Advances in DFT-Based Single-Microphone Speech Enhancement

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. J. T. Fokkema,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 18 februari 2008 om 10:00 uur
door Richard Christian HENDRIKS
elektrotechnisch ingenieur
geboren te Schiedam.

Dit proefschrift is goedgekeurd door de promotor:
Prof. dr. ir. J. Biemond

Toegevoegd promotor:
Dr. ir. R. Heusdens

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. J. Biemond,	Technische Universiteit Delft, promotor
Dr. ir. R. Heusdens,	Technische Universiteit Delft, toegevoegd promotor
Prof. dr. ir. A.-J. van der Veen,	Technische Universiteit Delft
Prof. dr. ir. J. W. M. Bergmans,	Technische Universiteit Eindhoven
Prof. dr. M. A. Clements,	Georgia Institute of Technology, Atlanta, United States
Prof. dr. W. B. Kleijn,	Royal Institute of Technology, Stockholm, Sweden
Dr. U. Kjems,	Oticon, Copenhagen, Denmark
Prof. dr. ir. R. L. Lagendijk	Technische Universiteit Delft, reservelid

Dr. J. Jensen heeft als begeleider in belangrijke mate aan de totstandkoming van het proefschrift bijgedragen.

The work described in this thesis was financially supported by STW and Philips Research Laboratories.

ISBN: 978-90-9022690-3

Advances in DFT-Based Single-Microphone Speech Enhancement

To Rosalie,

to my mother,

and in memory to my father.

Summary

The interest in the field of speech enhancement emerges from the increased usage of digital speech processing applications like mobile telephony, digital hearing aids and human-machine communication systems in our daily life. The trend to make these applications mobile increases the variety of potential sources for quality degradation. Speech enhancement methods can be used to increase the quality of these speech processing devices and make them more robust under noisy conditions.

The name "speech enhancement" refers to a large group of methods that are all meant to improve certain quality aspects of these devices. Examples of speech enhancement algorithms are echo control, bandwidth extension, packet loss concealment and (additive) noise reduction. In this thesis we will focus on single-microphone additive noise reduction and aim at methods that work in the discrete Fourier transform (DFT) domain. The main objective of the presented research is to improve on existing single-microphone schemes for an extended range of noise types and noise levels, thereby making these methods more suitable for mobile speech communication applications than state-of-the-art algorithms.

The research topics in this thesis are three-fold. At first, we focus on improved estimation of the *a priori* signal-to-noise ratio (SNR) from the noisy speech. Good *a priori* SNR estimation is crucial for speech enhancement, since many speech enhancement estimators depend on this parameter. We focus on two aspects of *a priori* SNR estimation. Firstly, we present an adaptive time-segmentation algorithm, which we use to reduce the variance of the estimated *a priori* SNR. Secondly, an approach is presented to reduce the bias of the estimated *a priori* SNR, which is often present during transitions between speech sounds or transitions from noisy speech to noise-only and vice versa. The use of these improved *a priori* SNR estimators leads to both objective and subjective quality improvement.

Secondly, we investigate the derivation of clean speech estimators under models that take properties of speech into account. This problem is approached from two different angles. At first, we consider the derivation of clean speech estimators under the use of a combined stochastic/deterministic model for the complex DFT coefficients. The use of a deterministic model is based on the fact that certain speech sounds have a more deterministic character. Secondly, we focus on the derivation of complex DFT and magnitude DFT estimators under super-Gaussian densities. Derivation of clean speech estimators under these types of densities is based on measured histograms of speech DFT coefficients. We present two different type of estimators

under super-Gaussian densities. Minimum mean-square error (MMSE) estimators are derived under a generalized Gamma density for the clean speech DFT coefficients and DFT magnitudes. Maximum *a posteriori* (MAP) estimators are derived under the multivariate normal inverse Gaussian (MNIG) density for the clean speech DFT coefficients. Objective performance of the estimators derived under the MNIG density is slightly better than for the estimators derived under the generalized Gamma density. Moreover, estimators derived under the MNIG density have some theoretical advantages over estimators derived under the generalized Gamma density. More specifically, under the MNIG density the statistical models in the complex DFT domain and the polar domain are consistent, which is not the case for estimators derived under the generalized Gamma density. In addition, the MNIG density can model vector processes, which allows for taking into account the dependency between the real part and imaginary part of DFT coefficients.

Finally, we developed a method for tracking of the noise power spectral density (PSD). The fact that all clean speech estimators are dependent on the noise PSD makes this an important research topic. However, fast and accurate tracking of the noise PSD is very challenging. The developed method is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. This approach makes it possible, in contrast to most existing methods, to update the noise PSD even when speech is continuously present. Furthermore, the tracking delay is considerably reduced compared to state-of-the-art noise tracking algorithms.

Some of the contributions presented in this thesis can be combined into a complete speech enhancement system. A comparison is performed between a combination of these individual components and a state-of-the-art speech enhancement system from literature. Subjective experiments by means of a listening test show that the system based on contributions of this thesis improves significantly over the state-of-the-art speech enhancement system.

Table of Contents

Chapter 1

Introduction

Over the last two decades society experienced an increase in the use of speech-processing devices like cellular phones, digital hearing aids and all kind of human-to-machine speech-processing applications. With the increased use of these devices, also the variety of application environments increased. As a consequence, these speech processors are potentially exposed to a large variety of acoustic noise sources. Although most of these applications have originally been developed to work with noise-free signals, as is e.g. the case for most speech coding and speech recognition algorithms, there has been an increasing interest to make these systems robust to work under these noisy conditions as well. Speech enhancement methods can be used to improve the quality of these speech-processing devices. The term speech enhancement in fact refers to a large group of methods that all aim at improving the quality of speech signals in some way. Some examples of speech enhancement methods are bandwidth extension, acoustic echo control (dereverberation), packet loss concealment and noise reduction. In this thesis we use the expression *speech enhancement* in the meaning of additive noise reduction.

The group of noise reduction methods for speech enhancement can be divided into two broad classes; the class of single-microphone noise reduction and the class of multi-microphone noise reduction.

Single-microphone speech enhancement algorithms estimate the clean speech signal using a realization of the noisy speech that is obtained using one microphone. These methods have in general lower costs than multi-microphone algorithms. Moreover, single-microphone algorithms often impose less constraints on the system than multi-microphone systems, for example requirements on the distance between the microphones. Multi-microphone enhancement algorithms on the other hand use more than one microphone and can as such also exploit spatial information, and, as a consequence, their performance is in general better than single-microphone speech enhancement systems. However, due to physical size limitation it is not always obvious how to implement multi-microphone algorithms on small devices when one has to fulfill for example the microphone inter-distance requirements.

In this thesis we focus on single-microphone speech enhancement. However, notice that often single-microphone methods can be extended and used as a multi-

microphone system as well. Moreover, single-microphone based methods can often be combined with multi-microphone algorithms as a post-processor to obtain an even better noise reduction.

There are several ways to classify existing single-microphone speech enhancement algorithms. One way is to make a distinction between methods that are based on signal subspace decompositions, methods based on parametric models, and methods based on processing in the discrete Fourier transformation (DFT) domain. The above mentioned classes of enhancement methods are not strictly disjoint and there are algorithms which do not naturally fit into any of these classes. In the following we give a brief overview of these classes.

The application of signal subspace decompositions within the context of speech enhancement was proposed by Ephraim and Van Trees in [1]. The subspace based approaches exploit the fact that the covariance matrix of a noisy speech signal frame can be decomposed into two mutually orthogonal vector spaces: a signal (+noise) subspace and a noise-only subspace. Noise reduction is obtained by discarding the noise-only subspace completely, while modifying the noisy speech components in the signal (+noise) subspace. A basic limitation of subspace based speech enhancement is the relatively high computational complexity. More specifically, the method is based on eigenvalue decompositions of the noisy speech covariance matrix. These eigenvalue decompositions (EVD's) are computationally quite expensive when the dimension of the covariance matrices become large. Another important aspect is the fact that subspace based speech enhancement assumes the noise process to be white. Extensions of subspace based enhancement methods that work for colored noise have been proposed, see e.g. [2]. Other extensions that have been proposed take perceptual aspects into account [3][4][5].

The second class contains methods where parametric models are fitted to the speech signal and used in combination with a filter, e.g. a Wiener or Kalman filter, to estimate the clean speech signal, see e.g. [6][7][8][9]. These methods often apply certain constraints on the estimation process by using the fact that speech can be very well represented as an autoregressive (AR) process. As such these methods can exploit certain *a priori* information and can make sure that the enhanced speech signal satisfies certain spectral constraints or constraints on the time-evolution of the enhanced speech spectra. These methods often use hidden Markov models (HMMs) [8] or codebooks [7][10] in order to determine the parametric description of the speech signal. Some of these methods also use an HMM or codebook to model the noise process with a parametric model, see e.g. [8][11]. Clearly, modelling the noise process with an HMM or codebook constrains the system to work for certain noise-types only.

The final group that we mention here is the class of DFT-domain based methods. These methods transform the noisy speech signal frame-by-frame to the spectral domain, e.g. using a discrete Fourier transform (DFT). Here, complex-valued DFT coefficients of the clean signal are estimated by applying a gain function to the noisy DFT coefficients. Subsequently, enhanced time-domain frames are generated using the inverse DFT. The enhanced waveform is constructed by overlap-adding the enhanced frames. Initially, processing of the noisy DFT coefficients was mainly based on the spectral subtraction type of methods, see e.g. [12][13]. Later, somewhat more sophis-

ticated methods were proposed, where estimators were derived under a certain error criterion and by exploiting (assumed) densities of noise and speech DFT coefficients [14][15][16][17]. These estimators are a function of the distributional parameters, e.g., the variance of the noise and speech DFT coefficients. Also some variants have been proposed where it is tried to take perception into account, see e.g. [18][19][20]. DFT-domain based speech enhancement has received significant interest recently, partly due to their relatively good performance and low computational complexity.

An important difference between the DFT-domain and the subspace based approaches is that the latter is based on a spectral transformation that is signal dependent. Despite the possible gain of subspace based methods over DFT-domain based methods due to their somewhat more advanced signal transformation, the gain is rather low and the added complexity is often hard to justify.

Parametric methods can be implemented in the DFT domain as well and can as such be combined with other DFT-domain based methods. The advantage of parametric methods based on HMMs and codebooks is that they can incorporate good statistical models of the speech process. However, they generally need a statistical model of the noise process as well. This severely restricts the situations in which the enhancement system will work. To overcome this restriction, methods have been proposed that use a set of noise models and then use the noise model that fits best to the situation at hand [11]. This of course broadens the applicability to more noise types, but pays a price in terms of a increased complexity and memory usage. Moreover, there is no guarantee that the system can handle a practical noise situation that it is not trained for. Moreover, not all noise types can be described well with a low-order AR model.

Notice that the list of references that we have given here is far from exhaustive, since much research has been done in the field of speech enhancement. However, the field of single-channel speech enhancement is still very challenging. There are many scenarios, e.g. under low signal-to-noise ratio (SNR) or under non-stationary noise conditions where existing systems fail to lead to a satisfying result.

In this thesis we mainly focus on the class of DFT-domain based methods for single-channel speech enhancement. The work presented in this thesis was done within the project *single-microphone enhancement of noisy speech signals*, supported by STW and Philips research. The problem statement within this project was to develop methods that can improve on existing single-microphone schemes for an extended range of noise types and noise levels. From this problem statement, the following three research topics were derived:

- To investigate clean speech estimators based on models that give a good description of the speech process.

- To develop a method for tracking of noise statistics (for stationary as well as non-stationary noise sources) during speech activity and with a short delay.

- To improve estimation of parameters that are used to express speech estimators, e.g. the variance of speech DFT coefficients, by taking into account that speech is a time-varying process.

In the next section we will give an overview of a general DFT-domain based speech enhancement scheme and relate the aforementioned research topics to this scheme.

1.1 DFT-Domain Based Single-Channel Speech Enhancement

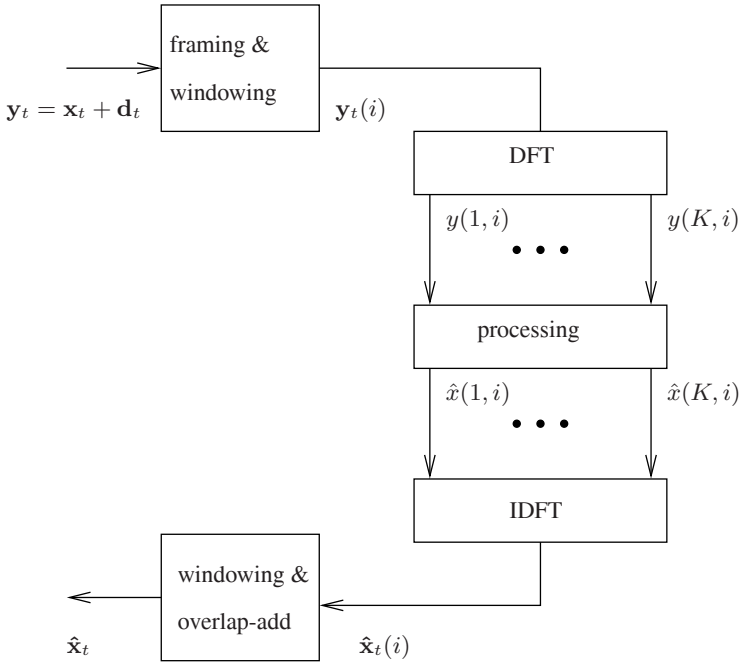

Figure 1.1: *Overview of a DFT-domain based single-channel speech enhancement system.*

In Fig. 1.1 the basic block-scheme of a DFT-domain based single-channel speech enhancement system is shown. The small letters indicate realizations of random variables and boldface letters indicate vectors. The input to this system is a signal $\mathbf{y}_t$, where the subscript t indicates that this is a sampled time-domain signal. This signal is a noisy version of the (unknown) clean speech signal $\mathbf{x}_t$. The purpose of this speech enhancement system is to make an estimate $\hat{\mathbf{x}}_t$ of $\mathbf{x}_t$ that satisfies certain quality criteria. Speech signals are non-stationary by nature. Therefore, processing of $\mathbf{y}_t$ is performed on a frame-by-frame basis, where in general the frames have a length of 10 up to 40 ms to satisfy quasi-stationarity conditions. A frame of noisy speech is indicated by $\mathbf{y}_t(i)$, where i indicates the frame number index. The frames have a length of K samples, and are selected from the noisy time signal with an overlap of P samples. By cutting out a frame from the signal $\mathbf{y}_t$, implicitly a so-called analysis window is applied. If no special actions are taken, this will be rectangular,

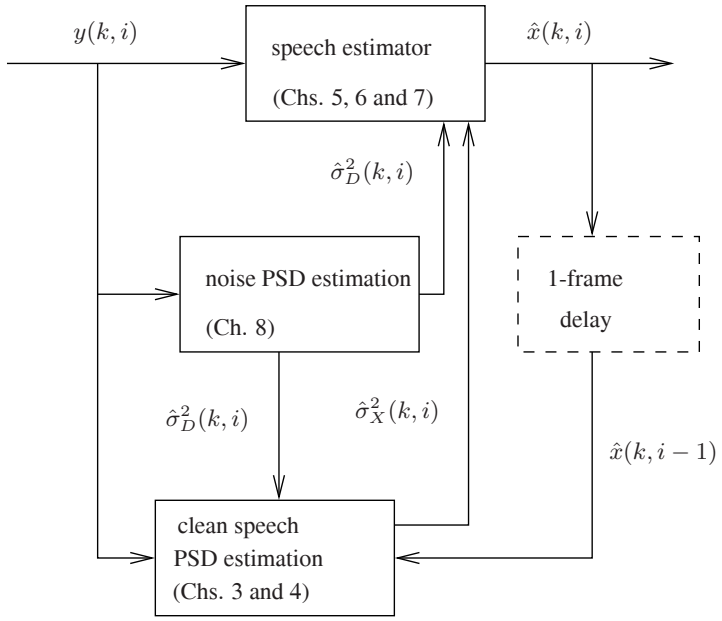

Figure 1.2: *Typical structure of a DFT-domain based single-channel speech enhancement system, with indication how the chapters in this thesis relate to the different blocks in the enhancement scheme.*

however, often more smooth windows are desirable like a Hann or Hamming window. The windowed frame $\mathbf{y}_t$ is transformed to the spectral domain by applying a discrete Fourier transform (DFT), leading to a set of K DFT coefficients for frame i, i.e. $y(k,i)$, with $k \in \{1,...,K\}$ the frequency bin index. In the next block, labelled as *processing*, clean speech DFT coefficients are estimated by processing the noisy DFT coefficients $y(k,i)$. Estimated clean speech DFT coefficients are denoted as $\hat{x}(k,i)$. This block is of most interest for us, since the following chapters will deal with how to process the noisy speech DFT coefficients, such that an estimate of the clean speech DFT coefficients is obtained. Subsequently, an inverse DFT is applied on the estimated clean speech DFT coefficients leading to an estimated clean speech time-domain frame $\hat{\mathbf{x}}_t(i)$. Possibly $\hat{\mathbf{x}}_t(i)$ is windowed again, using a so-called synthesis window, and the estimated clean speech signal is reconstructed using an overlap-add procedure. Often, the analysis and synthesis windows are chosen such that when no processing is performed in the block labelled as *processing*, a perfect (possibly delayed) reconstruction of the input signal is given at the output.

Although the variety in DFT-domain based enhancement schemes is large, in general a common structure for the processing block can be recognized, see Fig. 1.2. Most DFT based enhancement systems assume that the DFT coefficients are independent over frequency bins and time-frame indices and therefore process the noisy DFT coefficients independently. The scheme in Fig. 1.2 is therefore drawn for a single DFT coefficient at frequency bin k and frame i. The steps should be repeated for all other

bins $k \in \{1, ..., K\}$ and time-frames.

In general, a single-channel DFT-domain speech enhancement algorithm consists of three main components. The actual estimation of the clean speech DFT coefficients is performed in the *speech estimator* block in Fig. 1.2, leading to an estimate $\hat{x}(k,i)$ of $x(k,i)$. The first research topic from the list on page 3 is related to this block. Many procedures exist to obtain estimators for clean speech DFT coefficients. Some are based on more or less heuristic argumentations, for example the spectral subtraction based methods [12][13], while others consider the clean speech and noise DFT coefficients as random variables $X(k,i)$ and $D(k,i)$, respectively, with a certain density and employ so-called Bayes estimators, e.g. the minimum mean-square error (MMSE) estimator [14] or the maximum *a posteriori* (MAP) estimator [17]. Most of these estimators can be expressed in terms of the variance of the noise DFT coefficient $\sigma_D^2(k,i)$ and the variance of the clean speech DFT coefficient $\sigma_X^2(k,i)$. These variances are computed over the ensembles of the stochastic processes $D(k,i)$ and $X(k,i)$, respectively. The variances $\sigma_D^2(k,i)$ and $\sigma_X^2(k,i)$ are also referred to as the noise and clean speech power spectral density (PSD), respectively. Often, these two quantities are expressed as a ratio, namely as the *a priori* SNR $\xi(k,i)$, which is defined as

$$\xi(k,i) = \frac{\sigma_X^2(k,i)}{\sigma_D^2(k,i)}.$$

Since both these quantities are unknown, estimation from the noisy data is necessary. This is done in the two other blocks in the block diagram of Fig. 1.2.

