hz, $k_1 = 200$ Hz.

To have a selective noise suppression, similar to conventional STFT-based speech enhancement methods [1,19], we utilize the SPP computed in the DS domain as described in 4.2. The SPP is applied as a scaling factor on the cascade weighting outcome, which gives us the gain function for ADSW as

$$G(q,k) = \text{SPP} \cdot W_e(q,k) W_q(q,k). \tag{5.13}$$

Finally, we restrict the SPP-driven $G(q,k)$ to a lower limit $G_{\min} = 0.1778 \triangleq -15$ dB according to [55]. This serves the purpose of reducing the noise to a certain comfort noise level, which may be more convenient to a listener. Applying the lower limit to the gain function, this yields

$$G(q,k) = G_{\min} \quad \text{if} \quad G(q,k) < G_{\min}. \tag{5.14}$$

To conclude this Section, Figure 5.8 presents a block diagram that shows the essential steps in a DS-based speech enhancement system using the ADSW algorithm.

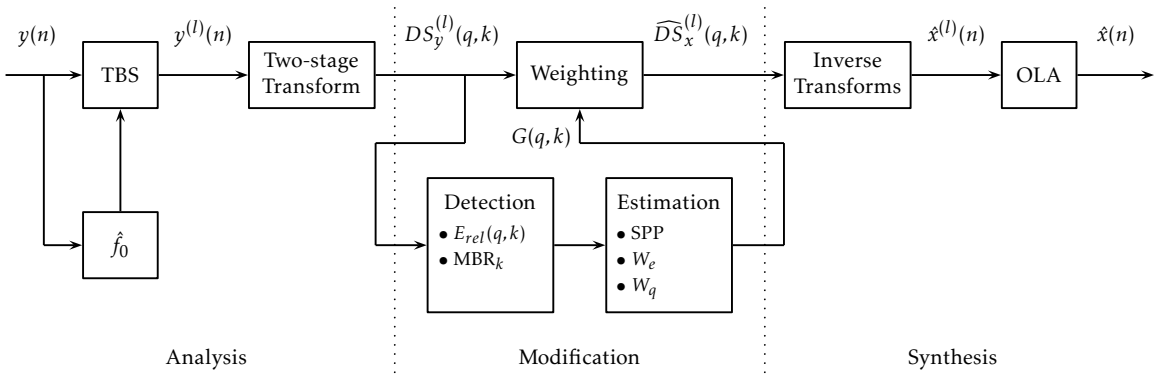


Figure 5.8: Framework of the ADSW speech enhancement algorithm. The modification stage is comprised of a detection block to compute the relative DS energy $E_{rel}(q,k)$, and of an estimation block to compute the SPP and the two weighting functions used in the design of the gain function $G(q,k)$.

5.3 The Wiener Solution for Noise Suppression in Double Spectrum

In the previous Sections, we described three noise suppression rules: a fixed weighting and an adaptive weighting scheme according to [33], and our proposed ADSW. The clean estimates in the DS domain were obtained by applying a gain function $G(q,k)$ to the noisy DS as described in Equation (5.3). By rewriting this weighting rule it can be shown that the multiplication by $G(q,k)$ is equivalent to a DS spectral subtraction rule defined as

$$\widehat{DS}_x(q,k) = DS_y(q,k) - \widehat{DS}_d(q,k) = DS_y(q,k)\left(1 - \frac{\widehat{DS}_d(q,k)}{DS_y(q,k)}\right), \tag{5.15}$$

where

$$G(q,k) = \left(1 - \frac{\widehat{DS}_d(q,k)}{DS_y(q,k)}\right). \tag{5.16}$$

By assuming linearity in the DS domain, the optimal gain function $G(q,k)$ takes into account a noise estimate $\widehat{DS}_d(q,k)$. Finding an accurate noise estimate is hence another way to design an effective speech enhancement algorithm. Typically, in STFT-based methods, the noise power spectral density (PSD) can be estimated using either of the following approaches [1]:

a) based on a voice activity detector (VAD) approach, where regions of speech absence are used to update the noise estimate e.g. by recursive averaging,

b) based on minimum power level tracking, where the noise estimate is updated even during speech presence at frequency bins corresponding to valleys between strong spectral peaks,

c) based on SPP, where the noise estimate is update following a soft decision rule between speech presence and speech absence.

Motivated by another work of Huang [34], we discuss the principles of a Wiener filter and derive its gain function in the DS domain. Wiener filters [50] are based on the MMSE optimisation criterion, i.e. minimizing the MMSE between the output signal $\hat{x}(n)$ of a given system and the desired reference signal $x(n)$. Wiener filters can be implemented in either time of frequency [3]. In our case, the DS domain Wiener filter $H(q,k)$ is designed to minimize the mean-square error (MSE) $\epsilon(q,k)$ between the clean DS, $DS_x(q,k)$ and the estimate of the clean target $\widehat{DS}_x(q,k)$ as

$$\epsilon(q,k) = \mathbb{E}\left\{|DS_x(q,k) - \widehat{DS}_x(q,k)|^2\right\} \tag{5.17}$$

$$= \mathbb{E}\left\{|DS_x(q,k) - H(q,k)DS_x(q,k)|^2\right\}, \tag{5.18}$$

where $\mathbb{E}\{.\}$ denotes the expectation operator. For further derivation steps the indices (q,k) are omitted for improved readability. Equation (5.18) is expanded to

$$\epsilon = \mathbb{E}\left\{|DS_x|^2 - 2H\,DS_xDS_y + H^2|DS_y|^2\right\} \tag{5.19}$$

$$= \mathbb{E}\left\{|DS_x|^2\right\} - 2H\,\mathbb{E}\Big\{DS_xDS_y\Big\} + H^2\,\mathbb{E}\Big\{|DS_y|^2\Big\}, \tag{5.20}$$

where the term $\mathbb{E}\{|DS_a|^2\} = DS_{aa}$ is the PSD and $\mathbb{E}\{DS_aDS_b\} = DS_{ab}$ is the cross-PSD in the DS domain. The optimal filter in a MMSE sense is found by minimizing the MSE, which is achieved by setting the derivative of ϵ to zero and solving for H. The general form is then given

by

$$H(q,k) = \frac{DS_{xy}}{DS_{yy}}. \tag{5.21}$$

The cross-PSD denoted by DS_{xy} can be expressed as

$$\begin{aligned} DS_{xy} &= \mathbb{E}\left\{DS_x DS_y\right\} && (5.22)\\ &= \mathbb{E}\left\{\big(DS_y - DS_d\big)\big(DS_x + DS_d\big)\right\} && (5.23)\\ &= \mathbb{E}\left\{DS_y\big(DS_x + DS_d\big)\right\} - \mathbb{E}\Big\{DS_d DS_x\Big\} - \mathbb{E}\Big\{|DS_d|^2\Big\}. && (5.24) \end{aligned}$$

Assuming that the clean signal and the noise signal are stochastically uncorrelated in both time and DS domains, the cross-PSD $\mathbb{E}\Big\{DS_n DS_x\Big\}$ cancels, which yields

$$\begin{aligned} DS_{xy} &= \mathbb{E}\Big\{|DS_y|^2\Big\} - \mathbb{E}\Big\{|DS_d|^2\Big\} && (5.25)\\ &= DS_{yy} - DS_{dd}. && (5.26) \end{aligned}$$

Inserting (5.26) into (5.21), the Wiener solution is simplified to

$$H(q,k) = \frac{DS_{yy} - DS_{dd}}{DS_{yy}}. \tag{5.27}$$

In practice, $H(q,k)$ can be expressed in terms of magnitude DS as

$$G_w(q,k) = \frac{|DS_y|^2 - |DS_d|^2}{|DS_y|^2}, \tag{5.28}$$

where $G_w(q,k)$ is the Wiener gain function.

Similar to [46,56] we introduce the *a priori* SNR ξ (also known as *prior* SNR) and the *a posteriori* SNR γ (also known as *posterior* SNR). The prior SNR defines the ratio between the clean and the noise PSD, which in the DS domain is

$$\xi(q,k) = \frac{DS_{xx}}{DS_{dd}}. \tag{5.29}$$

The posterior SNR is the ratio between noisy and noise PSD:

$$\gamma(q,k) = \frac{DS_{yy}}{DS_{dd}}. \tag{5.30}$$

Using the definition of the prior SNR, we can rewrite Equation (5.27) as

$$G_w(q,k) = \left(1 + \frac{1}{\xi(q,k)}\right)^{-1}, \tag{5.31}$$

to get the Wiener gain function in terms of ξ as

$$G_w(q,k) = \frac{\xi(q,k)}{1 + \xi(q,k)}. \tag{5.32}$$

Since DS_{xx} and DS_{dd} cannot be directly accessed, the prior SNR needs to be estimated. One popular method for computing ξ is the *decision-directed* approach by Ephraim and Malah (1984) [46]. In their proposal, an estimate of ξ is obtained by recursive averaging of the prior SNR of

the past frame. In the DS domain, the decision-directed method is defined as

$$\xi^{(l)}(q,k) = \alpha \frac{|\widehat{DS}_x^{(l-1)}(q,k)|^2}{|\widehat{DS}_d^{(l-1)}(q,k)|^2} + (1-\alpha)\max\left(\frac{|DS_y^{(l)}(q,k)|^2}{|\widehat{DS}_d^{(l)}(q,k)|^2} - 1, 0\right), \tag{5.33}$$

where α is a smoothing constant which works as a forgetting factor in the recursive averaging procedure. The term $|DS_y(q,k)|^2/|\widehat{DS}_d(q,k)|^2$ is an approximation of the posterior SNR γ, which is subtracted by 1 to give the so called *instantaneous* SNR, interpreted after Ephraim and Malah [46]. The instantaneous SNR is used to update the prior SNR estimation. The advantage of using the decision-directed method is that the "musical noise" distortion [57] can be effectively suppressed due to temporal smoothing. However, the degree of musical noise elimination depends on the choice of α, on the lower bound of the prior SNR ξ_{min} which determines the residual noise level, and on the type of noise [55, 58].

In Figure 5.9 we present the framework of our proposed speech enhancement method, Double Spectrum Wiener Filter (DS-Wiener). A detailed description about parameter choice and noise estimation is given in Section 6.1.2.

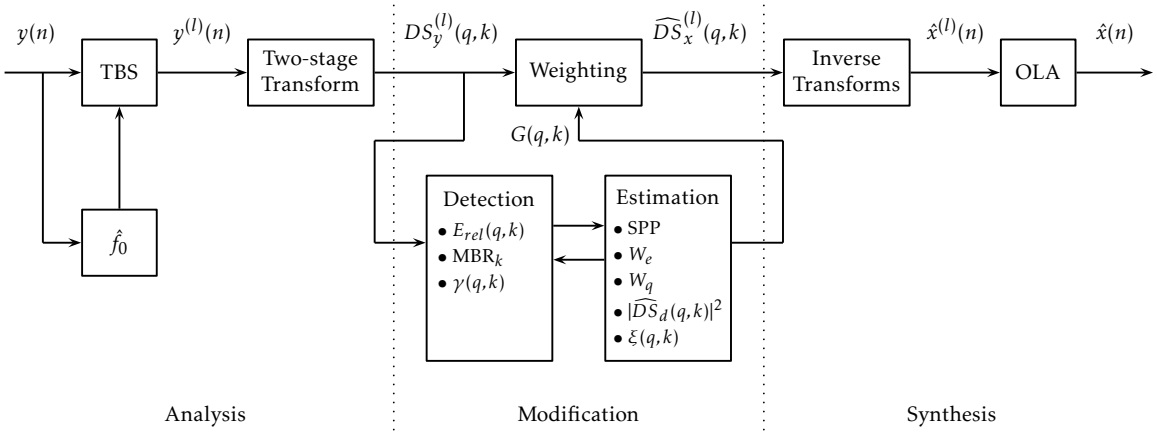


Figure 5.9: Framework of the DS-Wiener speech enhancement algorithm. In the modification stage we use a joint detection-estimation procedure in combination with the decision-directed approach for the estimation of the prior SNR ξ and for the design of the Gain function $G(q,k)$.

6

Experiments and Results

In this Chapter, we demonstrate the effectiveness of the proposed DS-based speech enhancement methods. First, we will compare the performance of the four DS-based methods, described in Chapter 5, in a blind scenario. In addition, we will discuss the upper bounds of the performance in terms of the f_0-oracle scenario where f_0 is estimated from the clean signal. Second, we choose the most effective DS-based method and compare its performance against the STFT-based and STSM-based benchmarks. Finally, we discuss the potentials and limits of the proposed DS speech enhancement systems.

6.1 Experimental Setup

Before we evaluate the different speech enhancement algorithms, we describe the setup chosen for our experiments. For this purpose, a detailed summary of the speech and noise databases, speech enhancement methods, parameter setup, and evaluation criteria are given in this Section.

6.1.1 Speech and Noise Databases

In our experiments we use two speech databases for performance evaluation. For the experiments described in Section 6.2.1 and Section 6.2.2, clean and noisy speech utterances were taken from NOIZEUS speech corpus [47] consisting of 30 phonetically-balanced sentences uttered by three male and three female speakers (average length of 2.6 seconds). The speech files were downsampled from the original sampling frequency of 25 kHz to 8 kHz and filtered to simulate telephony speech. The NOIZEUS corpus comes with non-stationary noises mixed at different SNRs taken from the AURORA database [59]. For the noisy speech scenario we use three noise types: white, babble and car noise. In the experiment of Section 6.2.3 we conduct performance evaluation on 18 sentences of the TIMIT corpus [60]. TIMIT contains a total of 6300 sentences (10 sentences spoken by each of 630 speakers from 8 major dialect regions of the USA). In our test set we use 8 utterances by a male and 10 utterances by a female speaker, both from the dialect region "New England". A list of the sentences used is given in the Appendix. For noisy utterances we use the files provided by TIMIT, corrupted with white, babble and factory noise. In all experiments the noisy files were mixed at an input SNR range of 0 to 10 dB, in steps of 5 dB to simulate heavy, medium and mild noise scenarios.

6.1.2 Speech Enhancement Methods and Parameter Setup

In this Section, we describe implementation aspects and parameter choice for each algorithm used in the performance evaluation. Every DS-based method uses the DS-f_0 algorithm for the TBS procedure, as described in Sections 4.1.1 and 4.1.2, and operates with a fixed number of modulation bands which is set to $Q = 4$, according to Section 3.5.2. The overlap between time

blocks is chosen as $(Q-1)\tilde{P}_0$, which relates to a step size of $\tilde{P}_0$. The window used in the analysis and synthesis stage is a square-root Hann window as described in Section 3.1.2.

DS-PE-fxd

The fixed weighting method proposed by Huang et. al [33], which we implemented using the DS framework (Figure 5.2), is referred to as Double Spectrum Periodicity Enhancement using Fixed Weighting (DS-PE-fxd). In contrast to the original implementation, where the weights were chosen according to Equation (5.2), we compute the fixed weights as

$$W_q = \max\left(1 - \frac{q}{2}, 0\right), \tag{6.1}$$

i.e. $W_0 = 1$, $W_1 = 1/2$, and $W_q = 0$ for $q \geq 2$. We justify this choice by taking into account the fixed number of modulation bands $Q = 4$. As reasoned in Section 3.5.3, setting $Q = 4$ leads to modulation frequencies which are higher than 16 Hz for $q = 2$, and thus of little relevance for speech information.