In the block labelled *noise PSD estimation*, the noise power spectral density is estimated. Estimation of the noise power spectral density is related to the second research topic on page 3. The estimated noise power spectral density is denoted by $\hat{\sigma}_D^2$. A common method for estimation of the noise PSD is to exploit speech pauses, where noise statistics can be measured. Detection of these pauses can be done using a voice activity detector (VAD) [21][22]. However, this method is only valid for stationary noise. Somewhat more advanced methods for noise PSD estimation are based on so-called minimum statistics [23][24][25]. The minimum statistics based methods do not need a VAD to estimate the noise PSD, but track the minimum power level in a particular frequency bin seen across a sufficiently long time interval and compute the noise PSD from this minimum.

Besides the noise PSD, most DFT-domain based noise reduction methods also require an estimate of the clean speech PSD. This estimate is obtained in the block labelled as *clean speech PSD estimation* and is related to the third and final research topic on page 3. Under certain assumptions, which will be specified in the next chapter, the speech PSD can be estimated by subtracting an estimate of the noise PSD from the noisy speech PSD. Since the latter is unknown as well, it is often estimated by averaging the power of noisy DFT coefficients from a few consecutive frames over time, see e.g. [12]. Often, the estimated speech PSD shows variations due to random fluctuation of the noisy speech realization. Since these variations can lead to perceptually annoying artifacts, other methods have been proposed that lead to estimates of the speech PSD which exhibit smoother time variation. A very popular method that leads to relatively smooth estimates is the so-called decision-directed approach [14].

It makes use of an estimate of the clean speech magnitude spectrum from the previous frame, indicated by the 1-frame delay block in Fig. 1.2, in combination with the estimated noise PSD and a periodogram estimate of the noisy speech PSD, to obtain smooth estimates of the current clean speech PSD.

1.2 Contributions

In this thesis we mainly deal with DFT-domain based single-microphone speech enhancement. Our main focus is on three different, but related, topics within single-microphone speech enhancement. First, we investigate improved estimation of the *a priori* SNR. Secondly, our research focusses on speech enhancement estimators that can take (statistical) properties of speech into account and as such lead to a better estimate of the clean speech signal. Thirdly, research aimed at developing a noise tracking method which can track the noise statistics when speech is continuously present. More specifically, the main contributions of this thesis are the following:

1. **Adaptive time segmentation for speech enhancement**

   We present an algorithm that can be used to obtain an adaptive time segmentation for noisy speech. The segmentation algorithm determines for each frame of noisy speech a corresponding stationary segment. Such a segment can be used to obtain an improved estimate of the noisy speech PSD, since it takes the region in which the (noisy) data is stationary into account.

2. **Improved *a priori* SNR estimation**

   Often, speech enhancement estimators are expressed in terms of the *a priori* SNR. Since this quantity is unknown in advance, it is often estimated using the so-called decision-directed [14] approach. There are two important aspects related to the estimated *a priori* SNR. First, the estimated *a priori* SNR can show some variations over time, leading to a rather annoying type of residual noise. Secondly, the decision-directed approach can lead to underestimates or overestimates around stationarity boundaries in the clean speech signal. With respect to the first aspect, the aforementioned improved estimate of the noisy speech PSD that is based on an adaptive time segmentation can be used within the decision-directed approach. This reduces the variance of the estimated *a priori* SNR.

   With respect to the second aspect, we present a so-called backward decision-directed approach. Combined with the standard decision-directed approach this can overcome overestimates and underestimates of the *a priori* SNR at the start of stationary regions.

3. **Clean speech DFT estimator under a combined stochastic-deterministic model**

   We present an MMSE estimator under a combined stochastic-deterministic model for the complex speech DFT coefficients. The use of the deterministic model is based on the observation that certain speech sounds have a more deterministic

character. Especially in frequency bins containing harmonics the deterministic speech model is more appropriate and leads to improvements over the use of a stochastic model.

4. **Clean speech DFT estimators derived under super-Gaussian densities**

   We present clean speech complex DFT coefficient and DFT magnitude estimators derived under two different densities that are known to be able to model super-Gaussian or (semi-)heavy-tailed processes very well. More specifically, we present MMSE estimators derived under the generalized Gamma density and MAP estimators derived under the multivariate normal inverse Gaussian (MNIG) density. These densities are of interest for the derivation of clean speech DFT estimators, because they provide good models for super-Gaussian or so-called (semi-)heavy tailed processes and show a much better fit to speech DFT histograms than more conventional densities like the Gaussian density. The presented MMSE estimator that is derived under the generalized Gamma density is a generalization of existing complex DFT and magnitude estimators, i.e. for specific parameter settings already existing estimators are obtained. The MAP estimator that we derive under the MNIG density has some advantages over the generalized Gamma density. At first, besides scalar processes it can model vector processes as well. As such, dependencies between vector elements, e.g. the real part and imaginary part of DFT coefficients, can be taken into account. With the generalized Gamma density this is in general not possible. Secondly, under the generalized Gamma density there is in general no consistency between the statistical models in the complex DFT domain and the polar domain. Under the MNIG density, on the other hand, the models are consistent.

5. **Tracking of noise statistics**

   A novel approach for noise tracking is proposed. In contrast to most existing noise tracking algorithms, this method can track the noise statistics also when speech is constantly present at a certain frequency bin. Moreover, the tracking delay in comparison to existing schemes is considerably reduced. An increase of the noise level of 15 dB per second can easily be tracked, which leads to an increase of the final enhancement performance in terms of segmental SNR of several dB's.

How these contributions fit into a general DFT-domain enhancement scheme and how they relate to each other can be indicated using the block-diagram in Fig. 1.2. Contribution 1 uses the noisy input DFT coefficients $y(k,i)$ to determine an adaptive time segmentation. It can not be directly related to any of the indicated blocks, but it can be used in combination with e.g. the blocks labelled as *clean speech PSD estimation* or *noise PSD estimation* to improve estimation of time-varying parameters. In this thesis, the adaptive time segmentation is used in contribution 2, which is related to the block labelled as *clean speech PSD estimation*, to improve estimation of the *a priori SNR*. From this (improved) estimate of the *a priori SNR* the speech PSD can be computed. The estimators in contributions 3 and 4 can be used in the block labelled

speech estimator to perform the actual estimation of the clean speech DFT coefficients. Finally, contribution 5 is used in the *noise PSD estimation*-block to estimate the noise PSD σ_D^2.

1.3 Thesis Outline

The notation and basic assumptions that we use throughout this thesis are introduced in Chapter 2. Further, Chapter 2 provides some background information on DFT-domain based speech enhancement and related topics on which other chapters are based.

In Chapter 3 an algorithm is presented that can be used to obtain an adaptive time segmentation based on noisy speech. We use this segmentation to obtain better estimates of the noisy speech PSD. Subsequently, this estimated noisy speech PSD is used in combination with the decision-directed approach in order to obtain improved estimates of the *a priori SNR*, which is a parameter that is frequently needed when computing speech enhancement gain functions.

In Chapter 4 another method is presented that aims at obtaining improved estimates of the *a priori SNR*. A property of the conventional decision-directed approach is that in general it leads to wrong estimates of the *a priori SNR* at each start of a stationary region. This behavior is related to the fact that the decision-directed approach makes use of clean speech estimates from the previous frame to make an estimate of the *a priori SNR* for the current frame. In Chapter 4 a backward decision-directed approach is presented where the *a priori SNR* is estimated using clean speech estimates from future frames. Estimates of the *a priori SNR* that are obtained using the conventional decision-directed approach and the presented backward decision-directed approach are combined into one single estimate by making use of the adaptive time-segmentation algorithm presented in Chapter 3.

Many DFT-domain based speech enhancement estimators are based on the assumption that speech DFT coefficients can be modelled as random variables with a certain density. However, it is known that some classes of speech sounds can be very well modelled with a deterministic model. Therefore, we investigate in Chapter 5 the use of a mixture of a deterministic and a stochastic speech model for speech DFT coefficients. Under this combined stochastic-deterministic model, an estimator for clean speech DFT coefficients is derived.

As mentioned above, an alternative to the use of such a combined stochastic-deterministic model is to consider speech DFT coefficients to be random variables. Several studies have been published where the density of speech DFT coefficients is studied, see e.g. [16][17]. From these studies it followed that the observed density of speech DFT coefficients has a so-called super-Gaussian shape, i.e. more heavy tailed and more peaked than a Gaussian density. To be able to exploit this knowledge, we derive in Chapters 6 and 7 speech estimators that can be used for a broad class of densities. More specifically, in Chapter 6 we derive MMSE estimators for complex DFT coefficients and DFT magnitudes under the generalized Gamma density. This leads to a generalization of the estimators derived in [14][16][17]. A potential weakness of the estimators derived under the generalized Gamma density is that real and imaginary parts of DFT coefficients are assumed to be independent, which will be shown

in Chapter 6 to be not completely in line with measured speech data. Further, under the generalized Gamma density there is no consistency between the models in the complex DFT domain and the polar domain for all parameter settings of the density.

In Chapter 7 MAP estimators for complex DFT coefficients and DFT magnitudes are derived by assuming that the complex DFT coefficients are distributed with a multivariate normal inverse Gaussian density. Estimators derived under this density eliminate the above mentioned potential weaknesses of the estimators under the generalized Gamma density.

In Chapter 8 a method is proposed for tracking of the noise PSD. This method is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. This approach makes it possible to update the noise PSD, even when speech is continuously present. Furthermore, the tracking delay is considerably reduced compared to a state-of-the-art noise tracking algorithm.

Finally, in Chapter 9 we summarize the main results of this thesis and discuss some directions that are interesting for future research.

1.4 List of Papers

The following papers have been published by the author of this thesis during his Ph.D. studies:

Journals

[A] R. C. Hendriks, J. Jensen and R. Heusdens. Noise Tracking using DFT Domain Subspace Decompositions, *IEEE Trans. Audio, Speech and Language Processing*, March 2008.

[B] J. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen. Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors, *IEEE Trans. Audio, Speech and Language Processing*, 15(6):1741 - 1752, August 2007.

[C] R. C. Hendriks and R. Martin. MAP Estimators for Speech Enhancement under Normal and Rayleigh Inverse Gaussian Distributions, *IEEE Trans. Audio, Speech and Language Processing*, 15(3):918 - 927, March 2007.

[D] R. C. Hendriks, R. Heusdens and J. Jensen. An MMSE Estimator for Speech Enhancement under a Combined Stochastic-Deterministic Speech Model, *IEEE Trans. Audio, Speech and Language Processing*, 15(2):406 - 415, Feb. 2007.

[E] R. C. Hendriks, R. Heusdens and J. Jensen. Adaptive Time Segmentation for Improved Speech Enhancement, *IEEE Trans. Audio, Speech and Language Processing*, 14(6):2064 - 2074, Nov. 2006.

Conferences

[a] R. C. Hendriks, J. Jensen and R. Heusdens. DFT Domain Subspace Based Noise Tracking for Speech Enhancement, *Proc. Interspeech*, pp. 830-833, August 2007.

[b] R. C. Hendriks, J. S. Erkelens, J. Jensen and R. Heusdens. Minimum mean-square error amplitude estimators for speech enhancement under the generalized Gamma distribution, In *Proc. Int. Workshop on Acoustic Echo and Noise Control*, Paris, France, September 2006.

[c] J. Jensen, R. C. Hendriks, J. S. Erkelens and R. Heusdens. MMSE estimation of complex-valued discrete Fourier coefficients with generalized Gamma priors, In *Proc. Interspeech*, pp. 257-260, September 2006.

[d] R. C. Hendriks, R. Heusdens and J. Jensen. Speech Enhancement Under a Combined Stochastic-Deterministic Model, In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. 1, pp. 453-546, Toulouse, France, May 14-19, 2006.

[e] R. C. Hendriks, R. Heusdens and J. Jensen. Improved Decision Directed Approach for Speech Enhancement Using an Adaptive Time Segmentation, In *Proc. Interspeech*, pp. 2101-2104, Lisboa, Portugal, September 4-8, 2005,

[f] R. C. Hendriks, R. Heusdens and J. Jensen. Forward-Backward Decision Directed Approach for Speech Enhancement, *Proc. Int. Workshop on Acoustic Echo and Noise Control*, pp. 109–112, Eindhoven, The Netherlands, September 12-15, 2005.

[g] R. C. Hendriks, R. Heusdens and J. Jensen. Adaptive Time Segmentation of Noisy Speech for Improved Speech Enhancement, In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. I, pp. 153-156, Philadelphia, PA, USA, March 18-23, 2005.

[h] R. C. Hendriks, R. Heusdens and J. Jensen. Improved Subspace Based Speech Enhancement Using an Adaptive Time Segmentation, In *Proc. IEEE First BENELUX/DSP Valley Signal Processing Symposium* , pp. 163-166, Antwerp, Belgium, April 19-20, 2005.

[i] J. Jensen, I. Batina, R. C. Hendriks and R. Heusdens. A Study of the Distribution of Time-Domain Speech Samples and Discrete Fourier Coefficients, In *Proc. IEEE First BENELUX/DSP Valley Signal Processing Symposium*, pp. 155-158, Antwerp, Belgium, April 19-20, 2005.

[j] R. C. Hendriks, R. Heusdens and J. Jensen. Perceptual linear predictive noise modelling for sinusoid-plus-noise audio coding, In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. IV, pp. 189-192, Montreal, Canada, May 17-21, 2004.

[k] J. Jensen, R. C. Hendriks, R. Heusdens and S. H. Jensen. Smoothed Subspace based Noise Suppression with Application to Speech Enhancement, In *Proc. European Signal Processing Conference*, Antalya, Turkey, September 4-8, 2005.

References

[1] Y. Ephraim and H. L. van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 3(4):251–266, July 1995.

[2] H. Lev-Ari and Y. Ephraim. Extension of the signal subspace speech enhancement approach to colored noise. *IEEE Signal Processing Letters*, 10(4):104–106, April 2003.

[3] F. Jabloun and B. Champagne. A perceptual signal subspace approach for speech enhancement in colored noise. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 569–572, May 2002.

[4] F. Jabloun and B. Champagne. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 2003.

[5] J. U. Kim, S. G. Kim, and C. D. Yoo. The incorporation of masking threshold to subspace speech enhancement. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 76–79, 2003.

[6] J. S. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-26(3):197–210, June 1978.

[7] J. H. L. Hansen and M. A. Clements. Constrained iterative speech enhancement with application to speech recognition. *IEEE Trans. Signal Processing*, 39(4):795–805, April 1991.

[8] Y. Ephraim. A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. on Signal Processing*, 40(4):725–735, April 1992.

[9] J. Jensen and J. H. L. Hansen. Speech enhancement using a constrained iterative sinusoidal model. *IEEE Trans. Speech Audio Processing*, 9(7):731–740, October 2001.

[10] T. V. Sreenivas and P. Kirnapure. Codebook constrained wiener filtering for speech enhancement. *IEEE Trans. Speech Audio Processing*, 4(5):383–389, September 1996.

[11] S. Srinivasan. *Knowledge-Based Speech Enhancement*. PhD thesis, 2005.

[12] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[13] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. of the IEEE*, 67(12):1586–1604, December 1979.

[14] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[15] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(2):443–445, April 1985.

[16] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[17] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[18] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis. Speech enhancement based on audible noise suppression. *IEEE Trans. Speech Audio Processing*, 5(6):497–514, November 1997.

[19] S. Gustafsson, P. Jax, and P. Vary. A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 397–400, May 1998.

[20] N. Virag. Signle channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Processing*, 7(2):126–137, March 1999.

[21] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Lett.*, 6(1), 1999.

[22] J. Chang, N. S. Kim, and S. K. Mitra. Voice activity detection based on multiple statistical models. *IEEE Trans. Signal Processing*, 54(6), 2006.

[23] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. Eur. Signal Processing Conf.*, pages 1182–1185, 1994.

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

[25] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Processing*, 11(5):446–475, September 2003.

Chapter 2

Background

In this chapter we provide background information on DFT-domain based speech enhancement necessary to read this thesis.

2.1 Notation and Basic Assumptions

In this section we introduce the notation and basic assumptions that we use in this thesis. We assume a signal model of the form

$$Y(k,i) = X(k,i) + D(k,i), \tag{2.1}$$

where $Y(k,i)$, $X(k,i)$ and $D(k,i)$ are DFT coefficients obtained at frequency bin index k, with $k \in \{1,...,K\}$ and in time-frame i from the noisy speech, clean speech and noise process, respectively. The signal model expressed in Eq. (2.1) is often referred to as the assumption of additive noise.

We assume that $Y(k,i)$, $X(k,i)$ and $D(k,i)$ are zero-mean complex random variables, unless stated otherwise. We use upper case letters to denote random variables and the corresponding lower case letters to denote their realizations. Vectors and matrices are indicated by boldface letters, e.g. $\mathbf{Y} \in \mathbb{C}^K$ is a K-dimensional complex random vector. We use the standard assumption that X and D are independent. As a consequence X and D are also uncorrelated, i.e.

$$E[X(k,i)D(k,i)] = 0 \,\forall\, k,i. \tag{2.2}$$

Notice that the aforementioned assumptions on additivity (Eq. (2.1)) and independence of X and D are reasonable in a wide range of applications where speech is distorted by environmental noise.

We use the following notation with respect to real and imaginary parts, as well as the magnitude[1] of the random variables in question

$$Y(k,i) = Y_{\Re}(k,i) + jY_{\Im}(k,i),\ |Y(k,i)| = R(k,i), \tag{2.3}$$

[1] We will use the words *magnitude* and *amplitude* interchangeably. They mean the same, namely the absolute value of a complex DFT coefficient.

$$X(k,i) = X_{\Re}(k,i) + jX_{\Im}(k,i),\ |X(k,i)| = A(k,i), \tag{2.4}$$

and

$$D(k,i) = D_{\Re}(k,i) + jD_{\Im}(k,i), \tag{2.5}$$

where $j = \sqrt{-1}$, and where the subscripts $\Re$ and $\Im$ indicate the real and imaginary part of a DFT coefficient and where R and A denote the magnitude of the noisy DFT coefficient and the clean speech DFT coefficient, respectively.

It is of our interest to make an estimate $\hat{x}(k,i)$ of the clean speech DFT coefficient $x(k,i)$. In order to obtain $\hat{x}(k,i)$, Bayes estimators that optimize a certain cost function are often employed. In Section 2.2 we briefly discuss the topic of Bayes estimation. It turns out that, in general the estimate $\hat{x}(k,i)$ is a function of the noise variance $\sigma_D^2(k,i) = E\left\{|D(k,i)|^2\right\}$, the speech variance $\sigma_X^2(k,i) = E\left\{|X(k,i)|^2\right\}$ and the noisy DFT coefficient $y(k,i)$, that is

$$\hat{x}(k,i) = f(\sigma_D^2(k,i),\ \sigma_X^2(k,i),\ y(k,i)). \tag{2.6}$$

The speech and noise variance are also often referred to as the speech and noise power spectral density (PSD), respectively. Notice that the power spectral density of a process D is defined as $S_{DD}(k,i) = \frac{1}{K}E\left\{|D(k,i)|^2\right\}$, where K is the length of the time frame in samples and where the expectation operator $E\{\cdot\}$ is computed over the ensemble of the random process $D(k,i)$. For notational convenience we will leave out the scaling $\frac{1}{K}$ and use both the terms variance and PSD to denote $\sigma_D^2(k,i) = E\left\{|D(k,i)|^2\right\}$.

An alternative notation that is frequently used for Eq. (2.6) is in terms of the *a posteriori* SNR $\zeta(k,i)$ and the *a priori* SNR $\xi(k,i)$, that is

$$\hat{x}(k,i) = f(\zeta(k,i),\ \xi(k,i),\ y(k,i)).$$

The *a posteriori* and *a priori* SNR are defined in [1] as

$$\zeta(k,i) = \frac{r^2(k,i)}{\sigma_D^2(k,i)} \tag{2.7}$$

and

$$\xi(k,i) = \frac{\sigma_X^2(k,i)}{\sigma_D^2(k,i)}, \tag{2.8}$$

respectively. The *a posteriori* $\zeta(k,i)$ is dependent on the noisy magnitude realization $r(k,i)$ and the noise PSD $\sigma_D^2(k,i)$. The realization $r(k,i)$ is known and can be measured from the noisy data. The noise PSD is an expected value that is in general unknown and needs to be estimated. The *a priori* SNR, on the other hand, is completely defined in terms of expected values, which means that in practice besides $\sigma_D^2(k,i)$, also $\sigma_X^2(k,i)$ needs to be estimated. Several methods exist for estimation of the *a priori* SNR, given an estimate of the noise PSD $\sigma_D^2(k,i)$. The most popular one is the so-called decision-directed approach that we will discuss in Section 2.4.

2.2 Bayes Estimation

To facilitate the discussion in Chapters 6 and 7 on the derivation of speech enhancement estimators under a minimum mean-square error (MMSE) and a maximum *a posteriori* (MAP) criterion, respectively, we review in this section the so-called Bayes estimators.

Let U and V be two related random variables and assume that we are interested in an estimate of U, that is $\hat{U}(V)$, while we can only observe V, e.g. $V = U + W$, where W can be seen as an additive distortion. Let $c(U, \hat{U}(V))$ be a specific non-negative cost function and let $f_{V,U}(v, u)$ be the joint density of V and U, $f_V(v)$ and $f_U(u)$ its marginal densities, and $f_{U|V}(u|v)$ the conditional density of U given V. A Bayes estimator can then be defined as the estimator that minimizes the expected costs [2]

$$E\left\{c(U, \hat{U}(V))\right\}, \tag{2.9}$$

which is defined as

$$\begin{aligned} E\left\{c(U, \hat{U}(V))\right\} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(u, \hat{u}(v)) f_{V,U}(v, u) dv du & (2.10)\\ &= \int_{-\infty}^{\infty} I\left(\hat{u}\left(v\right)\right) f_V(v) dv, & (2.11) \end{aligned}$$

with

$$I(\hat{u}(v)) = \int_{-\infty}^{\infty} c(u, \hat{u}(v)) f_{U|V}(u|v) du. \tag{2.12}$$

Although $\hat{U}(V)$ is a function of V, we leave out V for notational convenience and simply write $\hat{U}$. In order to minimize $E\left\{c(U, \hat{U})\right\}$ it is sufficient to minimize Eq. (2.12), because $I(\hat{u})$ is non-negative and probability density functions, by definition, are non-negative. A cost function that is of specific interest for speech enhancement is the square-error cost function, that is

$$c(U, \hat{U}) = |U - \hat{U}|^2.$$

The estimator under this cost function is often referred to as the MMSE estimator and is found by minimization of

$$I(\hat{u}) = \int_{-\infty}^{\infty} |u - \hat{u}|^2 f_{U|V}(u|v) du. \tag{2.13}$$

The solution can be shown to be equal to (see e.g. [3])

$$\hat{u} = \int_{-\infty}^{\infty} u f_{U|V}(u|v) du = E\{U|V = v\}, \tag{2.14}$$

that is, $\hat{u}$ is the conditional mean estimator. Using Bayes' theorem we can express

Eq. (2.14) in a somewhat more convenient way, that is

$$\begin{aligned}\hat{u} &= \frac{\int_{-\infty}^{\infty} u f_{V|U}(v|u) f_U(u) du}{f_V(v)} \\ &= \frac{\int_{-\infty}^{\infty} u f_{V|U}(v|u) f_U(u) du}{\int_{-\infty}^{\infty} f_{V|U}(v|u) f_U(u) du}.\end{aligned} \tag{2.15}$$

Estimators under the MMSE criterion can not always be derived analytically. In order to avoid computationally overwhelming solutions (like numerical integration), the uniform cost function is also often used as an alternative for derivation of speech enhancement estimators. This cost function is defined as

$$c(U, \hat{U}) = \begin{cases} 0, & |U - \hat{U}| < \epsilon, \\ 1, & \text{otherwise}, \end{cases} \tag{2.16}$$

with ϵ an arbitrarily small positive number. The estimator under this cost function is found by minimization of

$$I(\hat{u}) = 1 - \int_{|u-\hat{u}|<\epsilon} f_{U|V}(u|v) du. \tag{2.17}$$

Because the integral in Eq. (2.17) is computed over an arbitrarily small region around $\hat{u}$, the estimate $\hat{u}$ is obtained by maximizing the density $f_{U|V}(u|v)$, i.e.

$$\hat{u} = \arg\max_u f_{U|V}(u|v),$$

that is, $\hat{u}$ is the *maximum a posteriori* (MAP) estimator. Using Bayes rule we can write this as

$$\hat{u} = \arg\max_u \frac{f_{V|U}(v|u) f_U(u)}{f_V(v)}. \tag{2.18}$$

Because the denominator is independent of u, it is sufficient to maximize the numerator, i.e.

$$\hat{u} = \arg\max_u f_{V|U}(v|u) f_U(u). \tag{2.19}$$

Although the uniform weighting of the costs as in Eq. (2.16) might be less relevant than the quadratic cost function, sometimes this cost function leads to somewhat simpler and analytically better feasible solutions.

From Eqs. (2.15) and (2.19) we see that in order to compute the MMSE and MAP estimator, respectively, the *prior* density $f_U(u)$ and the density $f_{V|U}(v|u)$ are needed. Depending on whether the goal is to estimate clean speech complex DFT coefficients or to estimate the magnitude of the clean speech DFT coefficients we thus need the prior densities $f_X(x)$ or $f_A(a)$, and the densities $f_{Y|X}(y|x)$ or $f_{R|A}(r|a)$, respectively, to compute the corresponding MMSE or MAP estimators.

2.3 Probability Distributions of Noise and Speech DFT Coefficients

Based on the central limit theorem [4] it is often argued that the probability density function $f_{D(k,i)}(d(k,i))$ of a noise DFT coefficient $D(k,i)$ is zero-mean Gaussian, as each noise DFT coefficient is computed as a sum of time samples. This is true when the frame size $K \rightarrow \infty$ and when the time-span of dependency between the time-domain samples in the frame is short compared to the frame size K [5]. Moreover, none of the variances of the individual time samples should dominate the variance of the sum of the time samples.

For many noise sources the time-span of dependency is relatively short and, as a consequence, the distribution of noise DFT coefficients is often close to Gaussian [5]. Also, in many practical situations the observed noise process can be decomposed into a sum of several independent noise processes, leading to a faster convergence of the distribution of noise DFT coefficients to a Gaussian distribution. For these reasons, we model the complex noise DFT coefficients with a complex Gaussian density, i.e. the real and imaginary parts of D are jointly Gaussian, that is

$$f_D(d) = \frac{1}{\pi\sigma_D^2} \exp\left\{-\frac{|d|^2}{\sigma_D^2}\right\}.$$

The density $f_{Y|X}(y|x)$ is therefore complex Gaussian and can be written as

$$f_{Y|X}(y|x) = \frac{1}{\pi\sigma_D^2} \exp\left\{-\frac{|y-x|^2}{\sigma_D^2}\right\}. \tag{2.20}$$

Let the polar representations of X and Y be defined as $X = A\exp(j\Phi)$ and $Y = R\exp(j\Theta)$. In order to derive an expression for the density $f_{R|A}(r|a)$ we first write Eq. (2.20) as

$$f_{Y|A,\Phi}(y|a,\phi) = \frac{1}{\pi\sigma_D^2} \exp\left\{-\frac{a^2+r^2-2ar\cos(\theta-\phi)}{\sigma_D^2}\right\}. \tag{2.21}$$

Transformation of (2.21) into polar coordinates and using Jacobian R then leads to

$$f_{R,\Theta|A,\Phi}(r,\theta|a,\phi) = \frac{r}{\pi\sigma_D^2} \exp\left\{-\frac{a^2+r^2-2ar\cos(\theta-\phi)}{\sigma_D^2}\right\}.$$

Integrating out the noisy phase θ then gives

$$\begin{aligned} f_{R|A,\Phi}(r|a,\phi) &= \int_0^{2\pi} f_{R,\Theta|A,\Phi}(r,\theta|a,\phi)d\theta \\ &= \frac{2r}{\sigma_D^2} \exp\left\{-\frac{a^2+r^2}{\sigma_D^2}\right\} \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right), \end{aligned}$$

where $\mathcal{I}_0$ is the 0th order modified Bessel function of the first kind [6]. Finally, the

density $f_{R|A}(r|a)$ is found by

$$f_{R|A}(r|a) = \int_0^{2\pi} f_{R|A,\phi}(r|a,\phi) f_{\Phi|A}(\phi|a) d\phi \quad (2.22)$$

$$= \frac{2r}{\sigma_D^2} \exp\left\{-\frac{a^2+r^2}{\sigma_D^2}\right\} \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right). \quad (2.23)$$

The expression in Eq. (2.23) has been derived in [7] as well by assuming a uniform distribution for the clean speech phase. However, notice that here we made no assumption about the clean speech phase distribution to derive this expression.

Speech DFT coefficients have been assumed Gaussian distributed as well [1]. However, measured histograms of speech DFT coefficients and speech DFT magnitude coefficients have shown that the speech DFT coefficients can be better modelled using more leptokurtic or super-Gaussian pdfs [8][9]. Super-gaussian pdfs have in general somewhat more heavy tails than the Gaussian density. There are several explanations that play a role in these observed non-Gaussian densities. The first explanation is related to the time-span of the dependency, which for speech is in general relatively long compared to the frame size. Therefore, the central limit theorem is not applicable. Secondly, histograms of speech DFT coefficients as presented in [8][9] are measured conditioned on speech spectral variances estimated by the decision-directed approach, which might be different from the distribution of speech DFT coefficients conditioned on the true, but unknown, spectral variance. Thirdly, a frame of speech data is often to some degree non-stationary. Even if the speech data were truly Gaussian, estimating the pdf over a non-stationary signal region would lead to a density that is not Gaussian.

To be better in line with the observed super-Gaussian densities for the speech DFT coefficients, we derive in Chapters 6 and 7 MMSE and MAP estimators under generalized Gamma and multivariate normal inverse Gaussian densities, respectively.

2.4 Estimation of the A Priori SNR

Most of the DFT-domain based clean speech estimators are defined in terms of the *a priori* SNR $\xi = \sigma_X^2/\sigma_D^2$. In practice ξ is unknown and has to be estimated from the noisy speech data.

A method that can be used to make an estimate of ξ, denoted by $\hat{\xi}$, is the so-called maximum likelihood (ML) estimator [1]. To derive this ML estimator, the pdf of a vector of noisy DFT coefficients is considered, that is $\mathbf{Y}(k,i) = [Y(k,i-L), ..., Y(k,i)]$. It is assumed that the elements in the vector are independent from each other and Gaussian distributed. Notice that in practice the DFT coefficients $Y(k,i-L), ..., Y(k,i)$ are often computed using overlapping time frames. This will violate the assumption that the elements in vector $\mathbf{Y}(k,i)$ are independent. Nevertheless, under the given assumptions, the pdf of $\mathbf{Y}(k,i)$ conditioned on σ_D^2 and σ_X^2 is given by

$$f_{\mathbf{Y}(k,i)|\sigma_D^2,\sigma_X^2}(\mathbf{y}(k,i)|\sigma_D^2, \sigma_X^2) = \prod_{l=0}^{L-1} \frac{1}{\pi(\sigma_D^2+\sigma_X^2)} \exp\left[-\frac{|y(k,i-l)|^2}{\sigma_D^2+\sigma_X^2}\right]. \quad (2.24)$$

Maximization of Eq. (2.24) with respect to σ_X^2 leads to

$$\sigma_X^2(k,i) = \frac{1}{L}\sum_{l=0}^{L-1} |y(k,i-l)|^2 - \sigma_D^2(k,i). \tag{2.25}$$

Dividing Eq. (2.25) by $\sigma_D^2(k,i)$ then leads to

$$\hat{\xi}(k,i) = \max\left[\frac{1}{L}\sum_{l=0}^{L-1}\zeta(k,i-l) - 1,\, 0\right], \tag{2.26}$$

where the maximum operator is applied to make sure that the estimated *a priori* SNR is non-negative. The estimate in Eq. (2.26) is in fact based on an average of the *a posteriori* SNR. The number of terms L that is used in Eq. (2.26) is a compromise between conflicting requirements. On one hand, L cannot be chosen too large, since speech signals can in general only be considered to be short-time stationary. On the other hand, larger L leads to more reduction of the variance of the estimate $\hat{\xi}(k,i)$. Notice, that evaluation of Eq. (2.26) also implies knowledge on the noise PSD σ_D^2. Estimation of σ_D^2 can be performed using e.g. voice activity detection (VAD) [10][11], minimum statistics [12][13][14] or by employing the in Chapter 8 discussed method based on DFT-domain subspaces.

Although this estimator of ξ is relatively simple, and relatively easy to analyze, it is not commonly used in combination with a clean speech estimator. The reason for this is that the maximum likelihood estimate of ξ leads in general to a relatively large amount of musical noise. This annoying effect is introduced because clean speech DFT estimators are applied independently per frame. However, small variations on the noisy DFT coefficients $y(k,i)$ due to the noise process lead to variations in the sequence of estimated *a priori* SNR values. As a consequence, a sequence of estimated clean speech DFT coefficients will show variations over time as well (even if the original sequence of clean speech DFT coefficients was completely constant). These variations give rise to the effect that is known as musical noise.

A method that reduces the effect of musical noise and which is commonly used for *a priori* SNR estimation is the so-called decision-directed approach. The decision-directed approach was originally defined in [1] as a linear combination between two equally valid definitions of the a priori SNR, that are

$$\xi(k,i) = \frac{E\left[|X(k,i)|^2\right]}{\sigma_D^2(k,i)}$$

and

$$\xi(k,i) = E\left[\zeta(k,i) - 1\right].$$

The linear combination leads to

$$\xi(k,i) = E\left[\alpha\frac{|X(k,i)|^2}{\sigma_D^2(k,i)} + (1-\alpha)\left[\zeta(k,i) - 1, 0\right]\right], \tag{2.27}$$

with $0 \leq \alpha \leq 1$. For implementation of Eq. (2.27) some approximations were needed; the expectation operator was neglected and realizations of the random variables in

question were used. Since the clean speech DFT coefficient $x(k,i)$ of the current frame i is unknown, the estimate of the previous frame was used instead, i.e. $\hat{x}(k,i-1)$, which can be obtained from the noisy DFT coefficient in the previous frame by applying a clean speech estimator. Altogether this led to [1]

$$\hat{\xi}(k,i) = \alpha \frac{|\hat{x}(k,i-1)|^2}{\sigma_D^2(k,i)} + (1-\alpha)\max[\zeta(k,i)-1,0]. \tag{2.28}$$

The parameter α determines how smooth the estimates $\hat{\xi}(k,i)$ will be across time and is therefore often called the smoothing factor. The closer α is to one, the more smooth the sequence of estimates will become. In return for this decrease in variance, the price to pay is a delay in the estimation of ξ. This effect is eminent during transitions, i.e. when there is a sudden increase of decrease in the true, but unknown, *a priori SNR*. In that case Eq. (2.28) will lead to overestimates or underestimates of ξ and as consequence to under- or oversuppression of the noise, respectively. This issue will be discussed in more detail in Chapter 4. Similar as for the ML approach, the decision-directed approach also assumes that knowledge of the noise PSD σ_D^2 is available. The decision-directed approach is often preferred over the ML estimate of $\xi(k,i)$, because of its ability to highly reduce the effects of musical noise.

2.5 Overview of DFT-Domain Based Estimators

In this section we give a brief overview of existing clean speech estimators for DFT-domain based speech enhancement. We will not provide here a complete historical overview, but will discuss the most relevant methods for speech enhancement. One of the first methods that was used for noise reduction in noisy speech signals was spectral subtraction [15][16]. This method aims at estimating the clean speech DFT magnitude by subtracting a smoothed noise magnitude from the noisy speech DFT magnitude. Subsequently, the estimated complex clean speech DFT is reconstructed by adding the noisy phase to the estimated clean speech magnitude. The concept of spectral subtraction comes in a lot of varieties. A rather general formulation of an estimator based on spectral subtraction is given by

$$\hat{X}(k,i) = \left(\max\left[1 - b\frac{E\left\{|D(k,i)|^a\right\}}{|Y(k,i)|^a}, 0 \right] \right)^{1/a} Y(k,i). \tag{2.29}$$

The parameter b determines the amount of subtraction, i.e. $b > 1$ will lead to an over subtraction of the noise and thus a somewhat more aggressive noise reduction, while $b < 1$ leads to an under subtraction of the noise and will lead to a higher noise floor. Parameter a determines the type of spectral subtraction that is applied. Some special choices for a are $a = 1$ and $a = 2$ for which we obtain magnitude spectral subtraction and power spectral subtraction, respectively.

Another well-known estimator that has been applied for noise reduction in noisy speech signals is the Wiener filter [17]. Under the assumption of large frame size K,

the Wiener filter can be implemented in the DFT domain as

$$\hat{X}(k,i) = \frac{\sigma_X^2(k,i)}{\sigma_Y^2(k,i)} Y(k,i). \tag{2.30}$$

Using the assumption that speech and noise are uncorrelated we can write Eq. (2.30) as

$$\hat{X}(k,i) = \frac{\sigma_Y^2(k,i) - \sigma_D^2(k,i)}{\sigma_Y^2(k,i)} Y(k,i). \tag{2.31}$$

In practice, estimates of the clean speech and noisy speech PSD are used in order to compute Eq. (2.31), that is

$$\hat{X}(k,i) = \frac{\max\left[\hat{\sigma}_Y^2(k,i) - \hat{\sigma}_D^2(k,i), 0\right]}{\hat{\sigma}_Y^2(k,i)} Y(k,i). \tag{2.32}$$

The maximum operator is used to make sure that the estimate of the clean speech PSD, i.e.

$$\hat{\sigma}_X^2(k,i) = \hat{\sigma}_Y^2(k,i) - \hat{\sigma}_D^2(k,i),$$

is always non-negative. Notice that this is not always guaranteed when using estimates $\hat{\sigma}_Y^2(k,i)$ and $\hat{\sigma}_D^2(k,i)$.