DS-PE-adp

The adaptive counterpart of the fixed weighting method [33], implemented using the framework given in Figure 5.3, is called Double Spectrum Periodicity Enhancement using Adaptive Weighting (DS-PE-adp). Similar to the fixed case, the weighting function is modified so as to attenuate the energy for $q \geq 2$. This changes the sigmoid function s_2 used in the weighting scheme as

$$s_2(\tilde{E}_1) = -\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{1 + e^{\alpha(\tilde{E}_1 - \beta)}}. \tag{6.2}$$

In our implementation we set the attenuation parameter $A = 0.1778$, corresponding to a lower limit of -15 dB [55]. The slope and the threshold parameters were set as in the proposal as $\alpha = 0.1$ and $\beta = 73$, respectively [33].

ADSW

In the implementation of ADSW the SPP was calculated according to Section 4.2.2 for every time block. The weighting functions were designed as described in Section 5.2 and the lower limit of the gain function $G(q,k)$ was set to $G_{\min} = 0.1778 \triangleq -15$ dB [55].

DS-Wiener

The DS-Wiener algorithm is a more sophisticated version of the ADSW. The gain function derived for ADSW is used to obtain a noise estimate using Equation 5.16 rewritten as

$$\widehat{DS}_d(q,k) = DS_y(q,k) - G(q,k)DS_y(q,k). \tag{6.3}$$

To mitigate oversubtraction, which may lead to unwanted "musical noise" artifacts [57], we constrain the gain function in the subtraction process to

$$\tilde{G}(q,k) = \max\left(G(q,k), 1 - \varepsilon\right), \tag{6.4}$$

where ε is a small regularization constant to avoid division by zero in the computation of the prior or posterior SNR, ξ and γ, respectively. We argue that a noise estimate computed in this

way can be used as an alternative to existing noise estimation approaches listed in Section 5.3. The noise estimate thus obtained is then squared and recursively averaged in order to get a noise PSD estimate in the DS domain, which is needed for the calculation of ξ and γ.

Since the dimensions of DS change over time due to the time-varying nature of pitch, the update of the noise PSD needs to be interpolated. In other words, the number of frequency channels K of a noise estimate may be different at time blocks l and $l+1$. For this reason, a *global* noise estimate $\widehat{DS}_{d,glo}(q,k)$ is employed, which is obtained by interpolating the noise DS with K frequency channels to get a new DS with K_{glo} frequency channels. This is done by cubic interpolation along k, for each modulation band q. The global noise DS is initialized so as to have an adequate number for K_{glo}, which can be determined by taking into account the smallest f_0 in the pitch estimation procedure (Section 4.1.1). For example, let $f_{0,min} = 80$ Hz, then the maximum normalized pitch period yields $\tilde{P}_0 = 100$, which equals the maximum number of frequency channels K_{glo} occurring in the DS analysis stage. In our case we set $K_{glo} = 100$. The noise PSD in the DS domain, denoted by $\eta_d(q,k)$, is then estimated by recursive averaging as

$$\eta_d^{(l)}(q,k) = \lambda\, \eta_d^{(l-1)}(q,k) + (1-\lambda)\, \mathcal{I}\left\{ |\widehat{DS}_{d,glo}^{(l)}(q,k)|^2 \right\}, \tag{6.5}$$

where $\eta_d^{(l)}(q,k)$ is the estimated DS noise PSD of the l-th time block, λ is a smoothing constant, and $\mathcal{I}\{.\}$ denotes an operator that interpolates the given noise estimate to match the dimensions of the global estimate. Once the global noise PSD is computed for the current time block, its corresponding DS is interpolated again to come back to the actual dimension of the current DS. This interpolation procedure obviously introduces small errors to the noise estimate, which are however negligible compared to the noise estimation process itself. In our implementation the smoothing constant λ was set to 0.998 for white noise and 0.995 for car, babble and factory noise.

In the decision-directed estimation of the prior SNR ξ(q,k) we set the forgetting factor to $\lambda = 0.98$. The initialization of ξ is selected as

$$\xi^{(0)}(q,k) = \alpha + (1-\alpha)\max\left(\gamma^{(0)}(q,k) - 1\right) \tag{6.6}$$

following the suggestion of Ephraim and Malah [46]. The Wiener gain function $G_w(q,k)$ is bounded by the lower limit $G_{\min}$, which for white noise we chose corresponding to -25 dB, for car noise as -20 dB, and for babble noise as -15 dB.

To demonstrate the effectiveness of DS based methods, we include three benchmarks: 1) Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator (MMSE-STSA) [46] as an STFT-based speech enhancement benchmark, 2) Spectral Subtraction in the Short-Time Modulation Domain (ModSpecSub) as an STSM-based method relying on the spectral subtraction principle [4], and 3) Minimum Mean-Square Error Short-Time Spectral Modulation Magnitude Estimator with Speech Presence Uncertainty (MMESPU) as a statistical-model-based benchmark method in the STSM domain [25] taking into account the SPU for a fair comparison to DS-based methods relying on SPP.

MMSE-STSA

For MMSE-STSA the decision-directed approach for estimating the prior SNR was used according to [46]. The forgetting factor α used for recursive averaging was set to 0.98. As a noise estimator we chose the Minimum Statistics method [61]. The STFT frame setup was implemented with an analysis window of type Hamming, with a length of 32 ms, and a frame shift of 16 ms.

ModSpecSub

For ModSpecSub we used the implementation provided by Paliwal et. al [4]: The acoustic frame duration was set to 32 ms, with a frame shift of 8 ms, whereas the modulation frame duration was 256 ms with 32 ms frame shift. The Hamming window was used for both the acoustic and the modulation STFT setup. For the spectral subtraction rule according [57], magnitude-squared spectra were used in combination with a spectral floor parameter $\beta = 0.002$. The subtraction factor ρ was selected as described in [57]. The noise estimate of modulation magnitude spectra was obtained based on a decision from a simple VAD [19] applied in the modulation domain. The noise estimate was updated during speech absence using a recursive averaging rule with a forgetting factor $\lambda = 0.98$ [4].

MMESPU

For MMESPU we also used an implementation provided by Paliwal et. al [25]. In their study, the parameters were subjectively tuned and reported as follows: The acoustic frame duration was set to 32 ms, with a frame shift of 1 ms and the modulation frame duration was selected as 32 ms with a modulation frame shift of 2 ms. The Hamming window was used for both the acoustic and the modulation STFT setup. The noise estimate was obtained in a similar fashion as described in ModSpecSub, but with a VAD by Sohn (1999) [62] with a threshold set to 0.15 and a forgetting factor $\lambda = 0.98$. For estimation of the prior SNR, the decision-directed approach of [46] was used with a smoothing factor $\alpha = 0.995$ and a lower limit of $\xi_{\min} \triangleq -25$ dB. In the implementation of the gain function under assumption of SPU, the probability for speech presence in the modulation domain was set to 0.3.

6.1.3 Evaluation Criteria

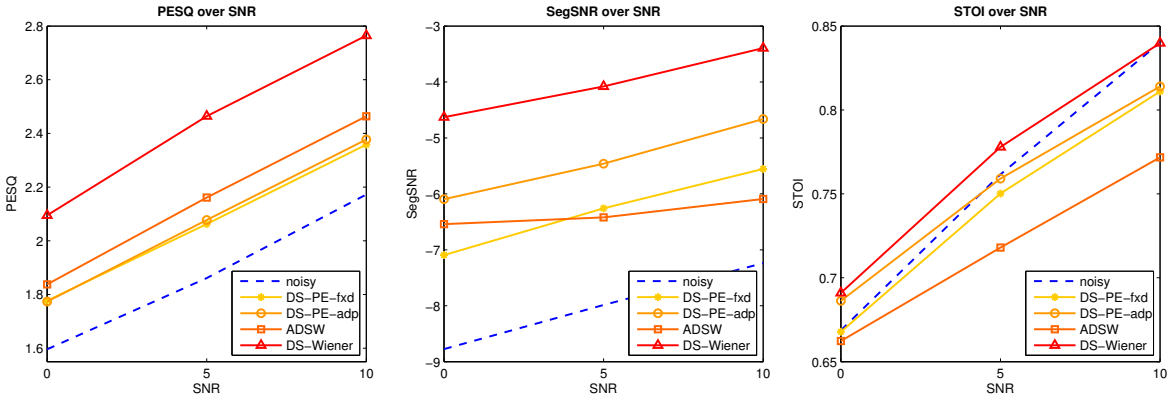
As evaluation criteria we chose the three following objective measures: Perceptual Evaluation of Speech Quality (PESQ) [63] and Segmental SNR (SegSNR) [64] as measures for speech quality, and Short-Time Objective Intelligibility (STOI) [65] as a speech intelligibility measure. We report speech enhancement results compared to the unprocessed noisy speech signal in terms of these measures as well as in terms of ΔPESQ, ΔSegSNR and ΔSTOI, which represent the absolute improvement compared to the noisy signal. The results obtained for every speech utterance evaluated at an SNR of 0 dB, 5 dB and 10 dB are averaged over all speakers and presented as a mean score.

6.2 Results

In this Section, we present the results of different speech enhancement experiments as outlined in the beginning of this Chapter.

6.2.1 Comparison of Double Spectrum-Based Methods

This Section deals with the comparison of DS-based speech enhancement algorithms described in Section 6.1.2. For the performance evaluation in terms of PESQ, SegSNR and STOI, we use 30 utterances from the NOIZEUS database. Figure 6.1 shows the results for white noise in a blind noise scenario, i.e. only the noisy signal is known. The corresponding Delta values of PESQ, SegSNR and STOI are presented in Table 6.1. Since the DS-based methods rely on the



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.1: Objective evaluation scores for DS-based methods averaged over 30 speakers of NOIZEUS database, blind scenario, white noise: (a) PESQ, (b) SegSNR, (c) STOI

TBS which depends on pitch estimation, we also provide the evaluation results of each method for the f_0-oracle scenario in Table 6.1.

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
DS-PE-fxd	0.18	0.20	0.19	1.68	1.73	1.69	0.00	-0.01	-0.03
f0-oracle, DS-PE-fxd	0.24	0.25	0.23	1.79	1.82	1.78	0.03	0.00	-0.02
DS-PE-adp	0.18	0.25	0.24	2.68	2.52	2.58	0.02	0.00	-0.03
f0-oracle, DS-PE-adp	0.22	0.25	0.24	3.03	2.93	2.89	0.03	0.00	-0.03
ADSW	0.24	0.30	0.29	2.23	1.56	1.15	-0.01	-0.04	-0.07
f0-oracle, ADSW	0.25	0.31	0.29	2.40	1.76	1.30	0.00	-0.04	-0.06
DS-Wiener	0.50	0.60	0.59	4.15	3.91	3.84	0.02	0.02	0.00
f0-oracle, DS-Wiener	0.55	0.63	0.60	4.07	3.86	3.85	0.03	0.02	0.00

Table 6.1: Delta scores for DS-based methods averaged over 30 speakers of NOIZEUS database, blind and f_0-oracle scenario, white noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

It can be observed that the DS-Wiener performs best in terms of PESQ, SegSNR and STOI in both blind and f_0-oracle scenario. Similar to conventional state-of-the-art speech enhancement algorithms, the DS-Wiener effectively suppresses noise without introducing perceptible speech distortion that degrades the intelligibility. For low and medium SNR levels, i.e. 0 dB and 5 dB, we have a joint improvement of quality and intelligibility in terms of PESQ and STOI. The periodicity enhancement methods DS-PE-fxd and DS-PE-adp perform similar, however since DS-PE-adp is capable of reducing noise during speech absence, this results in higher SegSNR scores. From a subjective perspective both methods introduce a harsh buzzy sound due to unnatural over-harmonization of the noise, which is less annoying in stimuli generated by the adaptive scheme. The ADSW method suppresses noise only to a certain residual noise floor and leaves room for improvement. This method is aggressive in terms of filtering aperiodic components. As a consequence the fine structure of the speech target is destroyed to some extent, which results in a low-pass characteristic of the enhanced signal. This is mainly the reason for degradation in STOI compared to the noisy signal. Perceptually, ADSW stimuli sound more pleasant than those of PE methods, since no unwanted over-harmonization occurs.

On the one hand, the f_0-oracle scenario serves as a realistic upper bound of our proposed methods and gives information about the robustness of the used f_0 estimator. On the other hand, it tells us about the dependency on f_0 for each method: The greater the gap between the blind and the f_0-oracle scenario in terms of performance measures, the more a DS-based method depends on accurate TBS. For a better visualization of this effect, Figures 6.2 and 6.3 show bar plots of the ΔPESQ and ΔSTOI results obtained in the blind and in the f_0-oracle scenario, respectively.

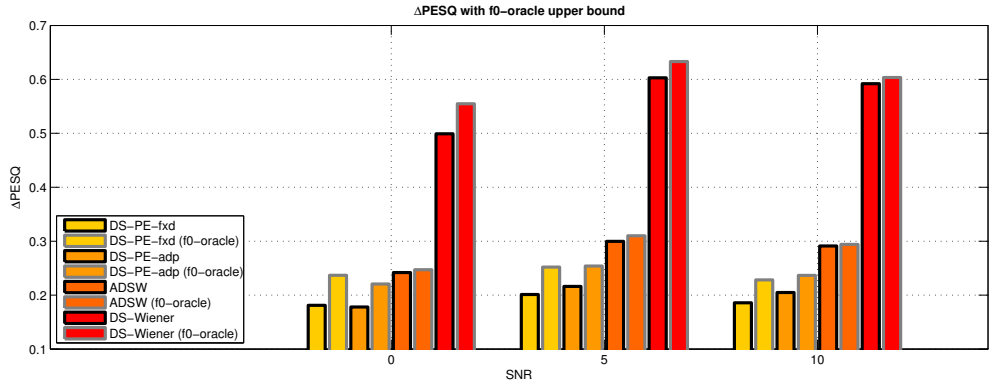


Figure 6.2: Comparison of $\Delta PESQ$ results of DS-based methods, obtained in a blind and in an f_0-oracle scenario. For each method, the right bar (grey frame) represents the f_0-oracle result.