Alternatively, Eq. (2.30) is also often written in terms of the *a priori* SNR $\xi(k,i)$ as

$$\hat{X}(k,i) = \frac{\xi(k,i)}{\xi(k,i)+1} Y(k,i), \tag{2.33}$$

which can be obtained from Eq. (2.30) by dividing both numerator and denominator by $\sigma_D^2(k,i)$. Among the linear estimators, the Wiener filter is the best estimator in terms of mean-square error (MSE). When the clean speech and noise DFT coefficients, respectively, are both Gaussian distributed, the Wiener filter is also the optimal non-linear estimator.

In [1] an MMSE magnitude estimator was proposed under the same statistical model as for the Wiener filter, i.e. both the speech and noise DFT coefficients were assumed Gaussian distributed. This implies that the clean speech DFT magnitude was assumed to be Rayleigh distributed. The reason to consider a magnitude estimator instead of an estimator for the complex DFT coefficients was based on the argumentation that the phase of speech DFT coefficients is perceptually less relevant than the magnitude. The choice for complex DFT or DFT magnitude estimators depends on the preference for the type of error criterion that is used and might be application dependent.

However, the use of a Gaussian density to model speech DFT coefficients is debatable, as also mentioned in Section 2.3. In [8], the density of speech DFT coefficients has been thoroughly investigated. It was concluded by measuring histograms of speech DFT coefficients conditioned on *a priori* SNR values that are estimated using the decision-directed approach, that the observed density of speech DFT coefficients is more super-Gaussian. It is important to realize that the preference for these super-Gaussian densities is influenced by the conditioning on the *a priori* SNR and on the method that is used to estimate the *a priori* SNR. More specifically, in [18] estimation

of the *a priori* SNR was based on so-called GARCH models, leading to a preference for a Gaussian density to model the speech DFT coefficients.

Because of the observed super-Gaussian densities for speech DFT coefficients, there has been an increased interest over the last years to derive estimators for the clean speech DFT coefficients under these densities. Important contributions with respect to derivation of estimators under super-Gaussian densities can be found in [8][9]. In [8] MMSE estimators for the complex DFT coefficients are proposed under Laplace and Gamma densities. In [9] MAP magnitude estimators under super-Gaussian approximations are presented. In Chapter 6 we generalize the results presented in [8] by deriving estimators under the generalized Gamma density. In Chapter 7 we present MAP estimators under a different type of density, namely the MNIG density. This density can model heavy-tailed processes very well and has some potential advantages over the generalized Gamma density.

References

[1] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[2] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1992.

[3] H. L. van Trees. *Detection, Estimation and Modulation Theory*, volume 1. John Wiley and Sons, 1968.

[4] H. Stark and J.W. Woods. *Probability, random processes, and estimation theory for engineers*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[5] D. R. Brillinger. *Time Series: Data Analysis and Theory*. SIAM, Philadelphia, 2001.

[6] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New-York, ninth dover printing, tenth gpo printing edition, 1964.

[7] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(2):137–145, April 1980.

[8] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[9] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[10] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Lett.*, 6(1), 1999.

[11] J. Chang, N. S. Kim, and S. K. Mitra. Voice activity detection based on multiple statistical models. *IEEE Trans. Signal Processing*, 54(6), 2006.

[12] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. Eur. Signal Processing Conf.*, pages 1182–1185, 1994.

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

[14] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Processing*, 11(5):446–475, September 2003.

[15] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[16] J. S. Lim and A. V. Oppenheim. Enhancement and bandwidth compression of noisy speech. *Proc. of the IEEE*, 67(12):1586–1604, December 1979.

[17] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*. Principles of Electrical Engineering Series. MIT Press, 1949.

[18] I. Cohen. Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. *Signal Processing*, 86(4):698–709, 2006.

Chapter 3

Adaptive Time Segmentation for Improved Speech Enhancement

This chapter is based on the article published as "Adaptive Time Segmentation for Improved Speech Enhancement", by R. C. Hendriks, R. Heusdens and J. Jensen in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 14, no. 6, pages 2064 - 2074, Nov. 2006.

3.1 Introduction

In Chapters 1 and 2 it was mentioned that speech enhancement estimators at frequency bin k and time-frame i are typically expressed in terms of the noise power spectral density (PSD) $\sigma_D^2(k,i)$ and the noisy speech PSD $\sigma_Y^2(k,i)$. The gain function can be expressed in terms of $\sigma_D^2(k,i)$ and $\sigma_Y^2(k,i)$ directly, or indirectly using the definition of the *a priori* SNR $\xi(k,i)$, that is

$$\xi(k,i) = \frac{\sigma_X^2(k,i)}{\sigma_D^2(k,i)} = \frac{\sigma_Y^2(k,i) - \sigma_D^2(k,i)}{\sigma_D^2(k,i)} \tag{3.1}$$

with $\sigma_X^2(k,i)$ the clean speech PSD and where we used the commonly made assumption that the noise and the clean speech are statistically uncorrelated. Clearly, in order to compute the gain function, it is necessary to estimate the PSD of the noisy speech as well as the PSD of the noise process. We denote estimates of the noisy speech PSD by $\hat{\sigma}_Y^2(k,i)$ and estimates of the noise PSD by $\hat{\sigma}_D^2(k,i)$.

While the problem of estimating and tracking the noise PSD in the presence of speech has received significant interest recently, see e.g. [1][2], methods for accurate estimation of the noisy speech PSD appear to have been less explored. A well-known estimator of the noisy speech PSD is the periodogram, computed as $\hat{\sigma}_{Y,P}^2(k,i) = |Y(k,i)|^2$, where $Y(k,i)$ is a Fourier coefficient of noisy speech and the subscript P indicates that it is a periodogram estimate. However, the periodogram estimator suffers from a variance proportional to $\sigma_Y^4(k,i)$ [3]. To reduce the variance, smoothing methods like the Bartlett method [3] can be used. The Bartlett method computes an estimated (smoothed) PSD by averaging periodograms of, say N, uncorrelated frames, hereby decreasing the variance of the PSD estimate by a factor N [3]. When the frames are correlated, the variance is still reduced, but by a factor smaller than N. The decrease in variance comes with a side effect; the frequency resolution is decreased as well. In this chapter we use the Bartlett method to estimate the noisy PSD for a frame using the noisy data from a certain segment, a sequence of consecutive frames positioned around the frame to be enhanced. Fig. 3.1 illustrates the terms frame and segment. A Bartlett estimate of a noisy PSD for a frame is thus computed using a segment consisting of N frames including the frame to be enhanced.

In Boll's work on spectral subtraction [4], the Bartlett method was used across segments consisting of 3 frames located symmetrically around the frame to be enhanced. Although this leads to a decrease in variance, this approach has a number of disadvantages. First, the position of the segment with respect to the underlying noisy frame that needs to be enhanced is fixed. However, if the onset of a speech sound is not aligned with the start of the segment, the onset will be oversuppressed and the noise only region preceding the onset will be undersuppressed, because of a wrong estimate of the noisy PSD is used in the speech enhancement gain function. Secondly, ideally, the length of segments should vary with speech sounds; some vowel sounds may be considered stationary up to 40-50 ms, while stop consonants may be stationary for less than 5 ms [5]. A fixed segment size has two potential drawbacks. First, in signal regions which can be considered stationary for longer time than the segment used, the variance of the spectral estimator is unnecessarily large. Secondly, if the stationarity of the speech sound is shorter than this fixed segment size, smoothing is applied

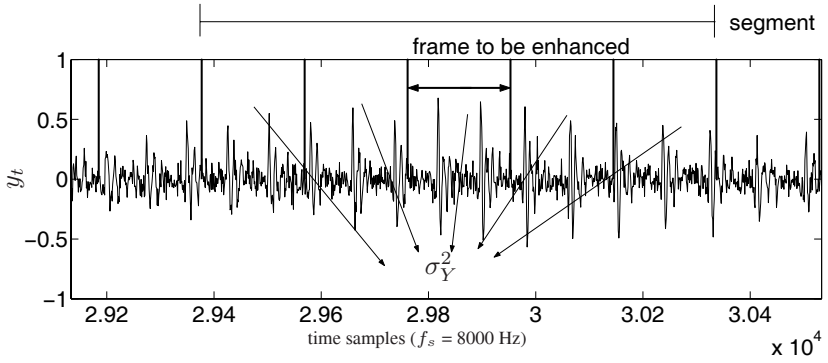

Figure 3.1: *Noisy speech signal with frame to be enhanced. In this example a segment consists of 5 consecutive frames.*

across stationarity boundaries resulting in oversuppression of transients, leading to a potential degradation of the speech intelligibility, and undersuppression of noise-only regions preceding those transients.

In this chapter, we propose an adaptive time segmentation for estimation of the noisy PSD to overcome the above mentioned problems. Notice that we keep the size and position of the frames fixed, but make the size and position of the segments adaptive. The proposed segmentation algorithm is very general. It can work as a front-end for most existing speech enhancement systems independently of the particular suppression rule that is used in the enhancement algorithm. To be more specific, the proposed method determines which noisy speech data should contribute in the estimation of the noisy speech PSD for a given frame, leading to better estimates of $\sigma_Y^2(k,i)$. The estimated $\sigma_Y^2(k,i)$ can then be used in, e.g. decision-directed (DD) approach [6] based or maximum likelihood [6] based schemes for estimating the *a priori* SNR $\xi(k,i)$.

The remainder of this chapter is organized as follows. In Section 3.2 we present an algorithm to determine an adaptive segmentation for speech enhancement. In Section 3.3 we show how this adaptive segmentation can be used to improve the estimation of the *a priori* SNR $\xi(k,i)$ using the DD approach. In Section 3.4 we evaluate the presented segmentation method by means of objective and subjective experiments. In Section 3.5 conclusions are drawn.

3.2 Adaptive Time Segmentation

To illustrate the impact of an adaptive segmentation within a speech enhancement context we compare time-domain waveforms of a noisy speech signal enhanced using an adaptive time segmentation with a scheme using a fixed segmentation. The noisy speech signal is constructed by degrading a clean speech signal by white Gaussian noise at an overall SNR of 10 dB. The enhanced signals are obtained using a Wiener filter where the noisy speech PSD was estimated using both a Bartlett estimate with

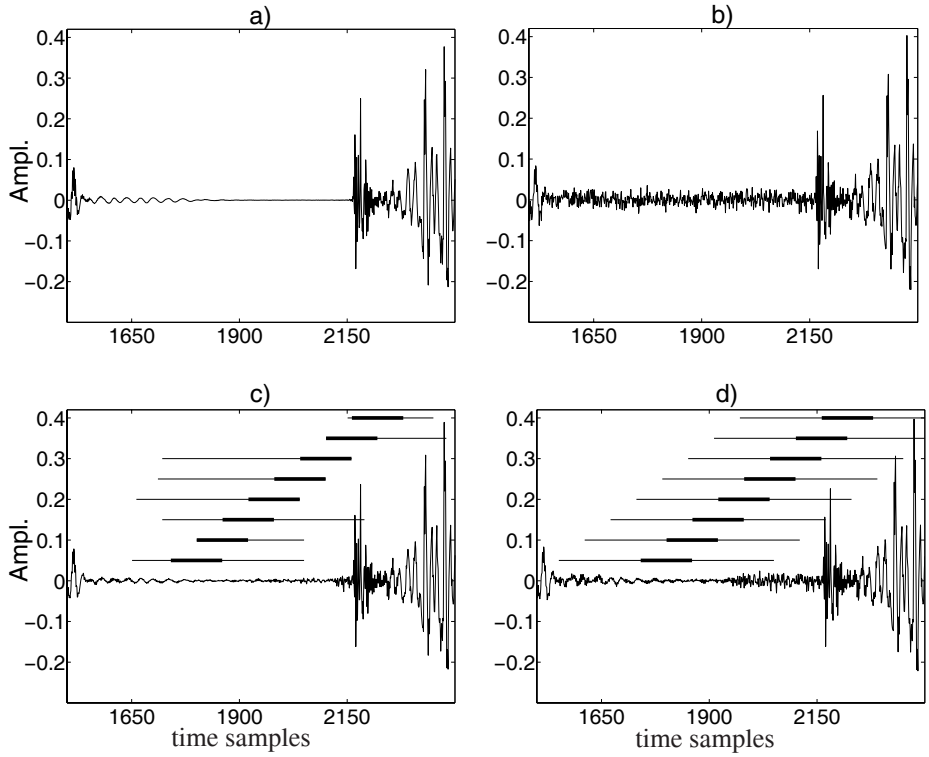

Figure 3.2: *a) Clean speech signal. b) Noisy speech signal with SNR of 10 dB. c) Enhanced noisy speech using an adaptive segmentation. d) Enhanced noisy speech using a fixed segmentation. (sample frequency is 8 kHz)*

adaptive time segmentation and a Bartlett estimate with fixed time segmentation. For ease of illustration, the adaptive segmentation was found here under ideal conditions (i.e. using the clean speech signal).

Fig. 3.2a shows the clean speech signal that contains a stop consonant. Fig. 3.2b shows the noisy speech signal. Fig. 3.2c shows the enhanced speech signal that was estimated using an adaptive time segmentation, whereas Fig. 3.2d shows the enhanced speech signal using a fixed segmentation. The thick lines mark the location of the signal frames, and the thin lines the segments that are used to estimate the noisy speech PSD for each frame. Comparing Fig. 3.2c and Fig. 3.2d it is clear that the fixed segmentation in Fig. 3.2d leads to a pre-echo present in front of the transient. This pre-echo is due to the fact that the PSD of the noisy signal, and consequently the gain value preceding the attack, is wrongly estimated and leads to undersuppression. With the adaptive segmentation in Fig. 3.2c the pre-echo is much reduced, because segment length and position of the segments are adapted to the speech signal.

Our goal in this section is to develop an adaptive segmentation algorithm that finds for each frame a corresponding segment containing noisy speech samples which can

be assumed stationary. To find an adaptive segmentation based on the noisy speech signal, we propose here a segmentation algorithm based on a probabilistic framework. The segment for a given frame is found based on the outcome of a sequence of hypothesis tests. We test the hypotheses whether two consecutive sequences of time-samples should be merged to form one segment. We regard the sequences as outcomes of random processes and search for sequences that can be considered stationary to a certain degree. Random variables will be denoted by capitals, whereas sample functions or realizations are denoted by small letters. In particular, we will use a test statistic based on a necessary condition for stationarity, namely that the zero-lag correlation coefficients of a random process $Y_t \equiv \{Y_t(m),\ m \in \mathbb{I}\}$, where $\mathbb{I}$ is an arbitrary index set, must remain invariant over time. Let E be the expectation operator. The correlation coefficient with lag 0 is defined as

$$C[0] = E\left\{|Y_t(m)|^2\right\}\ \forall\, m \in \mathbb{I}.$$

Let K be the frame size, P the frame shift and let us assume that within a frame the time samples are drawn from a wide sense stationary process. Let $\hat{C}^i[0]$ then be an estimator of $C[0]$ for frame i, defined by

$$\hat{C}^i[0] = \frac{1}{K}\sum_{m=1}^{K}|Y_t(m+iP)|^2.$$

Equivalently, using Parseval's identity [7], we can write this in the DFT domain as

$$\hat{C}^i[0] = \frac{1}{K}\sum_{k=1}^{K}|Y(k,i)|^2,$$

where $Y(k,i)$ is a random process representing DFT coefficients. The estimate $\hat{c}^i[0]$ of $\hat{C}^i[0]$ for a frame i can then be written as

$$\hat{c}^i[0] = \frac{1}{K}\sum_{m=1}^{K}|y_t(m+iP)|^2,$$

or, by again using Parseval's identity

$$\hat{c}^i[0] = \frac{1}{K}\sum_{k=1}^{K}|y(k,i)|^2, \tag{3.2}$$

where y_t and y denote realizations of the random processes Y_t and Y, respectively. Let $\mathbf{s}_1$ and $\mathbf{s}_2$ be two neighboring segments, which we assume to consist of statistically independent frames with frame numbers $i \in \{n, ..., n+n_0-1\}$ and $j \in \{n+n_0, ..., n+N-1\}$, respectively. Fig. 3.3 shows how $\mathbf{s}_1$ and $\mathbf{s}_2$ are defined. Let $\hat{c}_1^i[0]$ and $\hat{c}_2^j[0]$ denote estimates of $C[0]$ for frame i and j in segments $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively. We can interpret $\hat{c}_1^i[0]$ and $\hat{c}_2^j[0]$ as realizations of random variables $\hat{C}_1^i[0]$ and $\hat{C}_2^j[0]$, respectively. The two hypotheses then are:

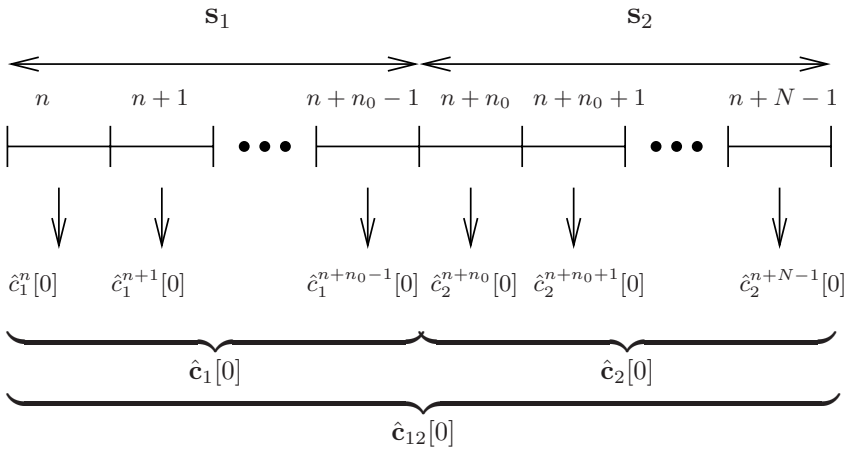

Figure 3.3: *Segments* $\mathbf{s}_1$ *and* $\mathbf{s}_2$ *with corresponding frames and estimates* $\hat{c}_1^i[0]$ *and* $\hat{c}_2^j[0]$.

$$\begin{aligned} H_0: &\quad \hat{C}_1[0] \text{ and } \hat{C}_2[0] \text{ have the same distribution} \\ &\quad (\,[\mathbf{s}_1, \mathbf{s}_2] \text{ is considered stationary}) \\ H_1: &\quad \hat{C}_1[0] \text{ and } \hat{C}_2[0] \text{ do not have the same distribution} \\ &\quad (\,[\mathbf{s}_1, \mathbf{s}_2] \text{ cannot be considered stationary}), \end{aligned} \tag{3.3}$$

where $[\mathbf{s}_1, \mathbf{s}_2]$ indicates the concatenation of $\mathbf{s}_1$ and $\mathbf{s}_2$. Let $\hat{\mathbf{c}}_1[0] \in \mathbb{R}^{n_0}$ and $\hat{\mathbf{c}}_2[0] \in \mathbb{R}^{N-n_0}$ be vectors containing n_0 realizations of $\hat{C}_1[0]$ and $N - n_0$ realizations of $\hat{C}_2[0]$, respectively, and let $\hat{\mathbf{c}}_{12}[0] = \left[\hat{\mathbf{c}}_1[0]^T, \hat{\mathbf{c}}_2[0]^T\right]^T \in \mathbb{R}^N$ be the concatenation of $\hat{\mathbf{c}}_1[0]$ and $\hat{\mathbf{c}}_2[0]$. Moreover, let λ_{th} be a decision threshold, and $f_{\hat{\mathbf{C}}_{12}[0]|H_0}(\hat{\mathbf{c}}_{12}[0]|H_0)$ and $f_{\hat{\mathbf{C}}_{12}[0]|H_1}(\hat{\mathbf{c}}_{12}[0]|H_1)$ the likelihood of observing the sequence $\hat{\mathbf{c}}_{12}[0]$ under hypothesis H_0 and H_1, respectively. The decision between the two hypotheses is then made by the following likelihood ratio test (LRT) [8],

$$\text{Reject } H_0 \text{ if} \qquad \frac{f_{\hat{\mathbf{C}}_{12}[0]}(\hat{\mathbf{c}}_{12}[0]|H_1)}{f_{\hat{\mathbf{C}}_{12}[0]}(\hat{\mathbf{c}}_{12}[0]|H_0)} > \lambda_{th}. \tag{3.4}$$

Assuming that the processes $\{\hat{C}_1^i[0],\ i = n, ..., n + n_0 - 1\}$ and $\{\hat{C}_2^j[0],\ j = n + n_0, ..., n + N - 1\}$ are iid, it follows that

$$f_{\hat{\mathbf{C}}_{12}[0]}(\hat{\mathbf{c}}_{12}[0]) = f_{\hat{\mathbf{C}}_1[0]}(\hat{\mathbf{c}}_1[0]) f_{\hat{\mathbf{C}}_2[0]}(\hat{\mathbf{c}}_2[0]), \tag{3.5}$$

$$f_{\hat{\mathbf{C}}_1[0]}(\hat{\mathbf{c}}_1[0]) = \prod_{i=n}^{n+n_0-1} f_{\hat{C}_1^i[0]}(\hat{c}_1^i[0]) \tag{3.6}$$

and

$$f_{\hat{\mathbf{C}}_2[0]}(\hat{\mathbf{c}}_2[0]) = \prod_{j=n+n_0}^{n+N-1} f_{\hat{C}_2^j[0]}(\hat{c}_2^j[0]). \tag{3.7}$$

Strictly speaking, the frames in Fig. 3.3 are assumed to be non-overlapping. However, in the experiments presented in Section 3.4 we used overlapping frames in order to increase the amount of data per segment, such that more data is available to estimate the parameters of the densities in (3.4).

We will argue in Section 3.2.1 that under certain assumptions, the pdfs $f_{\hat{C}_1^i[0]}$ and $f_{\hat{C}_2^i[0]}$ are Gaussian and that we therefore can use the standard procedure of the Generalized LRT [8], i.e., substitute unknown pdf parameters with their maximum likelihood estimates. In the case of Gaussian densities the unknown pdf parameters are the mean and variance. In Section 3.2.2 we use the derived densities for $\hat{C}_1^i[0]$ and $\hat{C}_2^i[0]$ to express the likelihood ratio into quantities that can be estimated from the noisy speech data. In Section 3.2.3 we present the algorithm that is used in combination with the LRT to find for each frame a corresponding segment.

3.2.1 Distribution of $\hat{C}[0]$

In this section we argue that it is reasonable to assume the pdf of $\hat{C}[0]$ to be Gaussian. To do so, we will reason that under certain conditions the DFT coefficients in a frame are independent from each other and then use the central limit theorem [9] to obtain the Gaussian density of $\hat{C}[0]$.

To derive the distribution of $\hat{C}[0]$, we use the aforementioned assumption that within a frame the time samples are drawn from a wide sense stationary process. Further, we assume the frames to be sufficiently long and assume that a K-dimensional vector (frame) of noisy time samples for frame i, that is $\mathbf{y}_t(i)$, is drawn from a K-dimensional multivariate Gaussian distribution, i.e. $\mathbf{Y}_t \sim N_K(0, \mathbf{C}_{Y_t})$, with $\mathbf{C}_{Y_t}$ a Toeplitz noisy speech correlation matrix. Let $\mathbf{F}$ be the DFT matrix. Then $\mathbf{Y} = \mathbf{F}\mathbf{Y}_t$ is distributed as $\mathbf{Y} \sim N_K(0, \mathbf{F}\mathbf{C}_{Y_t}\mathbf{F}^H) = N_K(0, \mathbf{C}_Y)$. Since the DFT matrix $\mathbf{F}$ is an asymptotic (in K) diagonalizer of any Toeplitz matrix, $\mathbf{C}_Y$ is asymptotically diagonal, which implies that $\mathbf{Y}$ is an uncorrelated multivariate Gaussian random vector and consequently also a vector of independent random variables. The estimator of $C^i[0]$,

$$\hat{C}^i[0] = \frac{1}{K}\sum_{k=1}^{K} |Y(k,i)|^2,$$

is therefore a sum of independent random variables. Using the central limit theorem [9] it then follows that $\hat{C}^i[0]$ approaches a Gaussian distribution for sufficiently large K.

In order to verify the density of $\hat{C}[0]$ we created a synthetic speech signal that was degraded by white noise at an SNR of 10 dB and measured the pdf of $\hat{C}^i[0]$ for this synthetic noisy signal. The synthetic speech signal was created by filtering an impulse train through a time-invariant LPC-synthesis filter whose coefficients were extracted from a natural speech signal. The pdf was measured by windowing the noisy speech data followed by computation of $\hat{C}^i[0]$ per windowed frame. The reason to use a synthetic speech signal is to be able to create a long, stationary sequence of speech with enough data to reliably estimate the density. In Fig. 3.4a the estimated pdf of $\hat{C}^i[0]$, based on the noisy synthetic speech signal shown in Fig. 3.4c, is compared to a

Gaussian distribution whose mean and variance are computed using $\hat{c}^i[0]$ values from the synthetic noisy speech data. It is shown that the measured pdf approximates the Gaussian distribution quite closely.

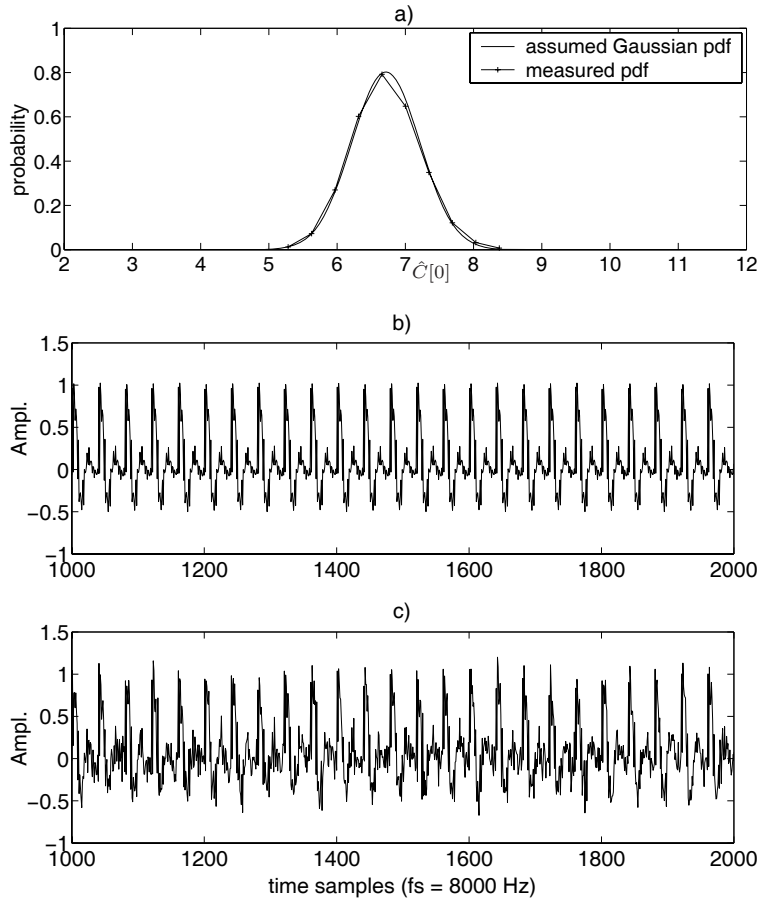

Figure 3.4: *a) Measured distribution of $\hat{C}[0]$ based on synthetic speech. b) Synthetic clean speech signal. c) noisy synthetic speech with SNR = 10 dB.*

3.2.2 Computation of the Likelihood Ratio

Using the argumentation from Section 3.2.1 it follows that $\hat{C}^i[0]$ can be assumed Gaussian distributed. The hypothesis H_0 and H_1 from Eq. (3.3) can now be expressed in terms of the Gaussian pdf. Let $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ and $\hat{\sigma}_{12}^2$ be maximum likelihood estimates of the variance of $\hat{C}[0]$ in $\mathbf{s}_1$, $\mathbf{s}_2$ and the concatenation of $\mathbf{s}_1$, and $\mathbf{s}_2$, respectively. Further, let $\hat{\mu}_1$, $\hat{\mu}_2$ and $\hat{\mu}_{12}$ be the corresponding estimates of the mean.

By substitution of the Gaussian densities from Eqs. (3.5)-(3.7) into Eq. (3.4) we

obtain the likelihood ratio

$$\frac{(2\pi\hat{\sigma}_1^2)^{-\frac{n_0}{2}} \exp\left[-\frac{\sum_{i=n}^{n+n_0-1}\left(\hat{c}[0]^i-\hat{\mu}_1\right)^2}{2\hat{\sigma}_1^2}\right] (2\pi\hat{\sigma}_2^2)^{-\frac{N-n_0}{2}} \exp\left[-\frac{\sum_{i=n+n_0}^{n+N-1}\left(\hat{c}[0]^i-\hat{\mu}_2\right)^2}{2\hat{\sigma}_2^2}\right]}{(2\pi\hat{\sigma}_{12}^2)^{-\frac{N}{2}} \exp\left[-\frac{\sum_{i=n}^{n+N-1}(\hat{c}[0]^i-\hat{\mu}_{12})^2}{2\hat{\sigma}_{12}^2}\right]}. \tag{3.8}$$

Substitution of maximum likelihood estimates of the model parameters

$$\begin{aligned} \hat{\mu}_1 &= \tfrac{1}{n_0}\textstyle\sum_{i=n}^{n+n_0-1}\hat{c}[0]^i \\ \hat{\mu}_2 &= \tfrac{1}{N-n_0}\textstyle\sum_{i=n+n_0}^{n+N-1}\hat{c}[0]^i \\ \hat{\mu}_{12} &= \tfrac{1}{N}\textstyle\sum_{i=n}^{n+N-1}\hat{c}[0]^i \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} \hat{\sigma}_1^2 &= \tfrac{1}{n_0}\textstyle\sum_{i=n}^{n+n_0-1}\left(\hat{c}[0]^i-\hat{\mu}_1\right)^2 \\ \hat{\sigma}_2^2 &= \tfrac{1}{N-n_0}\textstyle\sum_{i=n+n_0}^{n+N-1}\left(\hat{c}[0]^i-\hat{\mu}_2\right)^2 \\ \hat{\sigma}_{12}^2 &= \tfrac{1}{N}\textstyle\sum_{i=n}^{n+N-1}\left(\hat{c}[0]^i-\hat{\mu}_{12}\right)^2 \end{aligned} \tag{3.10}$$

then leads to the generalized LRT

$$\frac{\hat{\sigma}_{12}^N}{\hat{\sigma}_1^{n_0}\hat{\sigma}_2^{N-n_0}} > \lambda_{th}. \tag{3.11}$$

The likelihood ratio is compared to a fixed threshold λ_{th}. Alternatively, the LRT could also be applied using the Neyman-Pearson theorem, where the threshold λ_{th} is compared to a significance level η, also known as the false alarm probability $P(H_1|H_0)$. For derivations and more details on this relation we refer the reader to [10].

3.2.3 Segmentation Procedure

Given the LRT in Eq. (3.11) it is possible to find for a given frame a corresponding segment. To do so, we should in principle perform an exhaustive search over all possible segments. To avoid this computationally demanding full-search approach, we propose instead a computationally simpler algorithm. In [10] this computationally simpler segmentation algorithm was compared to an algorithm that uses the clean speech signal to obtain a segmentation that is optimal under the l_2-distortion measure between the clean speech signal and the clean speech estimate. It was shown that the difference in terms of segmental SNR between using this ideal segmentation and using the approach that we will follow here was on average smaller than 0.3 dB.

The procedure of the simplified algorithm is shown in Fig. 3.5. Initially, we start with a very short segment $\mathbf{s}_1$ that is assumed to be stationary and consists of multiple frames among which the frame to be enhanced, which is indicated in Fig. 3.5 by the shaded area. We form a similar short segment $\mathbf{s}_2$, which is positioned to the right of segment $\mathbf{s}_1$. Given the segments $\mathbf{s}_1$ and $\mathbf{s}_2$ we apply the LRT in Eq. (3.11). If the H_0 hypothesis is accepted we extend segment $\mathbf{s}_1$ with one frame from s_2 that is closest to segment $\mathbf{s}_1$. We then form a new segment $\mathbf{s}_2$, but now positioned to the left of the new extended segment s_1 and again apply the LRT in Eq. (3.11) and merge one frame

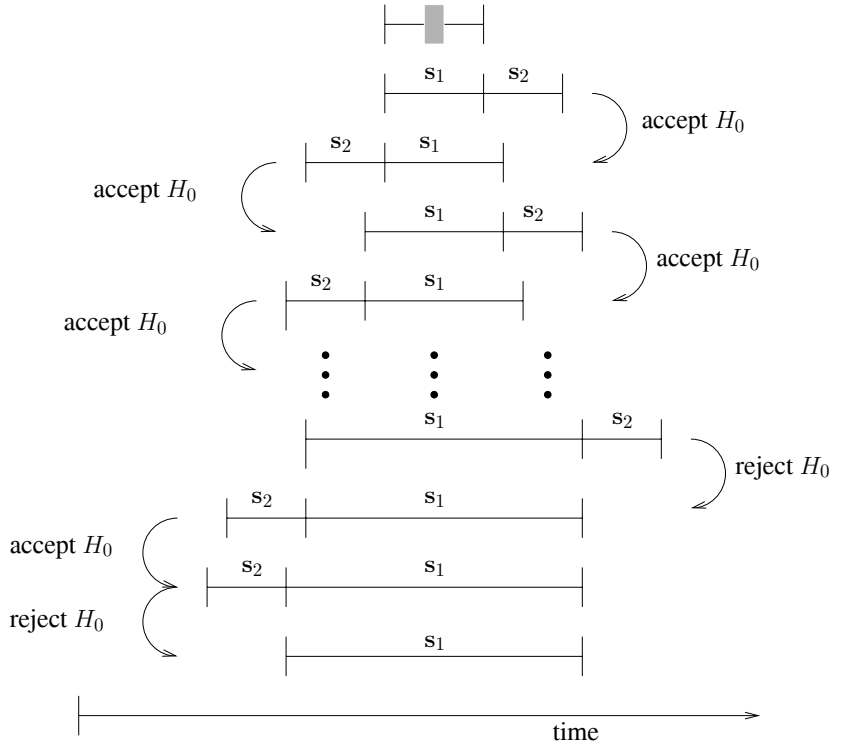

Figure 3.5: *Segmentation algorithm based on hypothesis.*

from segment $\mathbf{s}_2$ with segment $\mathbf{s}_1$ if the H_0 hypothesis is accepted. This procedure is iterated until on both the left and the right side of $\mathbf{s}_1$ the H_0 hypothesis is rejected. The final sequence $\mathbf{s}_1$, shown at the bottom of Fig. 3.5, is considered as the stationary segment that can be used in a Bartlett estimate of the noisy speech PSD. Note that the reason for having more frames in segment $\mathbf{s}_2$ than there are to be merged is the need for accurate estimates of the mean and variances in (3.9) and (3.10).

This segmentation algorithm can be generalized by dividing the frequency range in L sub-bands and determining a segmentation for each band independently. However, in this case less information is present per band to estimate maximum likelihood parameters of the Gaussian pdf. This, in turn, means that the variance of these estimates will be larger than in the full-band case. We expect that increasing the number of bands may be beneficial for a small number of bands, but for larger number of bands the advantage of having many bands may be overshadowed by the increased variance of the parameters of the Gaussian pdf that are estimated in each band.

Fig. 3.6 shows a block scheme of the proposed segmentation algorithm in combination with an enhancement algorithm, where we apply the adaptive segmentation procedure to different sub-bands. First the noisy signal $\mathbf{y}_t$ is divided into frames $\mathbf{y}_t(i)$, followed by a transform to the DFT domain, resulting in the DFT coefficients $y(1,i)$ up to $y(K,i)$. The whole frequency range is now divided into L sub-bands. Per sub-

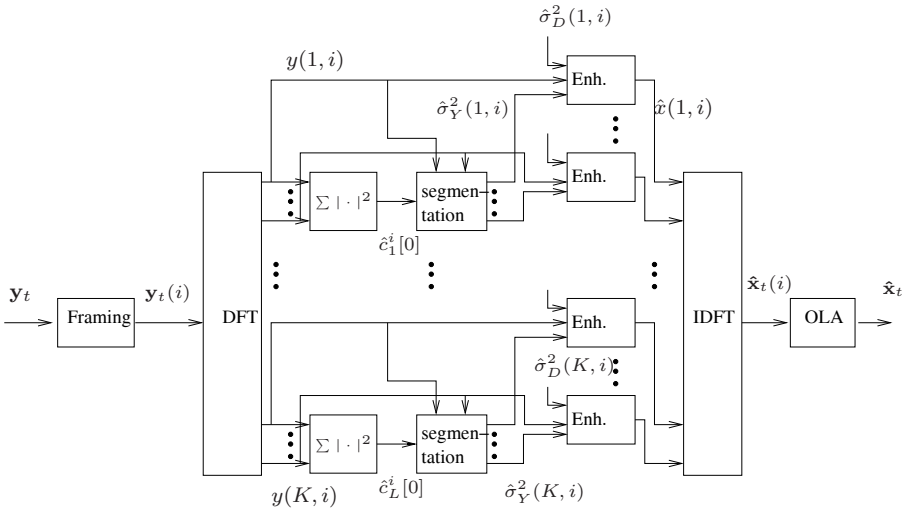

Figure 3.6: *Block diagram of adaptive segmentation speech enhancement system.*

band the correlation coefficients with lag zero are computed, that is, $\hat{c}_1^i[0]$ up to $\hat{c}_L^i[0]$. Per sub-band the segmentation is determined and used to compute $\hat{\sigma}_Y^2(k,i)$ for each frequency bin k in frame i. The clean speech DFT coefficients are then estimated in the block named *Enh.* for each frequency bin using the estimated noisy speech and noise PSDs, and the noisy speech DFT coefficient as input. The estimated clean speech DFT coefficients are transformed to the time domain using an inverse DFT followed by an overlap add resulting in the clean speech estimate $\hat{\mathbf{x}}_t$.