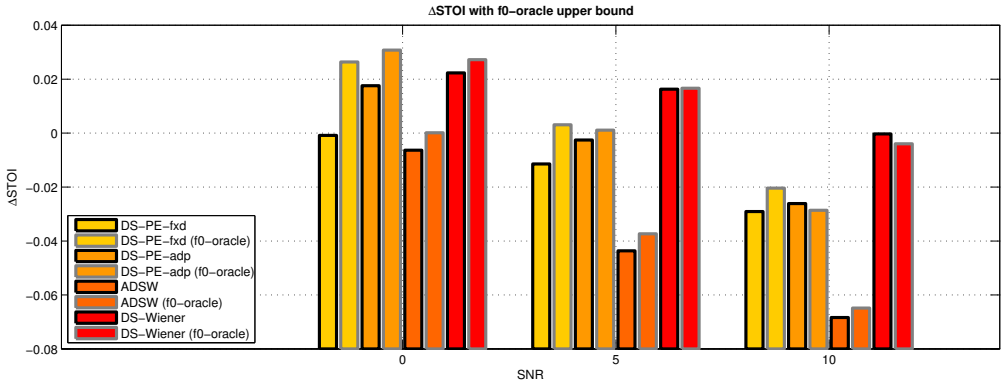
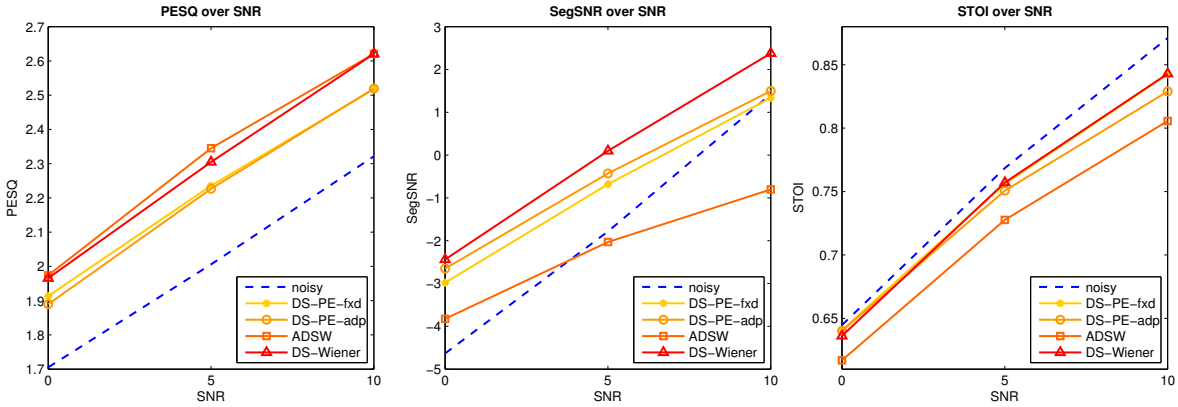


Figure 6.3: Comparison of $\Delta STOI$ results of DS-based methods, obtained in a blind and in an f_0-oracle scenario. For each method, the right bar (grey frame) represents the f_0-oracle result.

We observe relatively large differences between blind and f_0-oracle results for the PE methods. This seems plausible, since these algorithms are designed to reduce noise by modifying the energy distribution along modulation bands only. The ADSW algorithm shows no significant improvement in quality (PESQ), but slight improvement in intelligibility (STOI) assuming f_0 known from the clean signal. The DS-Wiener benefits from f_0-oracle in particular at low SNR levels. Additional bar plot figures illustrating the f_0 dependency of DS-based methods for other noise types (babble, car) are provided in the Appendix of this thesis.

In Figure 6.4 we present blind speech enhancement results for noisy speech utterances corrupted by babble noise. The quality and intelligibility improvement in terms of Delta scores is reported in Table 6.2 for both the blind and the f_0-oracle scenario.

Since babble noise is a non-stationary type of noise, it is more difficult to improve quality and intelligibility. The improvement in PESQ is approximately half compared to the evaluation results obtained in white noise. In this scenario, DS-based methods were not able to improve



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.4: Objective evaluation scores for DS-based methods averaged over 30 speakers of NOIZEUS database, blind scenario, babble noise: (a) PESQ, (b) SegSNR, (c) STOI

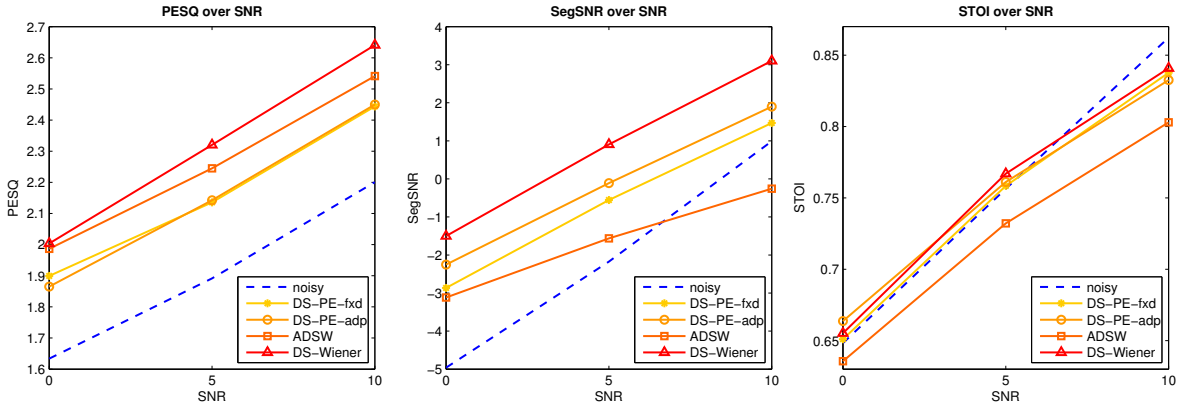
	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
DS-PE-fxd	0.21	0.23	0.20	1.65	1.10	-0.08	0.00	-0.01	-0.03
f0-oracle, DS-PE-fxd	0.35	0.32	0.28	2.75	2.05	0.67	0.06	0.02	-0.02
DS-PE-adp	0.18	0.22	0.20	1.98	1.35	0.08	0.00	-0.02	-0.04
f0-oracle, DS-PE-adp	0.32	0.31	0.28	3.35	2.56	1.12	0.06	0.02	-0.03
ADSW	0.27	0.34	0.30	0.81	-0.25	-2.22	-0.03	-0.04	-0.07
f0-oracle, ADSW	0.33	0.34	0.33	2.09	0.80	-1.31	0.02	-0.01	-0.05
DS-Wiener	0.26	0.30	0.39	2.19	1.88	0.96	-0.01	-0.01	-0.03
f0-oracle, DS-Wiener	0.34	0.35	0.36	3.26	2.66	1.54	0.01	0.00	-0.03

Table 6.2: Delta scores for DS-based methods averaged over 30 speakers of NOIZEUS database, blind and f_0-oracle scenario, babble noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

intelligibility in terms of STOI. Again, the DS-Wiener performs best on average. Due to temporal smoothing of the noise PSD estimate and the prior SNR, it is less prone to introducing speech distortion than the other algorithms.

We now continue our tests in presence of a quasi-stationary colored noise type, car noise. Due to its nature, we expect the performance measures to lie within the range of white noise and babble noise results. The performance evaluation results for the car noise scenario are given in Figure 6.5. Delta scores of blind and f_0-oracle experiments are presented in Table 6.3.

As expected, the speech quality improvement is increased compared to the babble noise scenario. The DS-Wiener and the PE methods enhanced the noisy signal in terms of STOI at low and medium SNR levels. Similar to the previous experiments, the performance of SegSNR and STOI degrades with increased SNR compared to the noisy signal. This may be a result of too aggressive filtering across modulation bands. The fine structure of speech including unvoiced transitions and onsets can be strongly represented in higher modulation bands. A way to deal with preserving such speech components would be to take into account a statistical model of DS coefficients at different SNR levels in the derivation of a noise suppression rule. This has not yet been done for either of the proposed DS-based algorithms.



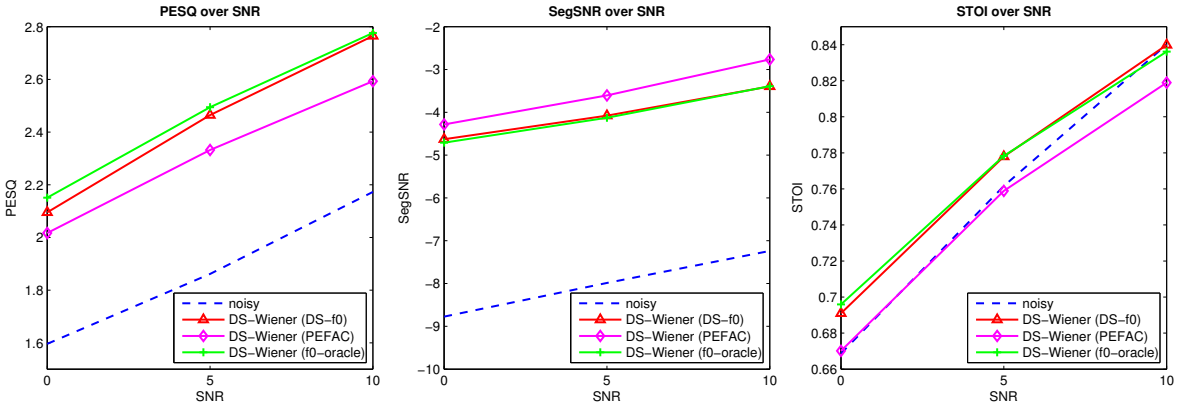
(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.5: Objective evaluation scores for DS-based methods averaged over 30 speakers of NOIZEUS database, blind scenario, car noise: (a) PESQ, (b) SegSNR, (c) STOI

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
DS-PE-fxd	0.27	0.24	0.24	2.09	1.62	0.48	0.00	0.00	-0.02
f0-oracle, DS-PE-fxd	0.31	0.30	0.28	2.83	2.20	0.96	0.04	0.02	-0.02
DS-PE-adp	0.23	0.25	0.25	2.71	2.06	0.91	0.02	0.00	-0.03
f0-oracle, DS-PE-adp	0.27	0.30	0.29	3.54	2.81	1.53	0.05	0.02	-0.03
ADSW	0.35	0.35	0.34	1.84	0.61	-1.25	-0.01	-0.02	-0.06
f0-oracle, ADSW	0.32	0.35	0.35	2.66	1.33	-0.63	0.02	-0.01	-0.05
DS-Wiener	0.37	0.43	0.44	3.46	3.08	2.11	0.01	0.01	-0.02
f0-oracle, DS-Wiener	0.44	0.49	0.47	4.41	3.78	2.46	0.02	0.01	-0.02

Table 6.3: Delta scores for DS-based methods averaged over 30 speakers of NOIZEUS database, car noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

To complete our experiments on the influence of f_0 on the noise reduction performance, we test the DS-Wiener under different conditions: Instead of using DS-f_0 for TBS we use f_0 estimated by PEFAC [44] to compute the pitch periods needed in the TBS procedure. We then compare both DS-Wiener implementations to the f_0-oracle upper bound in a white noise scenario in terms of PESQ, SegSNR and STOI. Note that the f_0-oracle version in this setup is based on DS-f_0 as well. The purpose of this experiment is to justify the choice of using DS-f_0 for TBS in the DS analysis stage, follow-up our observations from Section 4.1.2. The outcome is shown in Figure 6.6 with the respective Delta scores stored in Table 6.4.



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.6: Objective evaluation scores for different versions of DS-Wiener averaged over 30 speakers of NOIZEUS database, white noise: (a) PESQ, (b) SegSNR, (c) STOI

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
DS-f_0	0.50	0.60	0.59	4.15	3.91	3.84	0.02	0.02	0.00
PEFAC	0.42	0.47	0.42	4.49	4.38	4.47	0.00	0.00	-0.02
f_0-oracle	0.55	0.63	0.60	4.07	3.86	3.85	0.03	0.02	0.01

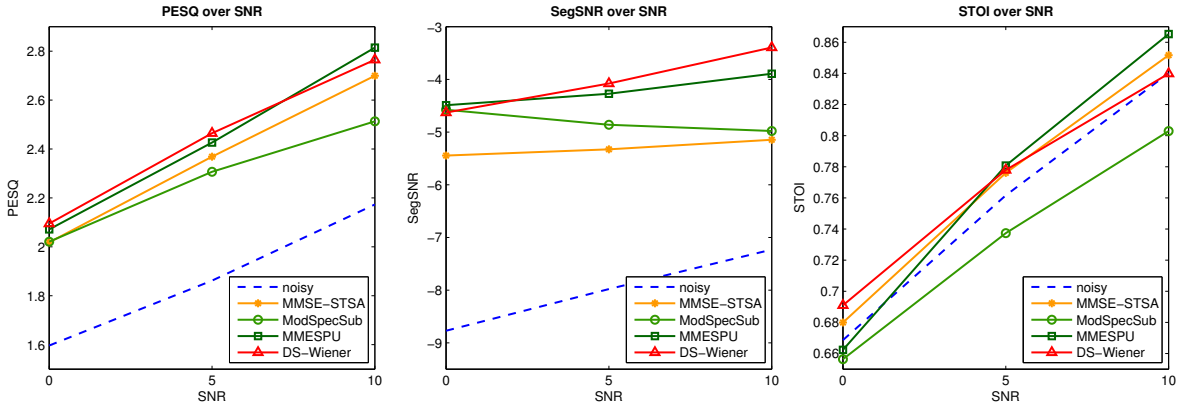
Table 6.4: Delta scores for different versions of DS-Wiener averaged over 30 speakers of NOIZEUS database, white noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

It can be observed, that the version of DS-Wiener implemented by using DS-f_0 clearly outperforms its PEFAC counterpart in terms of PESQ. What is more, the perceived intelligibility (STOI) could only be improved by our proposed DS-Wiener implementation. Interestingly, the DS-Wiener based on PEFAC beats the DS-f_0-based version in both blind and f_0-oracle scenarios. Since we are more interested in PESQ and STOI, we conclude that DS-f_0 is preferable when used in a speech enhancement application. Perceptually, we notice slightly more speech distortion and residual noise for stimuli produced by the PEFAC implementation. This possibly indicates that TBS performed with DS-f_0 yields a more compact DS representation suited for noise suppression.

In the following Sections, we select the DS-Wiener as the most effective of the described DS-based speech enhancement methods and compare it to prominent benchmark methods. For this purpose we will use utterances from both the NOIZEUS and the TIMIT database.

6.2.2 Comparison to Benchmark Methods - NOIZEUS Database

In this experiment we conduct a performance analysis in terms PESQ, SegSNR and STOI in blind scenarios only. As an STFT-based benchmark method we select the classical MMSE-STSA by Ephraim and Malah (1984) [46] with a Minimum Statistics noise estimator [61], as described in Section 6.1.2. The STSM-based benchmarks, ModSpecSub and MMESPU [4, 25] are algorithms operating in the modulation domain. In contrast to DS-based algorithms, these two methods employ STFT analysis with fixed time blocks, i.e. no pitch-synchronous framing. In Figure 6.7 we present the evaluation results for the white noise scenario. The corresponding Delta values are reported in Table 6.5.