In Figs. 3.7 and 3.8 we show the result of the above described hypothesis based segmentation algorithm applied to two different speech signals degraded by white noise at an SNR of 15 dB and 5 dB, respectively. In these two examples we used for ease of visualization the full-band version of proposed segmentation algorithm. In the figures the original clean speech signal is shown together with the resulting segmentation. The thick lines mark the frames in which the signal is divided for enhancement. The thin lines represent for each frame the corresponding segment that is found by the hypothesis based algorithm. In Fig. 3.7, the speech signal under consideration consists of four parts: an initial silence part, a transient, some ringing after the transient and a voiced part. We see that frames in the silence and voiced part have long segments associated which cover the whole silence and voiced part, respectively. Frames in the transient part on the other hand have rather short segments, which prevents oversuppression of the transient. Further, the onset of the voiced part is resolved. In Fig. 3.8 the speech signal under consideration also consists of four parts. A voiced speech sound, a silence region, another voiced sound and again a silence region. In Fig. 3.8 we see that the frames in the two voiced regions have corresponding segments that completely cover the whole voiced region. The frames in the two silence regions have segments that cover the complete silence region.

Notice that the segments are found using future information, which implies a cer-

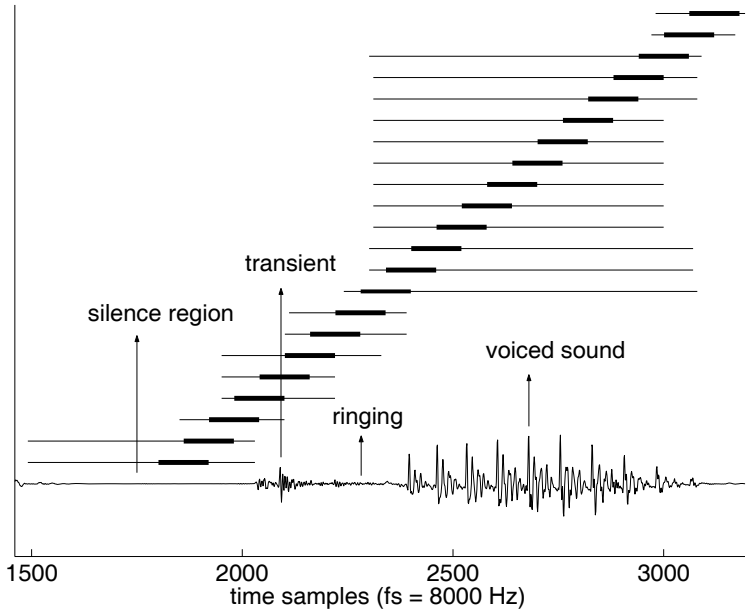

Figure 3.7: *Example Segmentation. Thick horizontal lines: Duration of frames. Thin horizontal lines: Corresponding segments. Input SNR = 15 dB. The shown signal is the original clean speech signal.*

tain latency or lookahead. However, the latency in the presented segmentation algorithm is adjustable. In Section 3.4 we demonstrate that also without or with limited latency, the use of an adaptive time segmentation leads to performance improvements.

3.3 A Priori SNR Estimation Using Adaptive Segmentation

As mentioned before, the Bartlett method reduces the variance of the estimate of σ_Y^2 by a factor N if N uncorrelated periodograms are averaged. In principle, the Bartlett estimate assumes no overlap and rectangularly windowed frames. However, the segments we find with the segmentation algorithm described in Section 3.2.3 may consist of overlapping frames and may be windowed using a non-rectangular window. Other methods, known as the Welch and Blackman-Tukey approach [3], are developed that do allow overlap and other windows than the rectangular window. A side-effect of increasing the overlap is that the decrease in variance will become smaller than a factor N.

The Bartlett estimate is computed by dividing a segment of M samples in frames of length K. The periodograms of these $N = \frac{M}{K}$ frames are then averaged. Note also

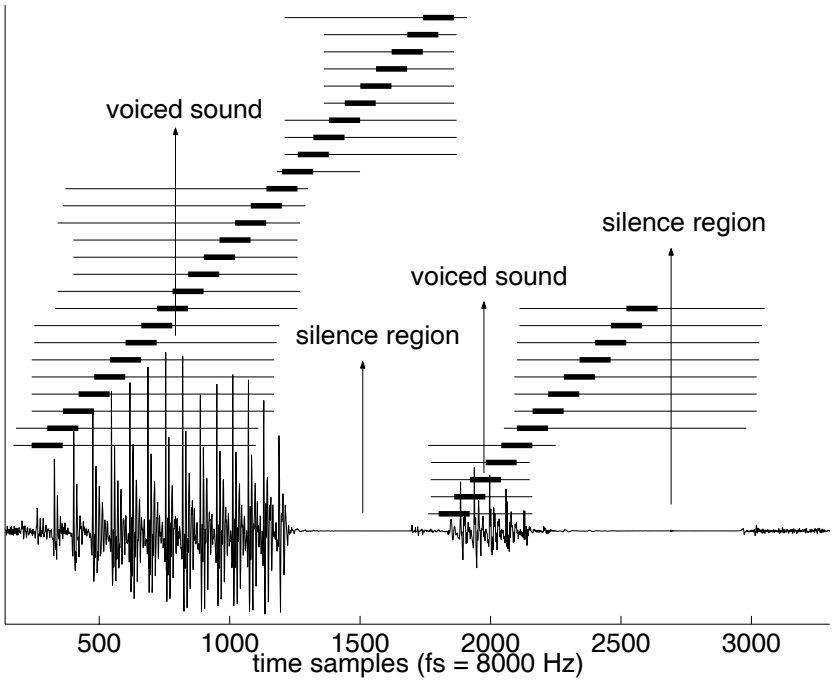

Figure 3.8: *Example Segmentation. Thick horizontal lines: Duration of frames. Thin horizontal lines: Corresponding segments. Input SNR = 5 dB. The shown signal is the original clean speech signal.*

that the decrease in variance comes with a side effect, namely that the frequency resolution of a periodogram based on a single frame is smaller than that of a periodogram based on the entire segment. Further, notice that the periodogram is asymptotically unbiased in the frame size K.

In conventional systems σ_Y^2 is estimated using a periodogram estimator [6], denoted by subscript P, i.e.

$$\hat{\sigma}_{Y,P}^2(k,i) = |Y(k,i)|^2, \tag{3.12}$$

or a Bartlett estimate with fixed segment length and fixed start and end positions $i-n$ and $i+n$, respectively [4][6], denoted by subscript B, i.e.

$$\hat{\sigma}_{Y,B}^2(k,i) = \frac{1}{2n+1} \sum_{j=i-n}^{i+n} \hat{\sigma}_{Y,P}^2(k,j). \tag{3.13}$$

The power spectral estimate can be improved by combining the Bartlett estimate with an adaptive segmentation. Let subscript A denote the use of adaptive time segmentation. The noisy speech PSD using the adaptive time segmentation can then be ex-

pressed as

$$\hat{\sigma}^2_{Y,A}(k,i) = \frac{1}{N} \sum_{j=i-n_1}^{i+n_2} \hat{\sigma}^2_{Y,P}(k,j), \tag{3.14}$$

with $i - n_1$ and $i + n_2$ denoting the start and end points of the segment with respect to frame number i and $N = n_2 + n_1 + 1$ the number of frames in the segment.

3.3.1 A Priori SNR Estimation Based on Improved Decision-Directed Approach.

The decision-directed approach [6], described in Section 2.4, is a well-known and often used method for estimation of the *a priori* SNR ξ, because it results in more natural residual noise than e.g. the maximum likelihood based scheme [4][6]. The decision-directed approach estimates ξ by taking a weighted average between two different estimates of ξ, namely

$$\hat{\xi}_1(k,i) = \frac{|\hat{x}(k,i-1)|^2}{\hat{\sigma}^2_D(k,i)}$$

and

$$\hat{\xi}_2(k,i) = \max\left[\frac{\hat{\sigma}^2_{Y,P}(k,i)}{\hat{\sigma}^2_D(k,i)} - 1, 0\right]. \tag{3.15}$$

See Section 2.4 for more details on the decision-directed approach. Instead of using estimate $\hat{\xi}_2(k,i)$ as in Eq. (3.15), that is based on the periodogram estimate we can use the PSD estimate based on the adaptive time segmentation defined in Eq. (3.14) as $\hat{\sigma}^2_{Y,A}(k,i)$. The decision-directed approach then becomes

$$\hat{\xi}(k,i) = \alpha \frac{|\hat{x}(k,i-1)|^2}{\hat{\sigma}^2_D(k,i)} + (1-\alpha)\max\left[\frac{\hat{\sigma}^2_{Y,A}(k,i)}{\hat{\sigma}^2_D(k,i)} - 1, 0\right]. \tag{3.16}$$

An advantage of (3.16) over the original decision-directed approach as specified in Eq. (2.28) is that the variance of the second term is decreased. Therefore, it is possible to decrease α, which means less influence of the first term in (3.16) and as a result a smaller tracking delay for the *a priori* SNR.

3.4 Objective and Subjective Simulation Experiments

We evaluate the presented adaptive time segmentation algorithm by means of objective and subjective simulation experiments. For objective evaluation we use SNR per frame, defined as

$$\text{SNR}(i) = 10\log_{10}\frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i) - \hat{\mathbf{x}}_t(i)\|^2}$$

and segmental SNR defined as [5]

$$\text{SNR}_{\text{seg}} = \frac{1}{N}\sum_{i=0}^{N-1} \text{SNR}(i).$$

In all experiments we use speech signals that originate from the Timit database [11] and that are re-sampled at 8 kHz. The frames have a size of 120 samples (15 ms) taken with 50% overlap. Noise statistics are measured during silence regions preceding speech activity and are assumed to be constant. In all experiments we use the Wiener filter as enhancement method and combine this with either the standard decision-directed approach [6] referred to as the DD approach or the decision-directed approach combined with adaptive time segmentation as presented in Eq. (3.16), referred to as the DDA approach.

The segmentation algorithm for the DDA approach follows the procedure in Fig. 3.5. All segments consist of frames that are taken with an overlap of approximately 90 %. The overlap is chosen in order to increase the amount of data per segment, such that more data is available to estimate the mean and the variances of the densities in Eq. (3.4), while still providing a good time resolution. For initialization of the segmentation algorithm both $\mathbf{s}_1$ and $\mathbf{s}_2$ consist of 5 (90% overlapping) frames. The threshold λ_{th} was chosen off line as $\lambda_{th} = 10^{7.5}$. This choice was based on experiments and led to a maximum average performance in terms of segmental SNR.

Objective Results

As a first experiment we study the influence of the number of sub-bands that is used in the segmentation algorithm on the speech enhancement performance. We expect a relation between the number of sub-bands that is used and the smoothing factor α in Eq. (3.16), because the larger α becomes, the smaller the influence of the second term in Eq. (3.16) becomes on the estimated *a priori* SNR. As a consequence, the influence of having multiple sub-bands on the performance will decrease. For this experiment we degraded speech signals by white noise at an SNR of 10 dB. For enhancement we use the DDA approach with $L = 1$, $L = 2$ and $L = 4$ sub-bands, respectively and compute the performance in terms of segmental SNR. Additionally we also compute the performance of the DD approach. In Fig. 3.9 we show the results of this experiment in terms of segmental SNR averaged over 6 different speakers versus α. We see indeed that the difference in terms of segmental SNR between using multiple sub-bands and using a full-band version of the segmentation algorithm decreases when α increases. For α in the range from 0.9 up to 1 we see that the difference between the several sub-band versions is negligible. Since this is the range of α values that is most interesting for speech enhancement we will use a full-band version of the segmentation algorithm in the following experiments. Notice, that $\alpha = 0$ in Fig. 3.9 corresponds to the maximum likelihood approach for *a priori* SNR estimation [6], for which it was also shown in [10] that an adaptive segmentation with multiple sub-bands leads to performance improvements.

As a second objective evaluation we compare the DD approach with the DDA approach in terms of segmental SNR as a function of the smoothing factor α, where we now focus on α-values in the range from 0.8 up to 1. The results are averaged over 6 different speakers, 3 male and 3 female. The speech signals are degraded by white noise at three different SNR levels, namely 5 dB, 10 dB and 15 dB. These results are shown in Fig. 3.10. From Fig. 3.10 it follows that combining the DD approach with an adaptive segmentation leads to an improved segmental SNR in the order of

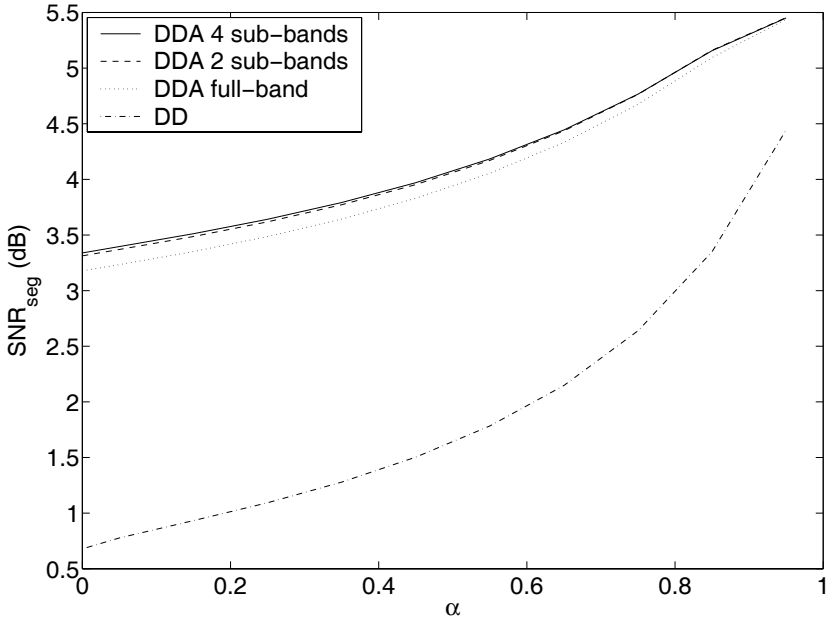

Figure 3.9: *Comparison between DD and DDA with* $L = 1$*,* $L = 2$ *and* $L = 4$ *sub-bands.*

approximately 0.7 dB. Further it can be seen that the DDA approach has its optimum at a lower α than the DD approach, which means less tracking delay in the estimation of ξ.

As a third evaluation we make a comparison between the DD approach and the DDA approach in terms of SNR per frame. To do so, a clean speech signal was degraded by white noise at an SNR of 15 dB. In this experiment we use for the smoothing factor in the DD approach $\alpha = 0.97$ as proposed in [6] for the Wiener filter. For the DDA approach we use $\alpha = 0.94$, which is based on optimal average performance in terms of segmental SNR. The comparison, which is depicted in Fig. 3.11, shows that the DDA approach generally leads to better performance: in onset regions, during sustained speech sounds and during silence intervals.

To demonstrate the influence of the latency, or lookahead, of the segmentation algorithm on the performance after enhancement we conducted an experiment as function of the maximum allowed latency. The signals in this experiment were degraded by white noise at an input SNR of 10 dB. During the experiments the maximum latency was limited to 0, 11, 24 and 36 ms and compared to the DD approach and the DDA approach with infinite latency. In Fig. 3.12 it is shown that the segmentation algorithm with a latency of approximately 36 ms has an almost similar performance as with an infinite latency. Even without a latency the use of the segmentation algorithm still leads to improvement compared to the DD approach without any adaptive segmentation. Notice that even when we do not use any lookahead, the segmentation

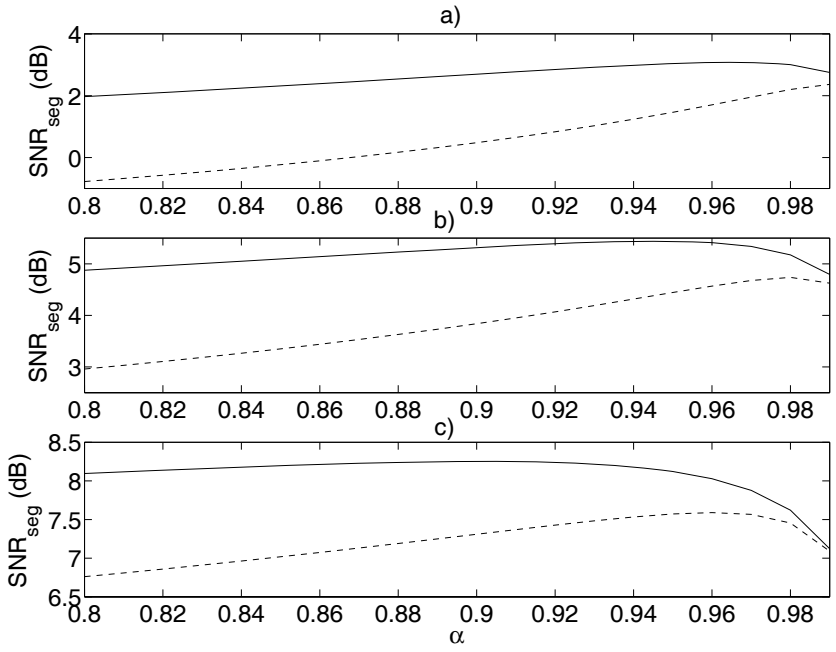

Figure 3.10: *Comparison between the standard DD approach (dashed) and the DDA approach (solid) in terms of segmental SNR as a function of the smoothing factor α. a) Input SNR of 5 dB ($SNR_{seg} = -8.23$ dB). b) Input SNR of 10 dB ($SNR_{seg} = -3.23$ dB). c) Input SNR of 15 dB ($SNR_{seg} = 1.77$ dB).*

algorithm still has access to frames from the past.

Residual Noise Analysis

Many enhancement methods suffer from a disturbing and unnatural sounding character of the residual noise. This is an important aspect of the quality of a speech enhancement algorithm and is often called *musical noise*.

In a simulation environment the residual noise signal can be computed by making a decomposition of the difference between the clean speech and its estimate into a speech distortion component and a noise residual component [12], that is

$$\begin{aligned} x(k,i) - \hat{x}(k,i) &= x(k,i) - y(k,i)G(k,i) \\ &= x(k,i) - \{x(k,i) + d(k,i)\}\, G(k,i) \\ &= \underbrace{x(k,i)\,\{1 - G(k,i)\}}_{\text{speech distortions}} - \underbrace{d(k,i)G(k,i)}_{\text{noise residual}}, \end{aligned}$$

where $x(k,i)$, $y(k,i)$ and $d(k,i)$ are realizations of Fourier coefficients and $G(k,i)$ is the value of the gain function. The normalized energy of the residual noise is then computed as $|d(k,i)G(k,i)|^2/\sigma_D^2(k,i)$.

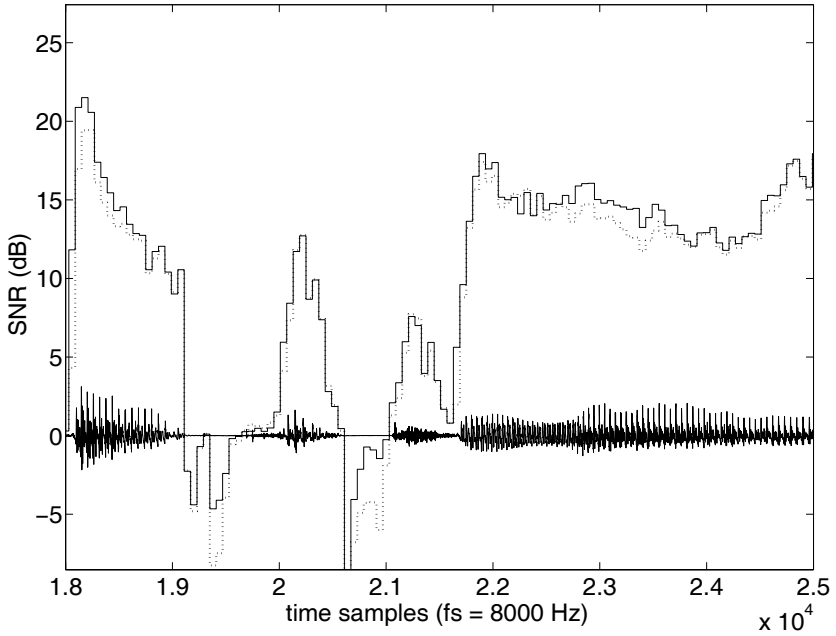

Figure 3.11: *SNR per frame after enhancement of noisy speech with 15 dB input SNR after using the DD approach (dotted) and the DDA approach (solid).*

We applied the above decomposition on a speech signal that was degraded by white noise at an SNR of 5 dB and compared the behavior of the residual noise between the DD and DDA approach, respectively. In Fig. 3.13a we show the time-domain waveform of the noisy speech signal. Fig. 3.13b shows the normalized energy of the noise residual for a typical frequency bin over several consecutive frames for both the DD (solid) approach and the DDA (dashed) approach. From Fig. 3.13b it is clear that the energy of the residual noise has a smoother character when using the DDA approach. With the DD approach the energy of the residual noise shows jumps and irregularities. Informal listening tests also confirmed that the DDA results in less residual noise with a less musical character. This is due to the decreased variance of the second term in Eq. (3.16).

Subjective Results

For subjective evaluation an OAB listening test was performed with 9 participants, the authors not included. Here, O is the original clean speech signal and A and B are two noisy signals that are enhanced using two different enhancement methods that we compare. The methods A and B are a Wiener filter where the *a priori SNR* was estimated with the DD approach and a Wiener filter where the *a priori SNR* was estimated with the DDA approach. In this listening test we used three different types of additive noise at two different SNRs, namely, white noise, car noise and F16-cockpit noise at

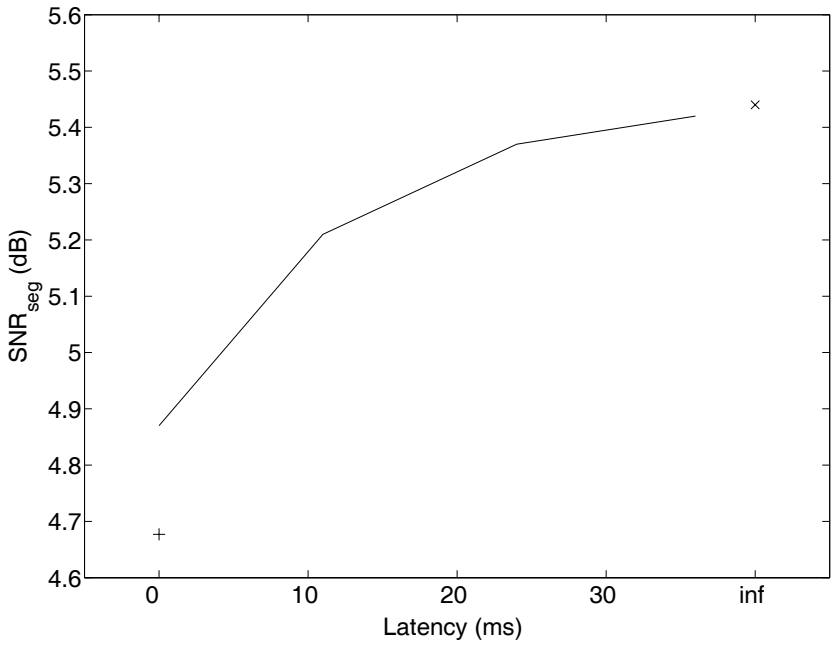

Figure 3.12: *Performance in terms of segmental SNR versus latency for the DD approach (+), DDA approach with finite latency (solid) and DDA approach with infinite latency (×).*

SNRs of 5 dB and 15 dB. Car noise and F16-cockpit noise originate from the Noisex-92 database [13]. For each noise type and noise power level we presented the listeners two female sentences and two male sentences. The listeners were presented first the original signal followed by the two different enhanced signals. The participants were asked to choose the signal that sounds best in comparison to the original. Each series was repeated 3 times with the enhanced signals being randomly assigned to the excerpts A and B. For speech signals corrupted with white noise at an SNR at 5 and 15 dB, the relative preference of the DDA approach over the DD approach was 80.6% and 70.4%, respectively. For speech signals corrupted with F16-cockpit noise at an SNR of 5 and 15 dB the DDA approach was preferred above the DD approach with 77.8% and 75%, respectively. A statistical Wilcoxon significance test revealed that the difference between the two methods is indeed significant at a significance level of $5 \cdot 10^{-3}$. The P-values of this test are tabulated in Table 3.1.

For speech signals corrupted by car noise the outcome of the listening test was close to 50%. In this case the statistical significance test (Wilcoxon test) was applied and revealed that the difference between the two methods indeed is insignificant, although objective tests done by the authors showed improvement in terms of SNR. This result can be explained by the fact that the energy of car noise is concentrated mainly in a small frequency band where in general the majority of speech energy is present. This means that most of the residual noise that is left after using the DD approach will

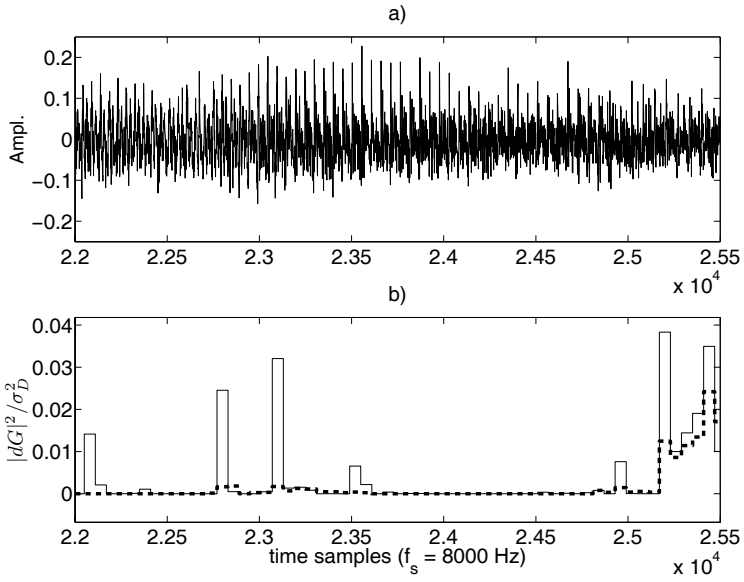

Figure 3.13: *a) Noisy speech signal at 5 dB SNR. b) Comparison between DD approach (solid) and the DDA approach (dashed) in terms of the normalized energy of the residual noise.*

be masked by the speech energy. As a result the perceptual difference between the DD approach and the DDA approach becomes smaller.

From the comments that were given by the participants of the listening test it was concluded that the main perceptual difference between the DD and DDA approach is the fact that the latter leads to a reduced level of musical noise.

3.5 Conclusions

We presented in this chapter an adaptive time segmentation for speech enhancement to improve estimation of the noisy speech PSD. The segmentation algorithm determines for each frame which segment of noisy data should be used to estimate the noisy speech PSD. The segments are formed based on the outcome of a sequence of hypothesis tests. We used this adaptive estimate of the noisy speech PSD to improve the decision-directed approach for speech enhancement methods. Objective experiments showed that usage of the adaptive time segmentation to improve decision-directed based speech enhancement leads to a better quality in terms of segmental SNR. Simulation experiments showed that the improved decision-directed approach results in less residual noise having a less musical character. Furthermore, subjective listening tests with speech signals degraded by various noise sources and noise levels showed that in terms of perceptual quality the decision-directed approach combined with the adaptive time segmentation algorithm is preferred over the usage of standard decision-directed

noise source	input SNR	P-value	significant
white noise	5 dB	$1.8 \cdot 10^{-5}$	yes
	15 dB	$3.53 \cdot 10^{-3}$	yes
car noise	5 dB	0.33	no
	15 dB	0.55	no
F16 noise	5 dB	$1.7 \cdot 10^{-4}$	yes
	15 dB	$1.5 \cdot 10^{-4}$	yes

Table 3.1: *Wilcoxon test results to determine the significance of the difference between the methods used in the listening experiment.*

approach.

References

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

[2] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Processing*, 11(5):446–475, September 2003.

[3] J. G. Proakis, C. M. Rader, F. Ling, C. L. Nikias, M. Moonen, and I. K. Proudler. *Algorithms for Statistical Signal Processing*. Prentice-Hall, Upper Saddle River, NJ, 2002.

[4] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[5] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

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

[7] H. Kwakernaak and R. Sivan. *Modern Signals and Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1991.

[8] H. L. van Trees. *Detection, Estimation and Modulation Theory*, volume 1. John Wiley and Sons, 1968.

[9] H. Stark and J.W. Woods. *Probability, random processes, and estimation theory for engineers*. Prentice-Hall, Englewood Cliffs, NJ, 1986.

[10] R. C. Hendriks, R. Heusdens, and J. Jensen. Adaptive time segmentation for improved speech enhancement. *IEEE Trans. Audio Speech and Language Processing*, 14(6):2064 – 2074, Nov. 2006.

[11] DARPA. Timit, Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc 1-1.1, October 1990.

[12] Y. Ephraim and H. L. van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Processing*, 3(4):251–266, July 1995.

[13] A. Varga and H. J. M. Steeneken. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–253, 1993.

Chapter 4

Forward-Backward Decision-Directed Approach for Speech Enhancement

This chapter is based on the article published as "Forward-Backward Decision Directed Approach for Speech Enhancement", by R. C. Hendriks, R. Heusdens and J. Jensen in the Proceedings of *Int. Workshop on Acoustic Echo and Noise Control*, pages 109-112, September 2005.

4.1 Introduction

Typically, DFT-domain based clean speech estimators can be written as a function of the *a priori* SNR ξ, see e.g. [1][2][3][4]. Because ξ is defined in terms of expected values, which are unknown in advance, estimation is necessary. Good estimation of ξ turns out to be crucial for the quality of the estimated clean speech signal [5][6]. In principle, there are two aspects related to *a priori* SNR estimation that influence the speech quality. The first aspect is related to the bias of the estimate $\hat{\xi}$ with respect to the true, but unknown, value of ξ. A biased estimate of ξ will lead to an overestimation or an underestimation of ξ and as a consequence to an undersuppression or oversuppression of the noise, respectively. A second aspect is the variance of $\hat{\xi}$, which has a huge impact on the perceptual quality of the estimated speech signal; the variance of $\hat{\xi}$ results in the introduction of musical noise. Often, musical noise is perceived as more annoying than the original noise. Several methods have been proposed for estimation of ξ. Among them are the maximum likelihood approach [1][7], which exploits properties of the Bartlett power spectral density estimate, and the decision-directed (DD) approach, presented in [1]. The DD approach has become quite popular for estimation of ξ, because in general it leads to less musical noise than the maximum likelihood approach.

The DD approach is defined as a weighted average between two different estimates of ξ, namely

$$\hat{\xi}_1(k,i) = \frac{|\hat{x}(k,i-1)|^2}{\hat{\sigma}_D^2(k,i)}$$

and

$$\hat{\xi}_2(k,i) = \max\left[\frac{\hat{\sigma}_{Y,P}^2(k,i)}{\hat{\sigma}_D^2(k,i)} - 1, 0\right], \tag{4.1}$$

that is,

$$\hat{\xi}_F(k,i) = \alpha\frac{|\hat{x}(k,i-1)|^2}{\hat{\sigma}_D^2(k,i)} + (1-\alpha)\max\left[\frac{\hat{\sigma}_{Y,P}^2(k,1)}{\hat{\sigma}_D^2(k,i)} - 1, 0\right]. \tag{4.2}$$

We see that the first term in Eq. (4.2) depends on an estimate of the clean speech DFT magnitude from the previous frame. Therefore, we refer to the traditional DD approach as the *forward decision-directed (FDD) approach* and denote the *a priori* SNR ξ estimated by the FDD approach as $\hat{\xi}_F$. The second term in Eq. (4.2) is an instantaneous estimate of the *a priori* SNR.

The FDD approach and in particular its ability to eliminate musical noise has been studied in detail by Cappé in [5], where two important observations were made. First, it was observed that the FDD based estimate $\hat{\xi}_F(k,i)$ leads to a highly smoothed version of $\hat{\xi}_2(k,i)$ in low SNR regions, which leads to a reduction of musical noise. Secondly, it was observed that Eq. (4.2) implicitly assumes the true underlying *a priori* SNR to be roughly constant across any two consecutive frames. If the underlying *a priori* SNR changes abruptly, e.g. in transitional regions between speech sounds or at speech onsets and offsets, a tracking delay in the estimated *a priori* SNR results.

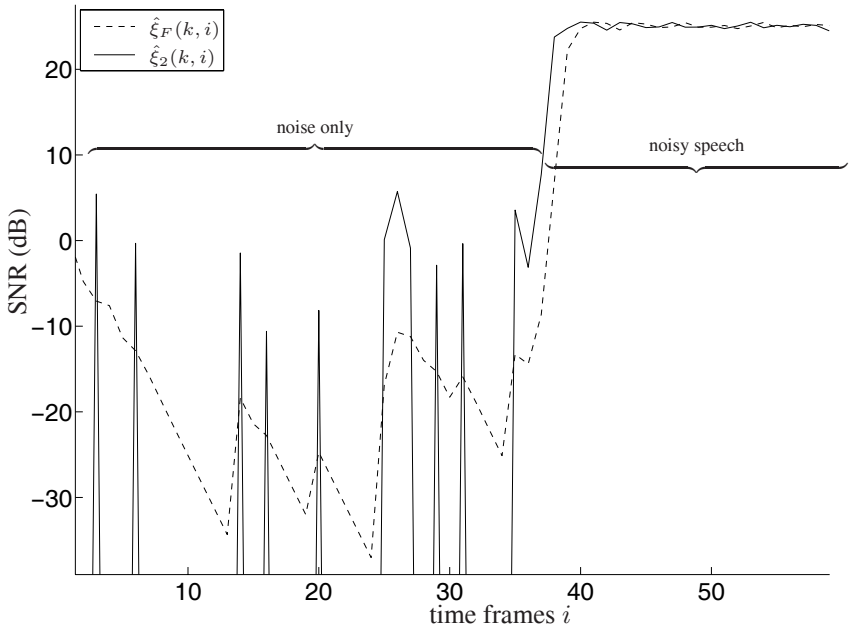

Figure 4.1: *Comparison between* $\hat{\xi}_F(k,i)$ *and* $\hat{\xi}_2(k,i)$.

To demonstrate these two influences of the FDD approach on the *a priori* SNR estimation we conducted a similar example for *a priori* SNR estimation as in [5]. In this example we created a synthetic noisy speech signal, consisting of a noise only region and a noisy voiced speech sound. The first 37 frames are noise only. Voiced speech is constructed by filtering a pulse-train through a time-invariant LPC-synthesis filter whose coefficients were extracted from a speech signal. We estimated ξ by $\hat{\xi}_F(k,i)$ using Eq. (4.2) with $\alpha = 0.98$ and plotted this together with the instantaneous SNR $\hat{\xi}_2(k,i)$ in Fig. 4.1 for a frequency bin containing a harmonic. From Fig. 4.1 we can see two effects. Firstly, in the first 37 frames, where the SNR is very low and the variance of $\hat{\xi}_2(k,i)$ rather large, the FDD based estimate $\hat{\xi}_F(k,i)$ leads to a rather smoothed version of $\hat{\xi}_2(k,i)$. Secondly, for higher SNRs, i.e. from frame 38 and further, we see that $\hat{\xi}_F(k,i)$ follows $\hat{\xi}_2(k,i)$ with one frame delay. Due to this delay, the estimated *a priori* SNR is biased directly after the transition. The closer the smoothing factor α is to one, the larger this delay becomes. The exact choice for α is a tradeoff; in stationary regions α should be close to one and in non-stationary regions a lower α should ideally be used. In a practical speech enhancement setup, the delay on the estimated *a priori* SNR $\hat{\xi}_F(k,i)$, will occur at each and every transition. Depending on the type of transition, i.e. a transition from low to high SNR or vise versa, $\hat{\xi}_F(k,i)$ will be underestimated or overestimated, respectively. As a consequence, the estimated gain function will be underestimated or overestimated,

respectively, as well.

In this chapter we present a *backward decision-directed* (BDD) approach to overcome over and underestimates of ξ at transitions when making use of the FDD approach. Instead of using the conventional order of time, we reverse the time index and make the estimate of ξ for the current frame dependent on clean speech DFT amplitude estimates from future frames. This implies the need for a (user-definable) algorithmic delay. We will show that this delay can be limited to several frames. Compared to the FDD approach, the BDD approach results in better estimates of ξ at the beginning of stationary regions and worse estimates of ξ at the end of stationary regions. By combining the BDD and FDD approach in a soft decision framework, a more efficient use of the noisy speech data is provided. This leads to better estimates of ξ at both the start and the end of stationary regions then when FDD is used alone. We will refer to this combination as the forward-backward DD (FBDD) approach. The combination between the estimates of ξ obtained by the BDD and the FDD approach is based on the time-adaptive segmentation algorithm for noisy speech, presented in Chapter 3.

4.2 The Backward Decision-Directed Approach

To motivate the use of the backward decision-directed approach we conducted an experiment, where we compute the true instantaneous SNR by $\frac{|x(k,i)|^2}{\hat{\sigma}_D^2(k,i)}$, and compare that with the *a priori* SNR that is estimated using the FDD approach. To do so, we created a piece-wise stationary signal consisting of a silence region, a synthetically created voiced speech region and again a silence region. Voiced speech is constructed by filtering a pulse-train through a time-invariant LPC-synthesis filter whose coefficients were extracted from a speech signal. This synthetic speech signal was degraded by white Gaussian noise.