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.7: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, white noise: (a) PESQ, (b) SegSNR, (c) STOI

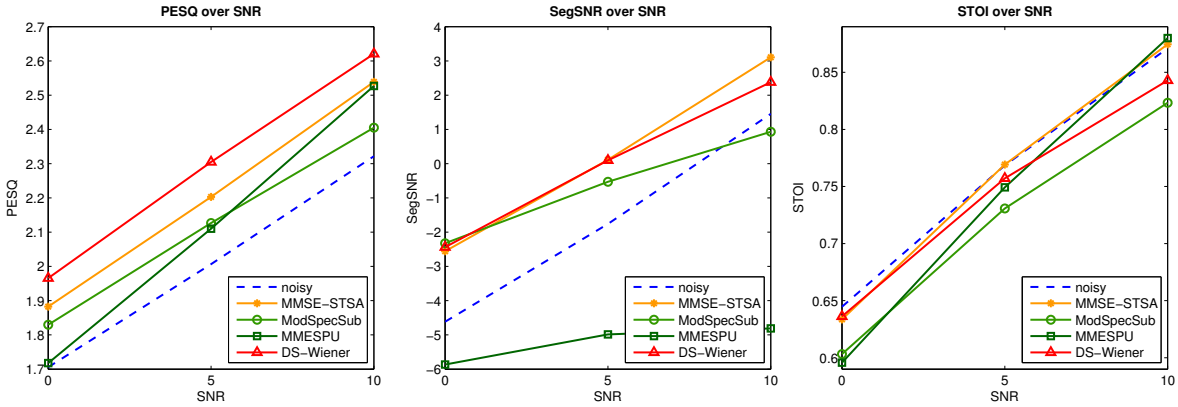
	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.42	0.51	0.53	3.33	2.65	2.09	0.01	0.01	0.01
ModSpecSub	0.43	0.45	0.34	4.20	3.12	2.26	-0.01	-0.02	-0.04
MMESPU	0.48	0.56	0.64	4.28	3.71	3.35	-0.01	0.02	0.03
DS-Wiener	0.50	0.60	0.59	4.15	3.91	3.84	0.02	0.02	0.00

Table 6.5: Delta scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, white noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

We observe that DS-Wiener and MMESPU perform best in terms of improving speech quality of a noisy signal corrupted by white noise. A joint improvement of quality and intelligibility is achieved by DS-Wiener and MMSE-STSA over the selected SNR range. For both STSM-based methods STOI is degraded at low SNR, however MMESPU effectively enhances the intelligibility at medium and high SNR levels. Perceptually, the DS-Wiener produces naturally sounding stimuli with little residual noise and no musical noise artifacts. Besides having a higher residual noise floor, ModSpecSub introduces slurring which degrades the intelligibility. MMESPU finds a better trade-off for these issues and removes noise to a great extent. However, this method often produces musical noise artifacts. Stimuli of MMSE-STSA feature relatively clear and natural speech, but are superimposed by harsh sounding residual noise.

We now test the performance of DS-Wiener and benchmark methods for babble noise. The results are shown in Figure 6.8. The improvement compared to the noisy signal is quantified as Delta values reported in Table 6.6. In this scenario, the DS-Wiener outperforms its STFT and STSM-based counterparts in terms of PESQ. For SegSNR DS-Wiener and MMSE-STSA show similar results, with slight advantages for MMSE-STSA at higher SNR levels. Interestingly, MMESPU degrades the SegSNR significantly with regard to the noisy signal. This may be due to severe noise estimation errors which cause this algorithm to filter out entire speech segments occasionally. Intelligibility in terms of STOI is best preserved by MMSE-STSA.

Subjectively, DS-Wiener and MMSE-STSA introduce least speech distortion in presence of babble noise. ModSpecSub is prone to noise overestimation and thus filters out relevant speech information.



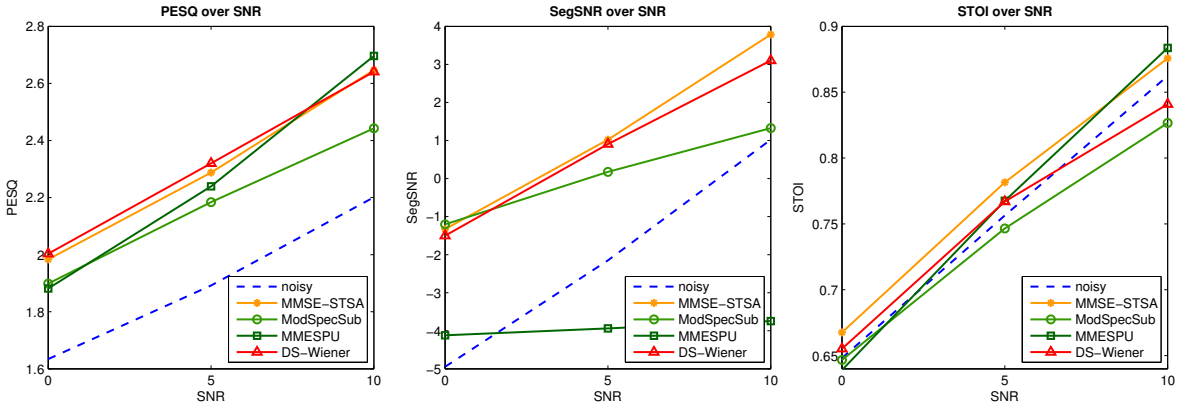
(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.8: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, babble noise: (a) PESQ, (b) SegSNR, (c) STOI

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.18	0.20	0.22	2.06	1.87	1.65	-0.01	0.00	0.00
ModSpecSub	0.12	0.12	0.08	2.31	1.25	-0.49	-0.04	-0.04	-0.05
MMESPU	0.01	0.10	0.21	-1.23	-3.20	-6.23	-0.05	-0.02	0.01
DS-Wiener	0.26	0.30	0.39	2.19	1.88	0.96	-0.01	-0.01	-0.03

Table 6.6: Delta scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, babble noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

Finally, we show performance evaluation results for car noise. In Figure 6.9 we report objective evaluation scores and in Table 6.7 the respective Delta values.



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.9: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, car noise: (a) PESQ, (b) SegSNR, (c) STOI

Again, DS-Wiener performs well in all categories. Joint PESQ and STOI improvement is achieved at low and medium SNR levels. Similar to the babble noise experiment, MMESPU

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.35	0.40	0.45	3.62	3.17	2.76	0.02	0.03	0.01
ModSpecSub	0.26	0.29	0.24	3.76	2.35	0.34	0.00	-0.01	-0.04
MMESPU	0.25	0.35	0.50	0.50	-2.06	-5.20	-0.01	0.01	0.02
DS-Wiener	0.37	0.43	0.44	3.46	3.08	2.11	0.01	0.01	-0.02

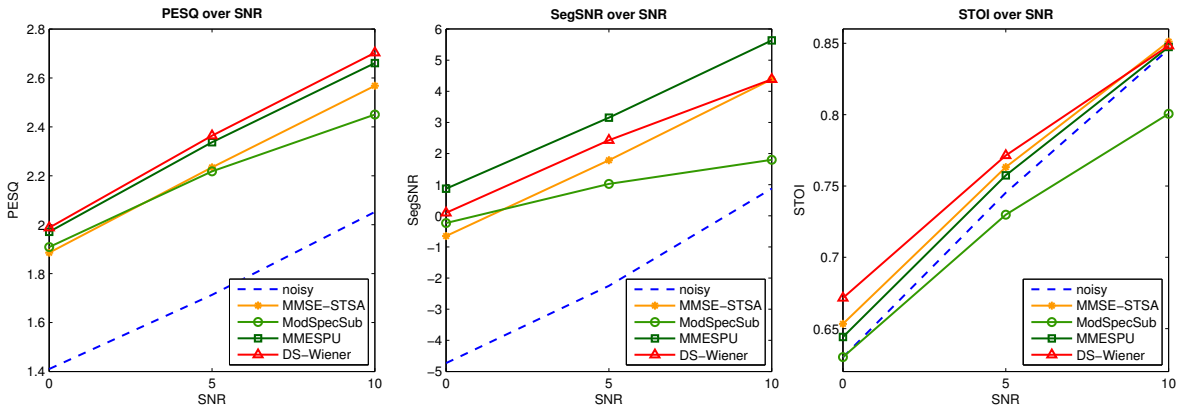
Table 6.7: Delta scores for DS-Wiener and selected benchmark methods averaged over 30 speakers of NOIZEUS database, car noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

degrades the perceived speech quality in terms of SegSNR, and MMSE-STSA yields the highest scores in terms of STOI. The stimuli generated by DS-Wiener show consistent and promising results. The subjective quality of MMESPU is only satisfactory, as long as the algorithm managed to preserve speech components through accurate noise estimation.

6.2.3 Comparison to Benchmark Methods - TIMIT Database

To cross-validate the results obtained with the NOIZEUS speech corpus, we run the same experiments on 18 utterances of the TIMIT corpus. Sentences taken from this database vary in length and speaking rate, and show more diversity in articulation and pronunciation. It is of great importance, that a speech enhancement method relying on pitch-synchronous time blocks operates robustly independent from language or dialect. For this reason, we want to investigate the performance of DS-Wiener under such conditions. In the following we compare DS-Wiener to the same benchmark methods as used in the previous Section for white, babble and factory noise. The latter represents another type of quasi-stationary noise in combination with occasional non-stationary components.

Figure 6.10 shows the evaluation results of the white noise scnerio. Table 6.8 presents the respective Delta scores of PESQ, SegSNR and STOI.



(a) PESQ *(b) SegSNR* *(c) STOI*

Figure 6.10: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, white noise: (a) PESQ, (b) SegSNR, (c) STOI

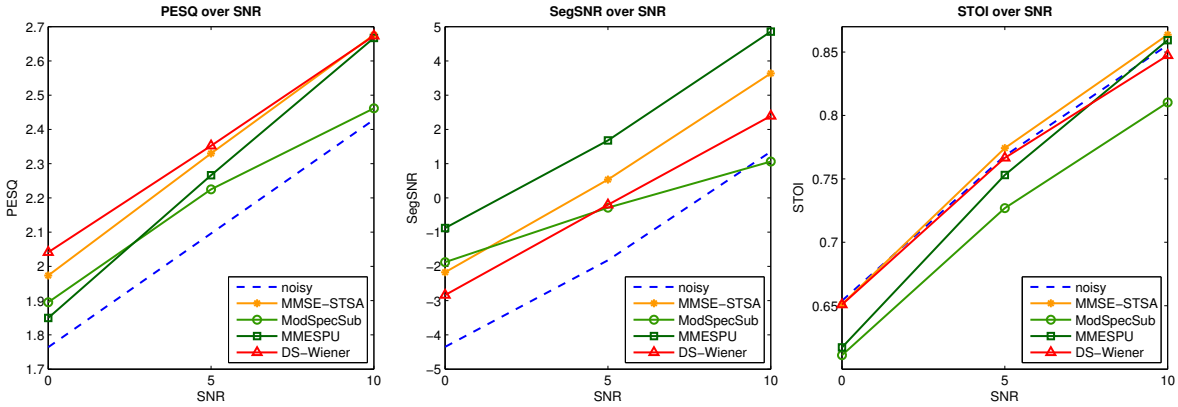
For white noise, it is confirmed that DS-Wiener also performs well on noisy speech signals of a different speech corpus. In contrast to the experiment using NOIZEUS database MMESPU

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.48	0.52	0.52	4.08	4.03	3.51	0.02	0.02	0.01
ModSpecSub	0.50	0.50	0.40	4.50	3.28	0.94	0.00	-0.02	-0.05
MMESPU	0.56	0.62	0.61	5.61	5.41	4.78	0.01	0.01	0.00
DS-Wiener	0.58	0.65	0.65	4.83	4.69	3.52	0.04	0.03	0.00

Table 6.8: Delta scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, white noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

outperforms the other methods in terms of SegSNR. However, intelligibility is most effectively improved by DS-Wiener, in particular at low SNR.

In Figure 6.11 the performance evaluation scores are shown for babble noise. Table 6.9 shows improvement in terms of Delta scores.



(a) PESQ *(b) SegSNR* *(c) STOI*

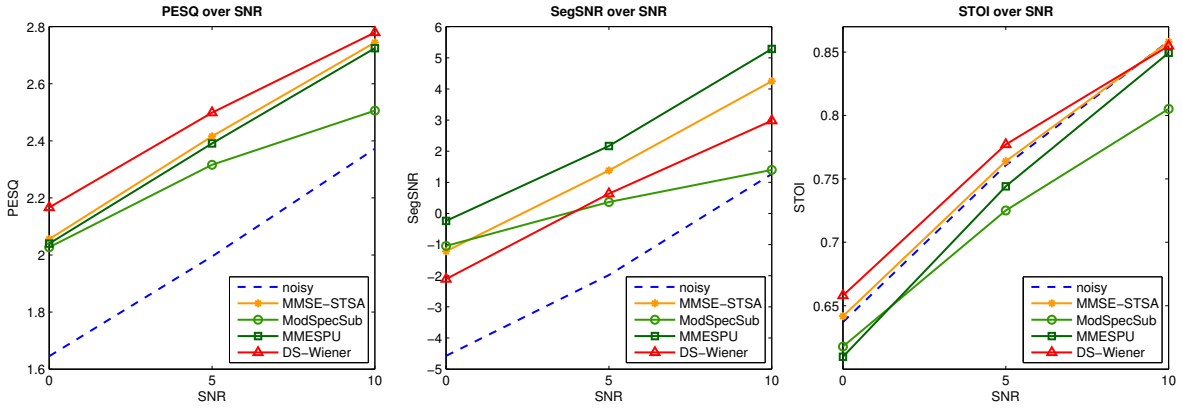
Figure 6.11: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, babble noise: (a) PESQ, (b) SegSNR, (c) STOI

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.21	0.23	0.25	2.18	2.36	2.29	0.00	0.01	0.01
ModSpecSub	0.13	0.13	0.03	2.48	1.54	-0.28	-0.04	-0.04	-0.05
MMESPU	0.09	0.17	0.24	3.48	3.51	3.52	-0.04	-0.02	0.00
DS-Wiener	0.28	0.26	0.24	1.52	1.63	1.06	0.00	0.00	-0.01

Table 6.9: Delta scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, babble noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

DS-Wiener shows consistent enhancement for babble noise. The PESQ improvement is similar as with MMSE-STSA, however MMESPU and MMSE-STSA outperform DS-Wiener in SegSNR. Regarding intelligibility scores, DS-Wiener performs better than its STSM-based counterparts.

In Figure 6.12 we present the evaluation results for factory noise. Table 6.10 shows the Delta scores of PESQ, SegSNR and STOI.



(a) PESQ *(b) SegSNR* *(c) STOI*

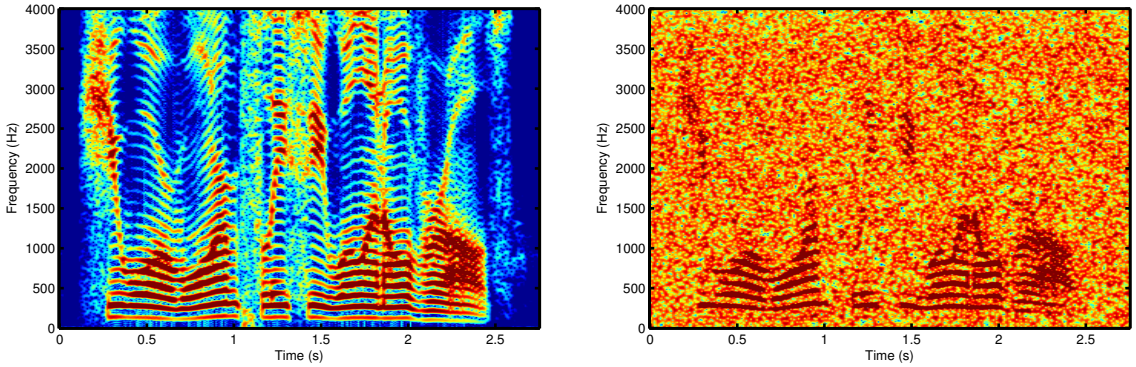
Figure 6.12: Objective evaluation scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, factory noise: (a) PESQ, (b) SegSNR, (c) STOI

	ΔPESQ			ΔSegSNR			ΔSTOI		
SNR level (dB)	0	5	10	0	5	10	0	5	10
MMSE-STSA	0.41	0.42	0.37	3.36	3.36	2.98	0.00	0.00	0.00
ModSpecSub	0.38	0.32	0.13	3.54	2.36	0.15	-0.02	-0.04	-0.05
MMESPU	0.39	0.40	0.35	4.34	4.15	4.03	-0.03	-0.02	-0.01
DS-Wiener	0.52	0.50	0.41	2.48	2.62	1.73	0.02	0.02	0.00

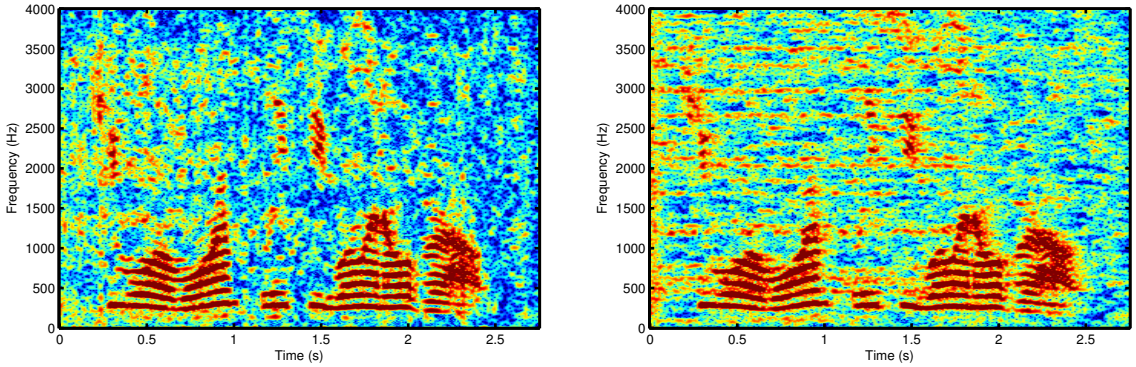
Table 6.10: Delta scores for DS-Wiener and selected benchmark methods averaged over 18 utterances of TIMIT database, factory noise: (Left) ΔPESQ, (Middle) ΔSegSNR, (Right) ΔSTOI.

For factory noise, we notice a similar trend compared to the babble noise scenario. DS-Wiener outperforms all benchmark methods in PESQ and STOI, but shows less speech quality improvement than MMESPU and MMSE-STSA in terms of SegSNR.

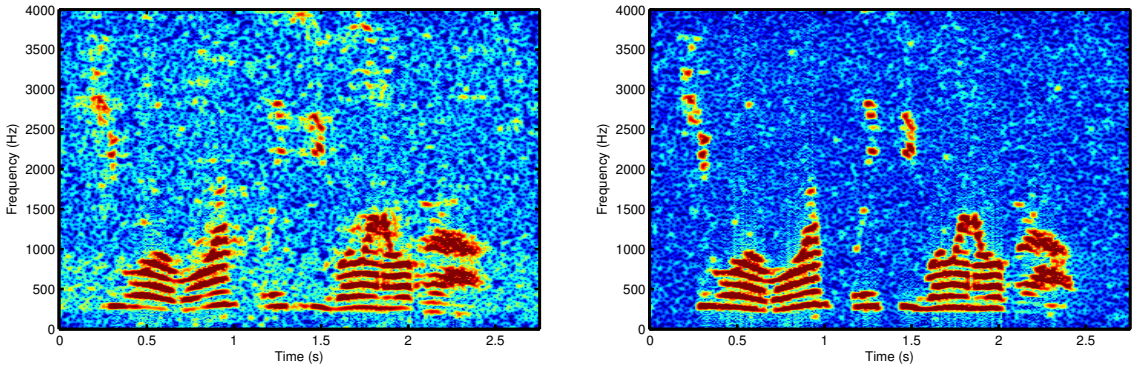
To conclude this Section about speech enhancement experiments and evaluation results, we present spectrograms of clean, noisy and enhanced signals as a visualization of the speech enhancement process. For this purpose, we selected a sentence of a male and a female speaker from the TIMIT test set used. The spectrograms are shown in Figures 6.13 and 6.14 for male and female speaker, respectively. The corresponding Delta scores for PESQ, SegSNR and STOI are reported in Tables 6.11 and 6.12, respectively.



(a) clean *(b) noisy*



(c) DS-Wiener *(d) MMSE-STSA*

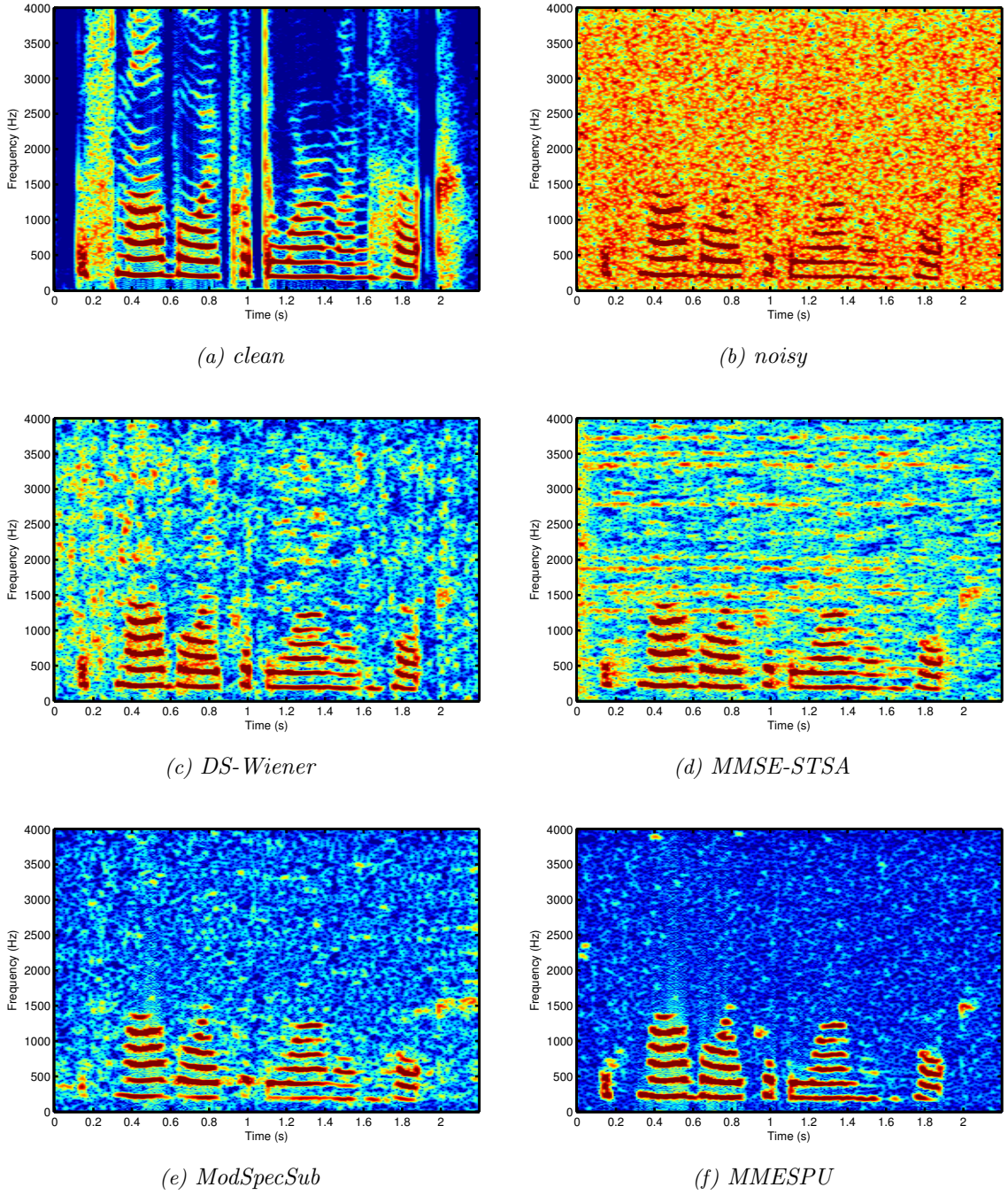


(e) ModSpecSub *(f) MMESPU*

Figure 6.13: Spectrograms of clean, noisy (white noise, 5 dB) and enhanced speech signals of the TIMIT sentence 'sx20' ("She wore warm, fleecy, woolen overalls."), uttered by a male speaker.

	ΔPESQ	ΔSegSNR	ΔSTOI
MMSE-STSA	0.54	3.92	0.01
ModSpecSub	0.23	2.64	-0.08
MMESPU	0.62	4.22	-0.02
DS-Wiener	0.70	4.49	0.03

Table 6.11: Delta scores for enhanced signals of the TIMIT sentence 'sx20'.



(a) clean *(b) noisy*

(c) DS-Wiener *(d) MMSE-STSA*

(e) ModSpecSub *(f) MMESPU*

Figure 6.14: Spectrograms of clean, noisy (white noise, 5 dB) and enhanced speech signals of the TIMIT sentence 'sx216' ("The small boy put the worm on the hook."), uttered by a female speaker.

	ΔPESQ	ΔSegSNR	ΔSTOI
MMSE-STSA	0.51	3.14	-0.01
ModSpecSub	0.15	1.91	-0.05
MMESPU	0.38	4.12	0.00
DS-Wiener	0.61	3.56	0.02

Table 6.12: Delta scores for enhanced signals of the TIMIT sentence 'sx216'.

6.2.4 Potentials and Limits

It has been shown, that DS-Wiener is an effective speech enhancement algorithm which can keep up with conventional state-of-the-art benchmark methods. Our proposed method was able to outperform its STFT and STSM-based counterparts in terms of objective speech evaluation scores in different scenarios. The potential of DS-based speech enhancement methods lies in the joint modification of acoustic frequency and modulation frequency. In addition, the pitch-synchronous nature of the DS analysis results in a harmonic filter bank structure, which tends to be beneficial for speech processing and modification.

The experiments conducted in the f_0-oracle scenario indicate that our proposed f_0 estimator, DS-f_0, performs well in the blind scenario. Besides, the results indicate that a more accurate and robust pitch estimator could further improve the achievable performance by the proposed DS-based speech enhancement methods. In practice, the performance of f_0 estimators is limited due to the impact of noise. That being said, we conclude that f_0-oracle will always be the upper bound for a noise suppression performance of DS-based speech enhancement methods.

The noise estimation method used in the implementation of DS-Wiener is not optimal. We believe, that a more sophisticated approach could boost the performance and robustness of our proposed method. For example, a noise estimator similar to *Minimum Statistics* [61] or *Improved Minima Controlled Recursive Averaging* [66] operating in the DS domain could be a promising alternative.

For our DS-based speech enhancement methods we did not use any statistical information about the distribution of DS coefficients of noisy or noise signals. To derive a statistical-based noise suppression rule, efforts have to be made for measuring the probability density function of DS coefficients. Assuming statistical knowledge, an estimator may be derived which is optimal in some statistical sense, e.g. MMSE. Early studies suggested a Gaussian distribution to describe (two-dimensional) DCT coefficients [37,38], whereas in more recent works a Laplacian distribution is said to be more appropriate [39,40]. In the latter case, modifications of the derivation of an MMSE amplitude estimator similar to [67–69] could be of interest.

By operating at a time block step size of $\tilde{P}_0$, i.e. an overlap of $(Q-1)\tilde{P}_0$, we impose a constraint on the real-time applicability of DS-based methods. Assuming $Q = 4$ as modulation band number and a speech signal of duration D, this introduces a worst case delay of

$$\tau_{max} \cong \frac{Q-1}{Q} \cdot D, \tag{6.7}$$

which corresponds to 75% of the signal length. Clearly, in this case only an offline implementation is feasible. Using a time block overlap of only $\tilde{P}_0$, we are able to reduce the processing delay to approximately 25% of the signal duration, but we may encounter the overlap mismatch issue addressed in Section 3.5.4. For this reason, extended research on how to improve the real-time implementation of DS-based systems is needed.

7

Conclusion and Future Outlook

In this thesis, we introduced Double Spectrum (DS) as a novel domain for the design of single-channel speech enhancement methods. Given the importance of temporal modulation in speech signals, it seems plausible to use speech representations which take into account not only spectral components, but also their modulation over time. This led to speech representations such as the STSM, where a two dimensional STFT analysis is proposed to describe the acoustic frequency and the modulation frequency components of speech. In studies by Kleijn [27,28] and Nilsson [16, 17], a canonical approach to represent speech was proposed by applying a two-stage transform: A pitch-synchronous transform to capture the harmonic nature of human speech, and a modulation transform to arrive at a compact representation describing how spectral components evolve over time. Applying this two-stage transform directly to speech signal segments in a block-wise fashion is the fundamental idea of DS.

In the design of DS-based speech enhancement methods we described the essential steps, namely pitch estimation, pitch-synchronous time block segmentation (TBS), DS analysis-synthesis, speech presence probability (SPP) estimation and noise suppression rules derived in the DS domain. In the course of this work an own f_0 estimator, called DS-f_0, was derived. The DS-f_0 proved to be more effective for TBS needed in DS analysis than conventional pitch estimators. This was verified by experiments on voiced-unvoiced classification (Section 4.2.1) and noise reduction performance (Section 6.2.1). Motivated by f_0 estimation using DS, also an SPP estimator was implemented in the DS domain, which was later used for different noise suppression rules of the proposed speech enhancement system.

By implementing and comparing different DS-based speech enhancement algorithms, we found Wiener filtering in the DS domain as most effective. The proposed method, DS-Wiener, was compared against popular speech enhancement benchmark methods. For a fair comparison, we included both an STFT-based method [46] and STSM-based methods [4,25] in our performance evaluation. The results obtained in our experiments demonstrate the effectiveness of DS-Wiener in reducing the noise and improving both perceived quality and intelligibility of speech. This was shown for stationary and non-stationary noise types at different SNR levels and was statistically validated by means of two different sets of speech utterances.

Future works may include the design of robust pitch estimators in the DS domain, which facilitate accurate TBS. Regarding DS-based speech enhancement methods, we see some room for improvement of noise estimation in DS and in the derivation of some statistically-based noise suppression rules, such as a DS-MMSE or DS-MAP estimator, which requires some prior estimates of the underlying probability density function assumed for DS magnitude. Other fields of application could be the speech quality estimation of speech enhancement methods by modulation-sensitive measures defined in DS [70] or single-channel source separation algorithms similar to [22]. Clearly, the full potential of the DS domain is yet to be explored.

A
Appendix

List of Abbreviations

ADSW	Adaptive Double Spectrum Weighting
AM	amplitude modulation
AMS	analysis-modification-synthesis
AR	autoregressive
ASR	automatic speech recognition
AWGN	additive white Gaussian noise
dB	decibel
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DS	Double Spectrum
DS-f_0	Double Spectrum Pitch Estimator
DS-PE-adp	Double Spectrum Periodicity Enhancement using Adaptive Weighting
DS-PE-fxd	Double Spectrum Periodicity Enhancement using Fixed Weighting
DS-SPP	Double Spectrum Speech Presence Probability Estimator
DS-Wiener	Double Spectrum Wiener Filter
DTFT	Discrete-Time Fourier Transform
KLT	Karhunen-Loève Transform
LP	Linear Prediction
LPC	Linear Prediction Coding
MBR	Modulation Band Ratio
MLT	modulated lapped transform
MMESPU	Minimum Mean-Square Error Short-Time Spectral Modulation Magnitude Estimator with Speech Presence Uncertainty
MMSE	Minimum Mean-Square Error
MMSE-STSA	Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator
ModSpecSub	Spectral Subtraction in the Short-Time Modulation Domain
MSE	mean-square error
OLA	overlap-and-add

PE	periodicity enhancement
PESQ	Perceptual Evaluation of Speech Quality
PSD	power spectral density
SegSNR	Segmental SNR
SNR	signal-to-noise ratio
SPP	speech presence probability
SPU	speech presence uncertainty
STFT	Short-Time Fourier Transform
STOI	Short-Time Objective Intelligibility
STSA	Short-Time Spectral Amplitude
STSM	Short-Time Spectral Modulation
TBS	time block segmentation
VAD	voice activity detector
VUV	voiced-unvoiced

List of Symbols

f_0	fundamental frequency
f_s	sampling frequency
f_q	modulation frequency
P_0	pitch period
$\tilde{P}_0$	normalized pitch period
n	time sample index
ℓ	time frame index
l	time block index
k	frequency channel index
q	modulation band index
$x(n)$	clean signal
$\hat{x}(n)$	enhanced signal, estimate of $x(n)$
$d(n)$	noise signal
$y(n)$	noisy signal
DS_x	clean Double Spectrum
$\widehat{DS}_x$	enhanced Double Spectrum, estimate of DS_x
DS_d	noise Double Spectrum
$\widehat{DS}_d$	noise Double Spectrum, estimate of DS_n
DS_y	noisy Double Spectrum
γ	a posteriori SNR
ξ	a priori SNR
α	forgetting factor used in the recursive averaging of the prior SNR
λ	smoothing constant used for recursive averaging of the noise estimate

A.1 Comparison of Double-Spectrum-Based Methods in Blind and Oracle Scenarios

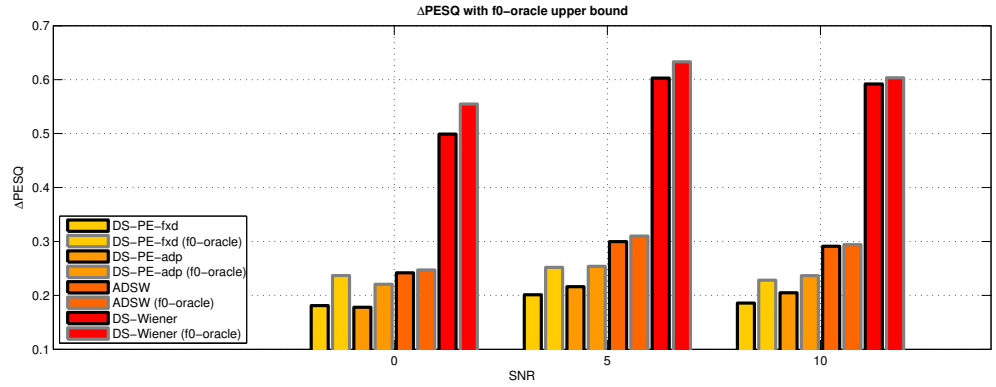


Figure A.1: Comparison of $\Delta PESQ$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, white noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

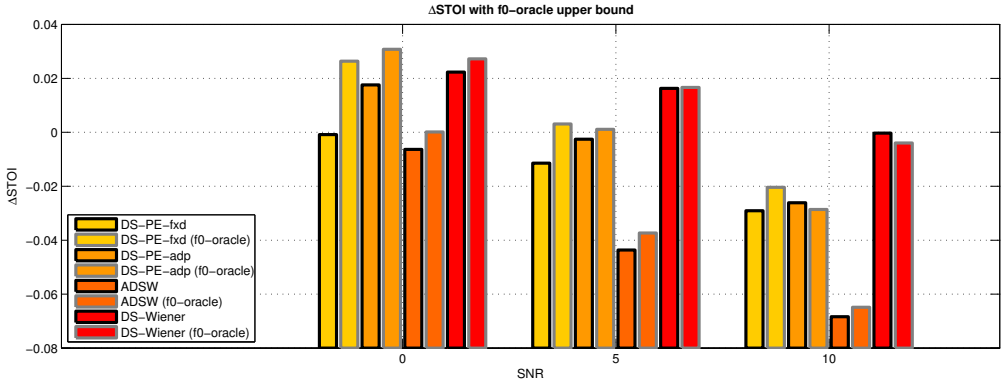


Figure A.2: Comparison of $\Delta STOI$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, white noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

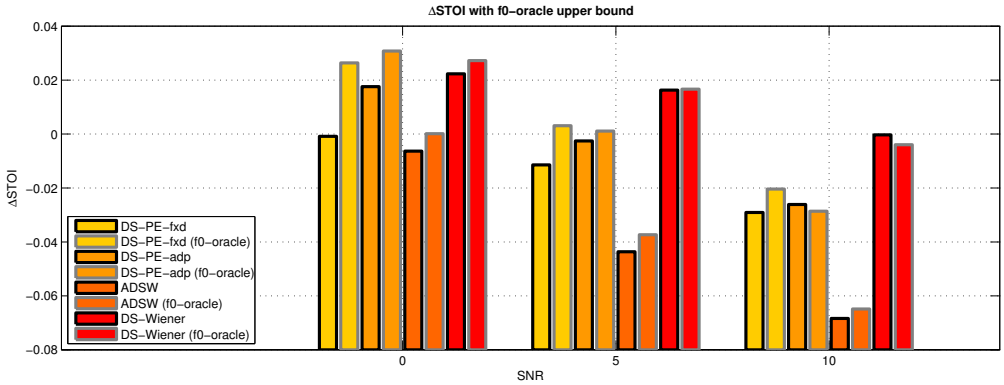


Figure A.3: Comparison of $\Delta SegSNR$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, white noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

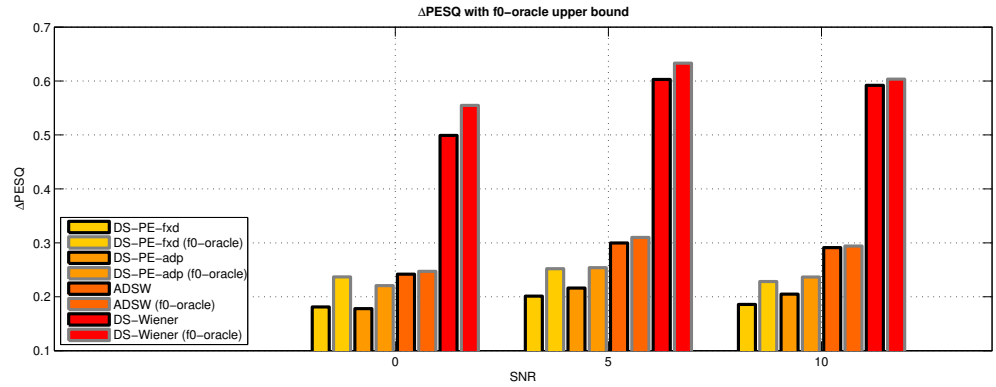


Figure A.4: Comparison of $\Delta PESQ$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, babble noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

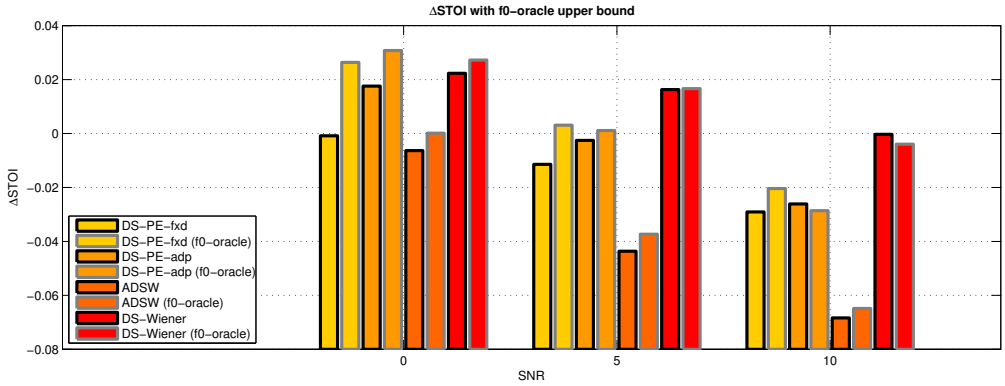


Figure A.5: Comparison of $\Delta STOI$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, babble noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

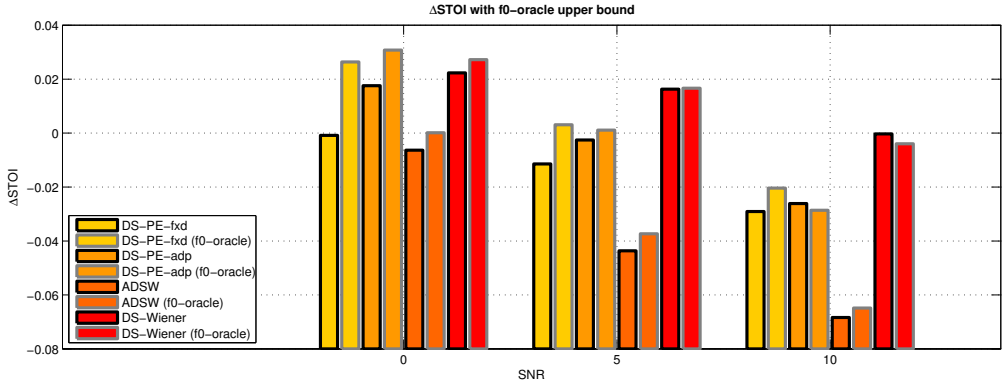


Figure A.6: Comparison of $\Delta SegSNR$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, babble noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

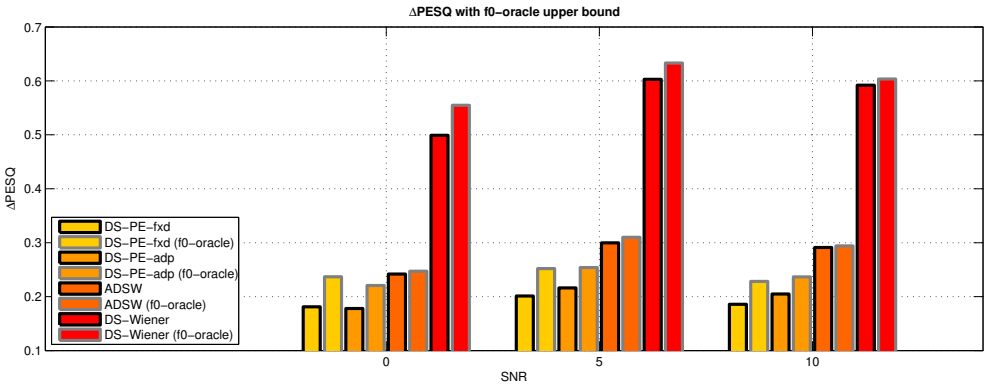


Figure A.7: Comparison of $\Delta PESQ$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, car noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

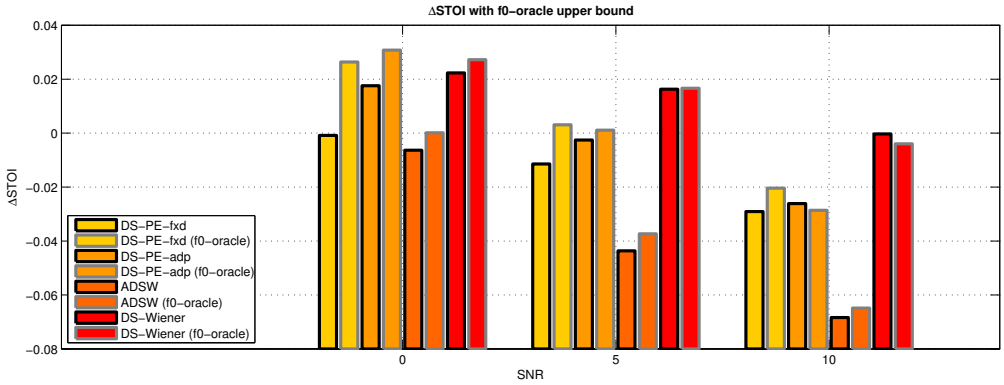


Figure A.8: Comparison of $\Delta STOI$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, car noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

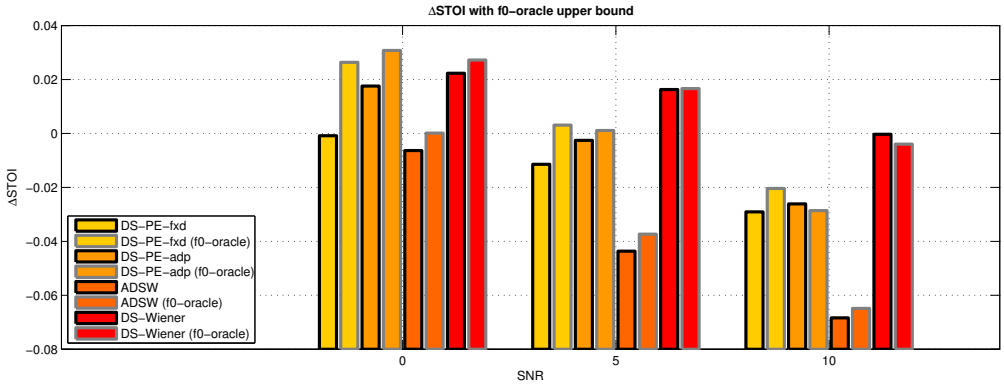


Figure A.9: Comparison of $\Delta SegSNR$ results of DS-based methods, obtained in both blind and f_0-oracle scenario, car noise. For each method, the right bar (grey frame) represents the f_0-oracle result.

A.2 List of TIMIT Sentences used in the Performance Evaluation

Male speaker

- The so-called vegetable ivory is the hard endosperm of the egg-sized seed. (si1010)
- How did one join them? (si1640)
- Dogs did something to one's ego. (si2270)
- The best way to learn is to solve extra problems. (sx110)
- She wore warm, fleecy, woolen overalls. (sx20)
- Ralph controlled the stopwatch from the bleachers. (sx200)
- This brochure is particularly informative for a prospective buyer. (sx290)
- Why charge money for such garbage? (sx380)

Female speaker

- She had your dark suit in greasy wash water all year. (sa1)
- Don't ask me to carry an oily rag like that. (sa2)
- In wage negotiations, the industry bargains as a unit with a single union. (si1386)
- Heave on those ropes; the boat's come unstuck. (si2016)
- Materials: ceramic modeling clay: red, white or buff. (si756)
- Artificial intelligence is for real. (sx126)
- The small boy put the worm on the hook. (sx216)
- A chosen few will become Generals. (sx306)
- Only the most accomplished artists obtain popularity. (sx36)
- The fish began to leap frantically on the surface of the small lake. (sx396)

Taken from "Documentation for TIMIT" [71]

A.3 Single-Channel Speech Enhancement Using Double Spectrum (Interspeech 2016 Conference Paper - under revision)

Single-Channel Speech Enhancement Using Double Spectrum

Martin Blass[†]*, Pejman Mowlaee*[†] *, W. Bastiaan Kleijn*[‡]

[†]Signal Processing and Speech Communication Lab, Graz University of Technology
[‡]School of Engineering and Computer Science, Victoria University of Wellington, New Zealand
mblass@student.tugraz.at pejman.mowlaee@tugraz.at bastiaan.kleijn@ecs.vuw.ac.nz

Abstract

Single-channel speech enhancement is often formulated in the *Short-Time Fourier Transform* (STFT) domain. As an alternative, several previous studies have reported advantages of speech processing using pitch-synchronous analysis and filtering in the modulation transform domain. We propose to use the *Double Spectrum* (DS) obtained by combining pitch-synchronous transform followed by modulation transform. The linearity and sparseness properties of DS domain are beneficial for single-channel speech enhancement. The effectiveness of the proposed DS-based speech enhancement is demonstrated by comparing it with STFT-based and modulation-based benchmarks. In contrast to the benchmark methods, the proposed method does not exploit any statistical information nor does it use temporal smoothing. The proposed method leads to an improvement of 0.3 PESQ on average for babble noise.
Index Terms: speech enhancement, double spectrum, modulation transform, pitch-synchronous analysis

1. Introduction

In various speech processing applications including speech coding, automatic speech recognition and speech synthesis the underlying signal representation determines the accuracy and efficiency of a certain algorithm. Good representations often require relatively few coefficients per unit time for an accurate description of the speech signal, but are complete and hence able to describe any signal. We argue that the *Short-Time Fourier Transform* (STFT), the predominant choice in speech enhancement (see e.g. [1] for an overview), while complete, generally does not lead to a sparse signal representation for speech.

An alternative to the STFT domain is pitch-synchronous analysis, with successful results reported both for speech coding [2, 3] and speech enhancement [4]. It was shown that frame theory can be used to understand this representation [3].

Another alternative is to process speech in the *Short-Time Modulation* (STM) domain. Speech enhancement proposals in modulation domain are spectral subtraction [5], *Minimum Mean Square Error* (MMSE) of *Short-Time Modulation Magnitude* (STMM) Spectrum [6], MMSE speech enhancement using real and imaginary parts of STM [7]. These STM-based methods, compared to their STFT counterparts, showed less musical noise or spectral distortion with improved perceived quality.

Inspired by the advantages of modulation and pitch-synchronous transforms, a key research question is then how to exploit these in a speech enhancement framework. In this paper, therefore, we propose *Double Spectrum* (DS) signal representation consisting of pitch-synchronous and modulation transforms. We propose single-channel speech enhancement in DS domain. To demonstrate the potentials and advantages of the proposed method, we compare its performance versus the previous STFT-based and modulation-based benchmarks.

The remainder of the paper is organized as follows; Section 2 places our work in the context of earlier work. In Section 3 we provide fundamentals of the *Double Spectrum* (DS) approach. Section 4 presents the proposed Double Spectrum speech enhancement, Section 5 shows the results and Section 6 provides conclusions.

The work was supported by Austrian Science Fund (P28070-N33).

2. Relation to Previous Works

Separating slowly varying and rapidly varying pitch-cycle waveform components formed the basis of *Waveform Interpolation* (WI), which resulted in high quality speech coding [2]. A more general pitch-synchronous modulation representation was introduced in [3]. This two-stage transform representation was further refined by Nilsson et al. [8]. The two-stage transform led to a solid performance in speech coding and prosodic modification. In such speech representation the fundamental frequency is the key feature resulting in a sparse speech-signal representation. The block diagram for the two-stage transform representation, shown in Figure 1, consists of four processing blocks: *Linear Prediction* (LP) analysis, constant pitch warping, pitch-synchronous transform and modulation transform.

The two-stage transform, consisting of pitch-synchronous and modulation transforms exploits the features of the warped residual to achieve a highly energy concentrated representation and will be described in more detail in Section 3.2. The combination of pitch-synchronous and modulation transform results in lapped frequency transforms, which approximates the *Karhunen-Loève Transform* (KLT) for stationary signal segments [9]. The KLT maximizes the coding gain, which can be seen as a particular form of energy concentration [8].

The two-stage transform was extended to speech enhancement [4], where its ability to separate periodic and aperiodic signals were exploited to improve speech quality. Noise reduction was achieved by adaptive weighting of the coefficients in different modulation bands, which restored harmonicity of noise corrupted speech. The method was capable of separating the speech signal into voiced and unvoiced components using a best-basis selection that optimized the energy concentration of the transform coefficients.

Throughout this paper, the signal representation obtained by two-stage transform (pitch-synchronous and modulation transform) will be referred to as *Double Spectrum* (DS). Figure 1 shows the DS framework highlighted in a light gray block as the basis of the proposed speech enhancement system. Our goal is to find a framework where the two-stage transform is directly applied on the noisy signal. In contrast to [4, 8], our proposed method relies on fixed analysis time blocks (no LP

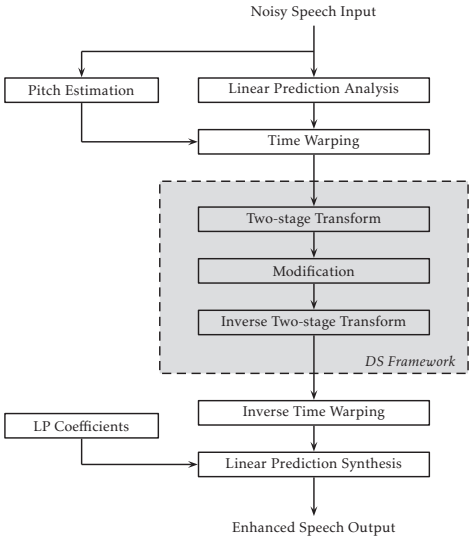


Figure 1: Block diagram for a canonical speech representation system [8]. The highlighted block shows DS framework using a two-stage transform and signal modification in DS domain.

analysis, nor time warping), which makes the method simpler and faster.

3. Double Spectrum: Fundamentals

First, the pitch is extracted and stored within the coefficients of the two-stage transform. Since pitch is time-varying and both transforms do not adapt to this property, we introduce block processing under the assumption of quasi-stationarity of speech, explained in the following.

3.1. Time Block Segmentation

Given a fundamental frequency f_0, the first step in calculating DS is pitch-synchronous *Time Block Segmentation* (TBS). The TBS step separates the input speech into L time blocks of variable length. The length of each time block is an integer multiple of $P_0 = f_s/f_0$, where f_s is the sampling frequency and P_0 is the fundamental period in samples. A time block is further subdivided into frames, each of length P_0. To avoid discontinuities at the transition of consecutive blocks overlapping is introduced.

3.2. Two-stage Transform

Each time block is analyzed in terms of a two-stage transform. The pitch-synchronous transform is implemented as a *Modulated Lapped Transform* (MLT) [9]. Since pitch varies over time, this means that we ignore its local variation of pitch during TBS. The MLT is implemented using a DCT-IV in combination with square-root Hann window following [8]. This facilitates a critically sampled uniform filter bank with coefficients that are localized in time and frequency. The usage of a square-root window at analysis and synthesis stage as a matched filter satisfies the power complementarity constraint needed for perfect reconstruction.

Let $\nu = 0, 1, \ldots, 2P_0 - 1$ be a time index and let $x_l(\nu)$ be the l'th pitch-synchronous time block, i.e. $x_l(\nu) = x(lP_0 + \nu)$. The first-stage transform coefficients $f(l, k)$ are then obtained as

$$f(l,k) = \sum_{\nu=0}^{2P_0-1} \tilde{x}_l(\nu)\sqrt{\frac{2}{P_0}}\cos\left(\frac{(2k+1)(2\nu - P_0 + 1)\pi}{4P_0}\right), \quad (1)$$

where $l = 0, 1, \ldots, L-1$ and $k = 0, 1, \ldots, P_0 - 1$ denote time block index and frequency band index, respectively, and $\tilde{x}_l(\nu) = x_l(\nu)w(\nu)$ as the windowed signal segment.

The output of the first transform is a sequence of MLT coefficients that evolve slowly over time for voiced speech but rapidly for unvoiced speech. Note that due to the pitch-synchronous nature of the time frames, the cardinality of the frequency bands is $K = P_0$.

The modulation transform is a DCT applied to a number of consecutive frames of the frequency coefficients obtained from pitch-synchronous transform [10]. To facilitate the implementation of the modulation transform as a critically sampled filter, we use DCT-II yielding the coefficients $g(q, k)$ given by

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

where $q = 0, 1, \ldots, Q-1$ is the modulation band index, $c(0) = 1/\sqrt{2}$ and $c(q) = 1$ for $q \neq 0$. The definition for *Double Spectrum* is now given by $DS(q, k)$, a matrix with K frequency bands as rows and Q modulation bands as columns. Figure 2 schematically visualizes a speech signal in terms of a sequence of Double Spectra, showing $DS^{(l)}(q, k)$ for a set of time blocks $l \in [0, L-1]$.

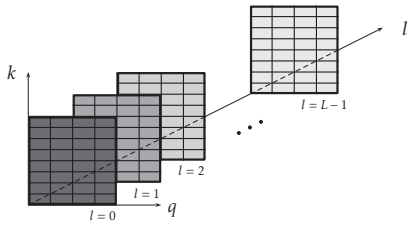


Figure 2: Illustration of a speech signal in Double Spectrum $DS^{(l)}(q, k)$ shown for time blocks $l = 0, 1, \ldots, L-1$.

3.3. Some Useful Properties of Double Spectrum

The useful properties of Double Spectrum are: sparsity, linearity, real-valued coefficients, and facilitates comb filtering.

3.3.1. Property I: Sparsity

For a periodic signal segment $DS(q, k)$ yields a high energy concentration at low modulation bands. In particular, the first modulation band $q = 0$ represents the periodic component of a signal, whereas the other modulation bands describe the aperiodic parts. This property can be explained by assuming a strictly periodic time signal, e.g., a pure sinusoid. Applying the pitch-synchronous transform yields MLT coefficients that are identical for consecutive frames. The subsequent modulation transform is hence applied to a constant data sequence, yielding only one non-zero coefficient for $q = 0$, which can be understood as the DC component of the DCT-II transform. This property may be exploited for voiced-unvoiced decomposition or for restoring the harmonicity of noise corrupted speech by finding an appropriate balance between low and high modulation bands [4].

3.3.2. Property II: Linearity

In the time domain, noisy signal $y(\nu)$ is a superposition of the clean signal $x(\nu)$ and the noise signal $d(\nu)$. In the DS domain this superposition is preserved, since DS is a linear operator:

$$y(\nu) = x(\nu) + d(\nu) \quad \circ\!\!-\!\!\bullet \quad DS_y = DS_x + DS_d, \tag{3}$$

where DS_y, DS_x and DS_d denote the DS representation of noisy, clean and noise signal, respectively. Figure 3 shows an example for DS_y, DS_x and DS_d of the same voiced speech segment to illustrate *linearity*.

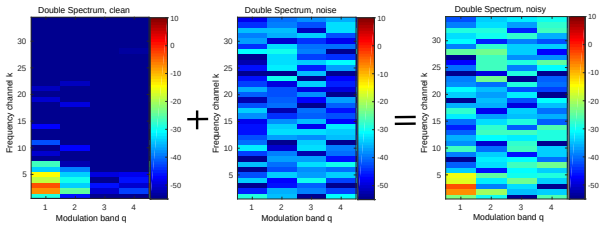


Figure 3: Linearity of DS operator given in (3): (Left) clean, (Middle) noise and (Right) noisy DS.

3.3.3. Property III: Real-Valued Coefficients

The coefficients of $DS(q,k)$ are real-valued and symmetrically distributed around zero as mean value.

3.3.4. Property IV: Facilitates Comb Filtering

Another property is the pitch-synchronous filter bank which allows comb filtering. Since an analysis frame of length of $2P_0$ yields $K = P_0$ frequency bands, $k_{f_0} = 2$ denotes the frequency band corresponding to f_0 and we have:

$$k_{f_0} = \frac{2K}{f_s} f_0. \tag{4}$$

4. Speech Enhancement in DS Domain

In this Section we present the essential tools for speech enhancement in DS domain comprised of pitch estimation, speech presence probability estimation, and the DS weighting function.

4.1. Pitch Estimation

The segmentation used in DS requires a fundamental frequency estimate. If the time blocks are segmented erroneously due to errors in pitch estimation, then the energy of periodic speech segments is no longer concentrated in the low modulation bands, but leaks into higher bands. We propose an f_0-estimator that relies on a periodicity measure calculated in the DS domain, called the *Modulation Band Ratio* (MBR). The MBR compares the summed energy of the first modulation band E_1 to the total energy $E_{1:Q}$

$$\text{MBR}(K) = \frac{E_1}{E_{1:Q}} = \frac{E_1}{E_1 + E_{2:Q}}\,, \tag{5}$$

where $E_1 = \sum_{k=0}^{K-1} |DS(0,k)|^2$ and $E_{1:Q} = \sum_{q=0}^{Q-1}\sum_{k=0}^{K-1} |DS(q,k)|^2$. For periodic frames the MBR reaches values close to 1, while for non-periodic frames the mean MBR is $1/Q$ (close to 0). This allows us to derive an f_0-estimator by searching for an optimal frequency index K^* that maximizes the MBR:

$$K^* = \arg\max_K \text{MBR(K)}. \tag{6}$$

Using (4), the fundamental frequency estimate is $f_0^* = \frac{f_s}{2K} K^*$.

4.2. Speech Presence Probability Estimation

Many common speech enhancement systems use information about the speech presence probability (SPP). In the design of our filter method we also take into account SPP to selectively modify regions of speech presence or absence. The SPP is computed in the DS domain using the MBR measure, which discriminates voiced and unvoiced speech even in heavy noise scenarios. MBR yields values close 1 for voiced and close to 0 for unvoiced, hence is a good measure for SPP.

4.3. Adaptive Weighting based on Energy Smoothing

Our proposed speech enhancement, referred to as *Double Spectrum Weighting* (DSW), is an adaptive weighting scheme corresponding to filtering in time domain. The weighting coefficients $G(q,k)$ are applied to the noisy coefficients $DS_y(q,k)$ and yield the clean speech estimate $\widehat{DS_x}(q,k)$:

$$\widehat{DS_x}(q,k) = G(q,k) DS_y(q,k), \tag{7}$$

where $G(q,k)$ is a cascade of two weighting schemes: $W_e(q,k)$ to dampen noise-dominant coefficients, and $W_q(q,k)$ to enhance harmonicity, each described in the following.

4.3.1. $W_e(q,k)$: Energy-based coefficient weighting

The first weighting, $W_e(q,k)$ is an energy based coefficient weighting $W_e(q,k)$ which compares the energy of each DS-coefficient with respect to the mean energy of $DS_y(q,k)$, resulting in the relative energy $E_{rel}(q,k)$ defined as

$$E_{rel}(q,k) = KQ\,\frac{|DS(q,k)|^2}{E_{1:Q}}. \tag{8}$$

Since E_{rel} shows a broad dynamic range, we apply the decadic logarithm as a non-linear mapping function. Additionally, we constrain the weights to 0 by adding 1 to E_{rel}:

$$W_e(q,k) = \log_{10}(E_{rel}(q,k) + 1). \tag{9}$$

4.3.2. $W_q(q,k)$: Harmonicity Enhancement

As the second weighting, we propose $W_q(q,k)$ to enhance the harmonicity of noisy speech. To this end, we need a harmonicity indicator. Similar to (5), we consider the Modulation Band Ratio of the respective frequency band, MBR_k given by

$$\text{MBR}_k = \frac{|DS(0,k)|^2}{\sum_{q=0}^{Q-1} |DS(q,k)|^2}. \tag{10}$$

In contrast to the fixed-weighting method in [4], we propose an exponentially decaying modulation weighting, which turned out to be a better choice in DS. Therefore, we use

$$W_q(q,k) = e^{-\text{MBR}_k q}, \tag{11}$$

where MBR_k serves as the decay factor of the exponential weighting. Figure 4 exemplifies the exponential decaying characteristic in $W_q(q,k)$ for different frequency channels k and

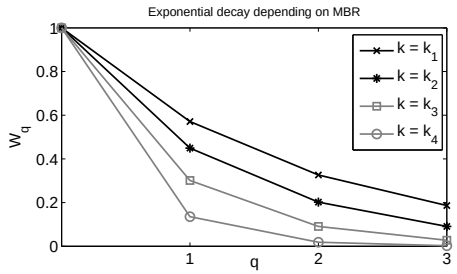


Figure 4: $W_q(q,k)$ as a function of q shown for different values of $k_1 = 2000$ Hz, $k_2 = 700$ Hz, $k_3 = 500$ Hz, $k_4 = 200$ Hz.

across all modulation bands q.

To have a selective noise suppression, similar to conventional DFT-based speech enhancement [1], we utilize DS-based SPP as described in 4.2 and apply it as a scaling factor on the cascade weighting outcome

$$G(q,k) = \text{SPP} \cdot W_e(q,k) W_q(q,k). \tag{12}$$

Finally, we restrict $G(q,k)$ to a lower limit $G_{\min} = 0.178 \triangleq -15$ dB [11] which yields

$$G(q,k) = G_{\min} \quad \text{if} \quad G(q,k) < G_{\min}. \tag{13}$$

Following (7) we apply these weighting coefficients on the noisy DS to obtain $\widehat{DS_x}$. To obtain the enhanced time signal inverse transforms are applied followed by an overlap-and-add routine.

5. Results

In this Section, we demonstrate the effectiveness of the proposed DS-based speech enhancement in a blind scenario and compare its performance versus the STFT-based and modulation-based benchmarks. To check the robustness of the method we provide results for f_0-known versus blind scenario.

5.1. Experimental Setup

Clean speech utterances were taken from Noizeus speech corpus [12] consisting of 30 phonetically-balanced sentences uttered by three males and three female speakers (average length of 2.6 seconds). The speech files were downsampled from the original sampling frequency of 25 kHz to 8 kHz to simulate telephony speech. To obtain noisy files, the clean speech was corrupted in babble noise mixed at SNRs of 0, 5 and 10 dB. As evaluation criteria, we chose *Perceptual Evaluation of Speech Quality* (PESQ) measure [13] and the *Short-Time Objective Intelligibility* (STOI) measure [14]. We report results in terms of improvement in ΔPESQ and ΔSTOI as comparison to the outcome from the noisy (unprocessed) input speech.

To demonstrate the effectiveness of the proposed method, we include three benchmarks: 1) *MMSE-STSA* [15], 2) *ModSpecSub* [5] referring to spectral subtraction in STM, as speech enhancement benchmark, and 3) we report results of fixed-weighting following specification in [4] without LP and time-warping stages. For *MMSE-STSA* a decision-directed scheme was used with a Minimum Statistics noise estimator [16] with a 16 ms frame shift, a 32 ms window length and a Hamming window. For *ModSpecSub* we used the implementation provided by Paliwal et al. [5].

The parameter setup used for the proposed DS-based speech enhancement is as follows. The length of the analysis window is $2P_0$ with 50% overlap, i.e., P_0 of the respective time block. Assuming stationarity for short time intervals [17] and taking a typical range for f_0 into account, we set the number of modulation bands to $Q = 4$.

	babble noise		
SNR-level (dB)	0	5	10
MMSE-STSA [15]	0.18	0.20	0.22
ModSpecSub [5]	0.12	0.12	0.08
Fixed weighting [4]	0.17	0.19	0.17
DSW (blind)	0.27	0.34	0.30
DSW (f_0-known)	0.37	0.38	0.35

Table 1: ΔPESQ results averaged over SNRs and utterances shown for babble noise and different methods.

SNR-level (dB)	0	5	10
MMSE-STSA [15]	-0.01	0.00	0.00
ModSpecSub [5]	-0.04	-0.04	-0.05
Fixed weighting [4]	0.00	-0.01	-0.02
DSW (blind)	-0.03	-0.04	-0.07
DSW (f_0-known)	0.03	0.00	-0.04

Table 2: ΔSTOI results averaged over SNRs and utterances shown for babble noise and different methods.

5.2. Speech Enhancement Results

Tables 1 and 2 report the averaged results of ΔPESQ and ΔSTOI for 30 speakers. The following observations are made:

- The proposed method (*DSW*) leads to a 0.3 improvement in PESQ, outperforming both the *MMSE-STSA* [15] and *ModSpecSub* [5] benchmarks.
- Our pitch estimator performs well. Using an oracle f_0 leads to only a minor improvement in performance in PESQ and STOI. For some audio examples we refer to https://www2.spsc.tugraz.at/people/pmowlaee/DS.html.
- In terms of intelligibility, a fixed weighting similar to [4] results in a better STOI compared to the proposed method at the expense of a lower improvement in the perceived quality predicted by PESQ.

6. Conclusions

In this paper, we proposed *Double Spectrum* (DS) speech enhancement that relies on pitch-synchronous and modulation transforms. The linearity of the DS operator results in a sparse representation of speech that provides a means for the identification and separation of rapidly-varying (noise and unvoiced speech) versus slowly varying (voiced speech) component. These properties facilitate selective noise reduction. Our experiments confirm that DS-based speech enhancement outperforms its STFT and modulation-only counterparts.

The linear property of DS suggests the study of DS subtraction as a direction for future work on the DS noise estimator.

7. References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013.

[2] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 1, no. 4, pp. 386–399, Oct 1993.

[3] ——, "A frame interpretation of sinusoidal coding and waveform interpolation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 3, 2000, pp. 1475–1478.

[4] F. Huang, T. Lee, W. B. Kleijn, and Y.-Y. Kong, "A method of speech periodicity enhancement using transform-domain signal decomposition," *Elsevier speech communication*, vol. 67, pp. 102–112, 2015.

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

[6] K. K. Paliwal, S. Belinda, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Elsevier speech communication*, vol. 54, no. 2, pp. 282–305, 2012.

[7] S. Belinda and K. K. Paliwal, "Using STFT real and imaginary parts of modulation signals for MMSE-based speech enhancement," *Elsevier speech communication*, vol. 58, pp. 49–68, 2014.

[8] M. Nilsson, B. Resch, M. Y. Kim, and W. B. Kleijn, "A canonical representation of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 849–852, 2007.

[9] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 38, no. 6, pp. 969–978, Jun 1990.

[10] M. Nilsson, "Entropy and speech," Ph.D. dissertation, Royal Institute of Technology (KTH), 2006.

[11] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

[12] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Elsevier speech communication*, vol. 49, no. 7–8, pp. 588–601, 2007.

[13] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 2, 2001, pp. 749–752.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

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

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

Bibliography

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain Based Single-Microphone Noise Reduction for Speech Enhancement*, ser. Synthesis Lectures on Speech and Audio Processing. Morgan & Claypool Publishers, 2013.

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

[3] S. Belinda, "Modulation domain based processing for speech enhancement," Ph.D. dissertation, Griffith University, Brisbane, 2012.

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

[5] H. Dudley, "The carrier nature of speech," *Bell System Technical Journal*, vol. 19, no. 4, pp. 495–515, 1940.

[6] E. Zwicker, "Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenz-Modulation eines Tones," *Acta Acustica united with Acustica*, vol. 2, no. 3, pp. 125–133, 1952.

[7] N. F. Viemeister, "Temporal modulation transfer functions based upon modulation thresholds," *The Journal of the Acoustical Society of America*, vol. 66, no. 5, pp. 1364–1380, 1979.

[8] 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.

[9] 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.

[10] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex. i. characteristics of single-unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[11] N. Mesgarani and S. Shamma, "Speech enhancement based on filtering the spectrotemporal modulations," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, vol. 1, March 2005, pp. 1105–1108.

[12] R. Drullman, J. M. 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.

[13] ——, "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.

[14] 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, Oct 1996, pp. 2490–2493 vol.4.

[15] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 7, pp. 1–8, 2003.

[16] M. Nilsson, "Entropy and speech," Ph.D. dissertation, Royal Institute of Technology (KTH), 2006.

[17] M. Nilsson, B. , M. Y. Kim, and W. B. Kleijn, "A canonical representation of speech," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, pp. 849–852, 2007.

[18] L. R. Rabiner and R. W. Schafer, *Theory and applications of digital speech processing.* Upper Saddle River: Pearson, 2011.

[19] P. C. Loizou, *Speech enhancement: theory and practice.* CRC press, 2013.

[20] S. W. Smith, "The scientist and engineer's guide to digital signal processing," accessed: 2016-05-03. [Online]. Available: http://www.dspguide.com

[21] L. A. Zadeh, "Frequency analysis of variable networks," *Proceedings of the IRE*, vol. 38, no. 3, pp. 291–299, March 1950.

[22] S. S. Les Atlas, S. Sukittanon, "Modulation frequency analysis and modification of signals," accessed: 2016-04-19. [Online]. Available: http://isdl.ee.washington.edu/projects/icassptutorial/Audio%20Coding.html

[23] L. Atlas, Q. Li, and J. Thompson, "Homomorphic modulation spectra," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP '04). Proceedings.*, vol. 2, May 2004, pp. ii–761–4 vol.2.

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

[25] K. K. Paliwal, S. Belinda, and K. Wójcicki, "Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator," *Elsevier speech communication*, vol. 54, no. 2, pp. 282–305, 2012.

[26] H. Hermansky, E. A. Wan, and C. Avendano, "Speech enhancement based on temporal processing," in *International Conference on Acoustics, Speech, and Signal Processing, 1995. (ICASSP),*, vol. 1, May 1995, pp. 405–408 vol.1.

[27] W. B. Kleijn, "Encoding speech using prototype waveforms," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 1, no. 4, pp. 386–399, Oct 1993.

[28] ——, "A frame interpretation of sinusoidal coding and waveform interpolation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Proceedings.*, vol. 3, 2000, pp. 1475–1478.

[29] J. Makhoul, "Spectral linear prediction: Properties and applications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 3, pp. 283–296, Jun 1975.

[30] B. Resch, M. Nilsson, A. Ekman, and W. B. Kleijn, "Estimation of the instantaneous pitch of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 813–822, March 2007.

[31] H. S. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 38, no. 6, pp. 969–978, Jun 1990.

[32] F. Huang, T. Lee, W. B. Kleijn, and Y.-Y. Kong, "A method of speech periodicity enhancement using transform-domain signal decomposition," *Elsevier speech communication*, vol. 67, pp. 102–112, 2015.

[33] F. Huang, T. Lee, and W. B. Kleijn, "Transform-domain speech periodicity enhancement with adaptive coefficient weighting," in *2011 International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS).*, Dec 2011, pp. 1–5.

[34] ——, "Transform-domain wiener filter for speech periodicity enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4577–4580.

[35] G. Strang, "The discrete cosine transform," *SIAM Review*, vol. 41, no. 1, pp. 135–147, 1999.

[36] N. Vasconcelos, "Discrete cosine transform," accessed: 2016-04-26. [Online]. Available: http://www.svcl.ucsd.edu/courses/ece161c/handouts/DCT.pdf

[37] R. Zelinski and P. Noll, "Adaptive transform coding of speech signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 4, pp. 299–309, 1977.

[38] R. Reininger and J. D. Gibson, "Distributions of the two-dimensional DCT coefficients for images," *IEEE Transactions on Communications*, vol. 31, no. 6, pp. 835–839, 1983.

[39] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, 2000.

[40] S. R. Smoot and L. A. Rowe, "DCT coefficient distributions," in *Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1996, pp. 403–411.

[41] D. Ealey, H. Kelleher, and D. Pearce, "Harmonic tunnelling: tracking non-stationary noises during speech." in *INTERSPEECH*, 2001, pp. 437–440.

[42] H. Traunmüller and A. Eriksson, "The frequency range of the voice fundamental in the speech of male and female adults," *Consulté le*, vol. 12, no. 02, p. 2013, 1995.

[43] R. Wang, "Heisenberg uncertainty principle," accessed: 2016-05-15. [Online]. Available: http://fourier.eng.hmc.edu/e161/lectures/fourier/node2.html

[44] S. Gonzalez and M. Brookes, "PEFAC - a pitch estimation algorithm robust to high levels of noise," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 22, no. 2, pp. 518–530, Feb 2014.

[45] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[46] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 32, no. 6, pp. 1109–1121, Dec 1984.

[47] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Elsevier speech communication*, vol. 49, no. 7–8, pp. 588–601, 2007.

[48] J. Castiñeira Moreira and P. G. Farrell, *Appendix A: Error Probability in the Transmission of Digital Signals*. John Wiley & Sons, Ltd, 2006, pp. 327–337.

[49] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[50] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series*. MIT press Cambridge, MA, 1949, vol. 2.

[51] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 4, pp. 251–266, 1995.

[52] M. Blass, P. Mowlaee, and W. B. Kleijn, "Single-channel speech enhancement using double spectrum," *submitted to INTERSPEECH*, 2016.

[53] J. G. Lyons and K. K. Paliwal, "Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement." in *INTERSPEECH*. Citeseer, 2008, pp. 387–390.

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

[55] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 2, no. 2, pp. 345–349, Apr 1994.

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

[57] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 1979, pp. 208–211.

[58] P. Scalart *et al.*, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1996). Conference Proceedings.*, vol. 2, 1996, pp. 629–632.

[59] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[60] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, Feb. 1993.

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

[62] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

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

[64] S. R. Quackenbush, . Barnwell, T. P. (Thomas Pinkney), and M. A. Clements, *Objective measures of speech quality.* Englewood Cliffs, N.J. : Prentice Hall, 1988, includes bibliographies and index.

[65] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2125–2136, Sept 2011.

[66] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[67] R. Martin, "Speech enhancement using mmse short time spectral estimation with gamma

distributed speech priors," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1. IEEE, 2002, pp. I–253.

[68] ——, "Speech enhancement based on minimum mean-square error estimation and super-gaussian priors," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 845–856, 2005.

[69] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1741–1752, 2007.

[70] W. B. Kleijn, J. B. Crespo, R. C. Hendriks, P. Petkov, B. Sauert, and P. Vary, "Optimizing speech intelligibility in a noisy environment: A unified view," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 43–54, 2015.

[71] "Documentation for TIMIT," accessed: 2016-05-22. [Online]. Available: https://catalog.ldc.upenn.edu/docs/LDC93S1/