The comparison between the true instantaneous SNR and the *a priori* SNR estimated by the FDD approach is shown in Fig. 4.2a for a representative frequency bin. During the start of the voiced sound and the second noise-only region it is shown that $\hat{\xi}_F$ lags behind the true instantaneous SNR for one or two frames. This is a typical behavior of the FDD approach as mentioned in the introduction. More specifically, the estimate $\hat{\xi}_F$ is always dependent on estimates from previous frames. However, the *a priori* SNR is not necessarily constant over time, i.e. the *a priori* SNR of the previous frame might be different from the current frame. The consequence then is a wrong estimate of $\xi(k,i)$, leading to an oversuppression or undersuppression of the noisy speech DFT coefficient, particularly in transitional signal regions. In stationary regions that are long enough this effect will die out after a couple of frames due to the first term in (4.2) that is weighted by the forgetting factor α.

Let us now consider a system where we reverse the processing order of frames, i.e. we make the estimate of $\xi(k,i)$ dependent on the estimate $|\hat{x}(k,i+1)|$ from the future frame. We then define the *backward decision-directed (BDD) approach* as

$$\hat{\xi}_B(k,i) = \alpha \frac{|\hat{x}(k,i+1)|^2}{\hat{\sigma}_D^2(k,i)} + (1-\alpha)\max\left[\frac{\hat{\sigma}_{Y,P}^2(k,i)}{\hat{\sigma}_D^2(k,i)} - 1, 0\right], \qquad (4.3)$$

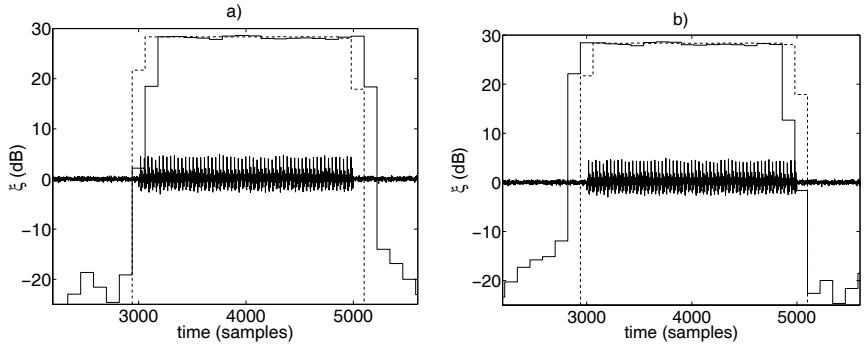

Figure 4.2: *Noisy speech signal, the true instantaneous SNR (dashed) and ξ estimated by a) FDD approach (solid) b) BDD approach (solid).*

where subscript B in $\hat{\xi}_B(k,i)$ denotes that this is the *a priori* SNR estimated using the BDD approach. From Eq. (4.3) we see that $\hat{\xi}_B(k,i)$ is dependent on future frames. A necessary assumption for implementation of (4.3) is an infinite delay. For now we will stick to this assumption, although later on we will show that this assumption can be weakened and that only a finite delay of a few frames is necessary. In Fig. 4.2b we consider the same example as before, but now ξ is estimated with the BDD approach. The estimate $\hat{\xi}_B$ at the start of stationary regions is now approximately equal to the true instantaneous SNR, while the bias in $\hat{\xi}_B$ is now present at the end of stationary regions.

To demonstrate that the difference in time dependency between FDD and BDD has an effect on the enhancement performance as well, we conducted a second experiment where a comparison is made between the FDD and the BDD approach. For this experiment we degraded a natural speech signal by white noise at an SNR of 10 dB and applied the MMSE amplitude estimator as proposed in [1] to estimate the clean speech DFT amplitudes. As is often done [1], the clean speech time frames are reconstructed by appending the noisy phase to the estimated clean speech DFT amplitudes followed by an inverse DFT. For *a priori* SNR estimation the BDD and FDD approach are both used with a smoothing factor $\alpha = 0.98$. The comparison is shown in Fig. 4.3 together with the original clean speech signal. The comparison is made in terms of the enhancement performance expressed by the SNR per frame.

From this example it is obvious that the FDD approach leads to better enhancement performance at the end of speech sounds, while the BDD approach leads to higher SNR values at the beginning of speech sounds. This difference between the FDD and BDD can be explained by the fact that the FDD approach estimates the *a priori* SNR using a clean speech DFT amplitude estimate from the previous frame. This means that it observes a transition at the beginning of stationary regions. The BDD approach makes an estimate of the *a priori* SNR using a clean speech DFT amplitude estimate from the next (future) frame, which means that it observes a transition at the end of stationary regions. This difference between the BDD and FDD approach suggests to combine them in such a way that the advantages of both methods can be exploited.

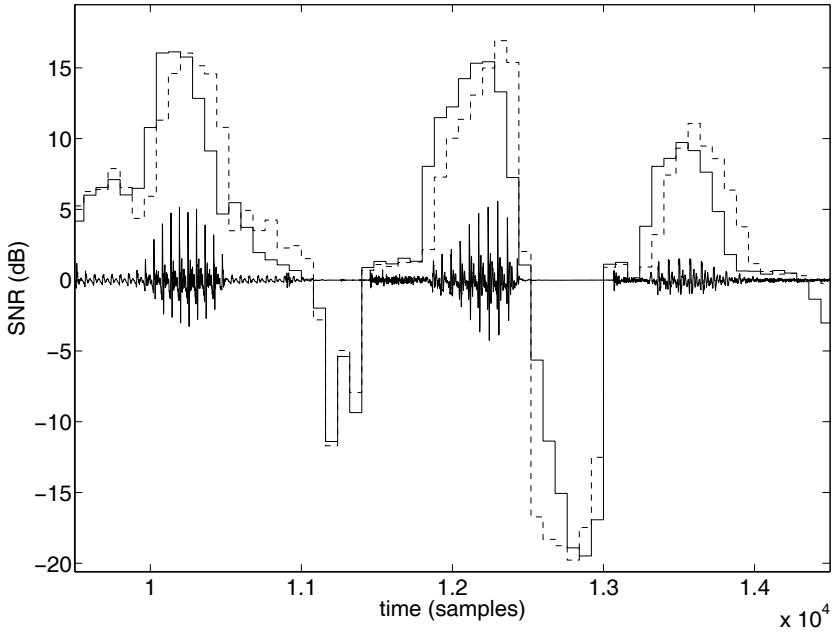

Figure 4.3: *Comparison between FDD approach (dashed) and BDD approach (solid) in terms of SNR.*

4.3 Forward-Backward Decision-Directed Approach

The examples given in Figs. 4.2 and 4.3 indicate that the preference for FDD or BDD depends on the position of the frame to be enhanced within a stationary region. We propose to combine the estimates $\hat{\xi}_F$ and $\hat{\xi}_B$ dependent on the position of the frame within the stationary region; at the beginning of a stationary region $\hat{\xi}_B$ is preferred, while at the end of stationary regions $\hat{\xi}_F$ offers a better alternative. To do so, let $N(i)$ be the length in samples of a stationary region in which time frame i is positioned and let $n_0(i)$ be the start-position in samples of the frame within the stationary region. To identify $N(i)$ we use the adaptive time-segmentation algorithm as described in the previous chapter. Fig. 4.4 illustrates the definition of $N(i)$ and $n_0(i)$ in a signal example. In this example the segment containing the stationary region is represented by the thin line and the frame to be enhanced is indicated by the thick line. Let $\beta\left(N(i), n_0(i)\right)$, $0 \leq \beta(N(i), n_0(i)) \leq 1$ be a weighting function. The estimates $\hat{\xi}_B$ and $\hat{\xi}_F$ can then be combined into a single estimate $\hat{\xi}_{FB}(k,i)$ by

$$\hat{\xi}_{FB}(k,i) \quad = \quad \hat{\xi}_F(k,i)\beta(N(i),n_0(i)) + \hat{\xi}_B(k,i)(1-\beta(N(i),n_0(i))). \quad (4.4)$$

In principle $\beta(N(i), n_0(i))$ can also be made frequency dependent. However, in this work we assume that $\beta\left(N(i), n_0(i)\right)$ is time-dependent only.

To study the behavior of the function $\beta(N(i), n_0(i))$, a training procedure was used where $\beta(N(i), n_0(i))$ was tabulated as a function of enhancement performance in terms of the SNR. To do so, β was sampled between 0 and 1. For each combination of β, N and n_0 the amount of improvement in terms of SNR, averaged over frames and frequencies, was computed using a training-set of 6 different speech signals. This leads to the average SNR improvement as a function of (β, N, n_0). The estimates $\hat{\xi}_B$ and $\hat{\xi}_F$ are then combined by selecting that β-value that leads for a given pair (N, n_0) to the largest SNR improvement based on the training data. It turned out that the value of $\beta(N(i), n_0(i))$ is especially important when $n_0(i)$ is positioned at the beginning of a stationary region (here $\hat{\xi}_B$ is typically a better estimate than $\hat{\xi}_F$, i.e. $\beta(N(i), n_0(i)) \approx 0$), and when $n_0(i)$ is close to the end of a stationary region, (here $\hat{\xi}_F$ is typically a better estimate than $\hat{\xi}_B$, i.e. $\beta(N(i), n_0(i)) \approx 1$). Based on these observations we define the following expression that fulfills these two requirements

$$\beta(N(i), n_0(i)) = \frac{1}{2}\left(\sin\left(1.5\pi + \pi\frac{n_0(i)-1}{N(i)-K}\right) + 1\right), \tag{4.5}$$

where K is the frame size. This function is also shown in Fig. 4.4, where it is indicated how β is chosen based on the values of $N(i)$ and $n_0(i)$. Experiments presented in [8], have shown that the use of Eq. (4.5) leads to enhancement performances in terms of segmental SNR that are as good as the training based procedure described above. Therefore, we use in the following sections Eq. (4.5) to combine $\hat{\xi}_F$ and $\hat{\xi}_B$. We will refer to this combination of the FDD and BDD approach as the FBDD approach.

4.3.1 Delay in the Backward Decision-Directed Approach

Until now, the BDD approach was assumed to run from the last frame, backwards in time to the first frame. As a consequence, an infinite lookahead is needed, which cannot be tolerated in many applications. As an alternative, we propose a method that limits this delay of the BDD approach to a user-defined delay of, say L, future frames. This method is based on an initialization of the BDD approach using the FDD approach. The procedure to do this is visualized in Fig. 4.5 for estimating the *a priori* SNR in frame i. First the FDD approach is run up to frame $i + L$ resulting in the estimate $\hat{\xi}_F(k, i + L)$. This estimate can then be used to compute an estimate of the clean speech DFT amplitude $|\hat{x}(k, i + L)|$ for frame $i + L$. The estimate $|\hat{x}(k, i + L)|$ can then be used to initialize the BDD approach that is run from frame $i+L$ backwards in time to frame i. This procedure limits the delay to L frames.

4.3.2 Iterative Forward-Backward DD Approach

The procedure described above can be further extended by an iterative procedure, such that the DD smoothing process is only applied within stationary regions. Both the FDD and the BDD approach are then run alternately across the frames in a stationary

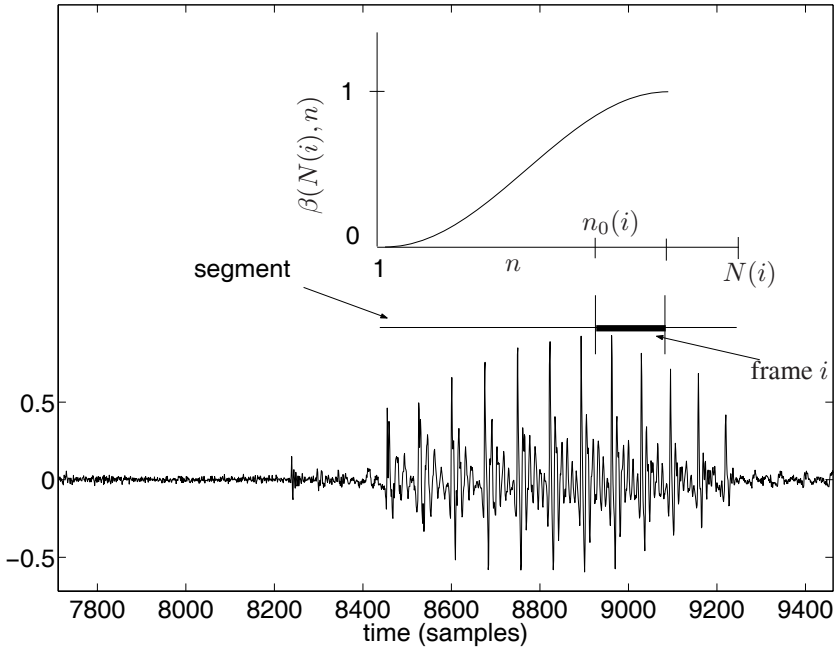

Figure 4.4: *Example of* $\beta(N(i), n_0(i))$ *selection.*

region that surrounds the frame to be enhanced. As mentioned before, these stationary regions can be identified with a segmentation algorithm presented in Chapter 3. The procedure for iterative forward-backward decision-directed (IFBDD) approach is illustrated in Fig. 4.6 for a setup with two iterations. First the FDD approach is run up to the frame with index $i + L$, where frame i is the frame to be enhanced. Then the BDD approach is initialized with the clean speech estimate based on the FDD approach and is run down to frame $i - L$. This ends the first iteration. Then the second iteration starts with the FDD approach, initialized with the clean speech DFT amplitude estimate based on the BDD approach of frame $i - L$. During this run, the estimate of the *a priori* SNR of frame i is then used as $\hat{\xi}_F$. Then as a final step the BDD approach is run for the last time, initialized with the clean speech estimate based on the FDD approach of frame $i + L$, until this run reaches frame i. The estimate of the *a priori* SNR of frame i is then used as $\hat{\xi}_B$, and so on. Experiments showed, however, that the number of iterations can be limited to two.

Simulation results confirmed that when the number of iterations is larger than 1, the difference between $\hat{\xi}_F$ and $\hat{\xi}_B$ is decreased and that the choice for β becomes less sensitive.

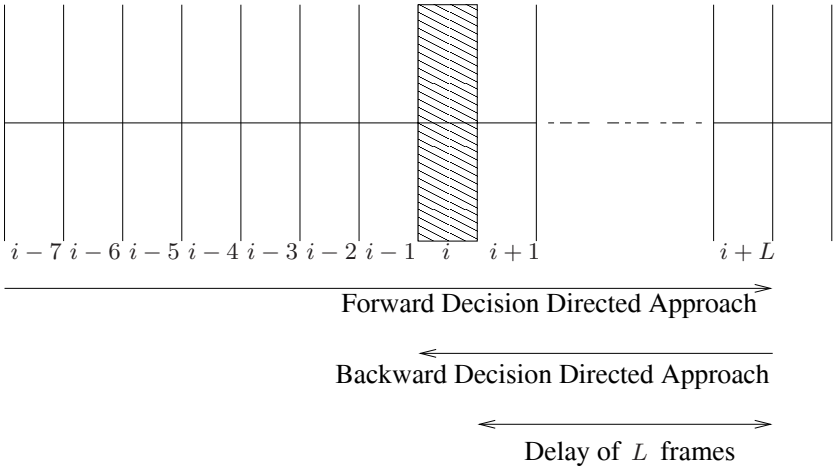

Figure 4.5: *Procedure to apply the BDD approach with limited delay.*

4.4 Experimental Results

In this section we evaluate the ideas presented in Section 4.2 and 4.3 by means of objective and subjective experiments. The speech signals that we use in the experiments originate from the Timit database [9] and are re-sampled at 8 kHz. All speech signals are degraded by white Gaussian noise. Noise statistics are measured during noise only regions, identified using an ideal voice activity detector, and are assumed to stay constant across time. The frame size is 160 samples and the frames are taken with 50 percent overlap. As enhancement method we use in all experiments the MMSE amplitude estimator as proposed in [1]. Furthermore, we use in all methods a smoothing factor $\alpha = 0.98$. Computation of N and n_0 is done per frame using the segmentation algorithm as presented in Chapter 3. For evaluation we use SNR per frame, defined as

$$\text{SNR}(i) = 10 \log_{10} \frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i) - \hat{\mathbf{x}}_t(i)\|^2}$$

where $\mathbf{x}_t(i)$ and $\hat{\mathbf{x}}_t(i)$ are vectors and denote frame i of the clean speech signal and the enhanced speech signal, respectively. Further, we use segmental SNR defined as [10]

$$\text{SNR}_{\text{seg}} = \frac{1}{N} \sum_{i=0}^{N-1} \text{SNR}(i),$$

and SNR per time-frequency point, that is

$$\text{SNR}(k,i) = 10 \log_{10} \frac{|x(k,i)|^2}{|x(k,i) - \hat{x}(k,i)|^2}.$$

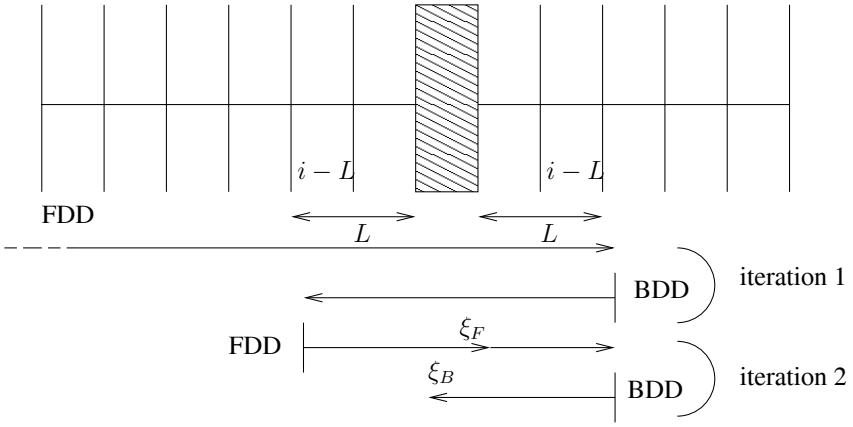

Figure 4.6: *Procedure for iterative combined forward-backward decision-directed approach.*

4.4.1 Objective Evaluation

Delay Limitation

As a first experiment we evaluate the influence of the delay limitation in the FBDD approach on the enhancement performance. In this experiment speech signals were degraded by white Gaussian noise at an SNR of 5 dB and enhanced using the framework proposed Section in 4.3.1 and illustrated in Fig. 4.5. The results are averaged over eight different speakers and expressed in terms of segmental SNR. The results are shown in Fig. 4.7 for the FBDD approach versus the delay L. From Fig. 4.7 it follows that a delay of two or three frames is already enough for good performance, and that using more delay even decreases the performance a little bit. This can be explained by the fact that for large L, the initialization of the BDD approach gets more dependent on estimated clean DFT amplitudes from different stationary regions in the future. Based on this experiment we use in the following experiments a delay of $L = 3$ frames.

IFBDD versus FDD

As a second experiment IFBDD and FDD are compared in terms of SNR over time. The signal under consideration originates from a female speaker degraded by white noise at an input SNR of 15 dB. The IFBDD approach was implemented as demonstrated in Fig. 4.6 using one iteration. Fig. 4.8 shows the comparison between IFBDD and FDD together with the clean speech signal. As expected, Fig. 4.8 demonstrates that the IFBDD approach performs better than the FDD approach, especially during the start of each speech sound, but also in more stationary regions.

To investigate in which parts of the spectrum the SNR improvements in Fig. 4.8 are obtained, we compute the difference in SNR per time-frequency point between the

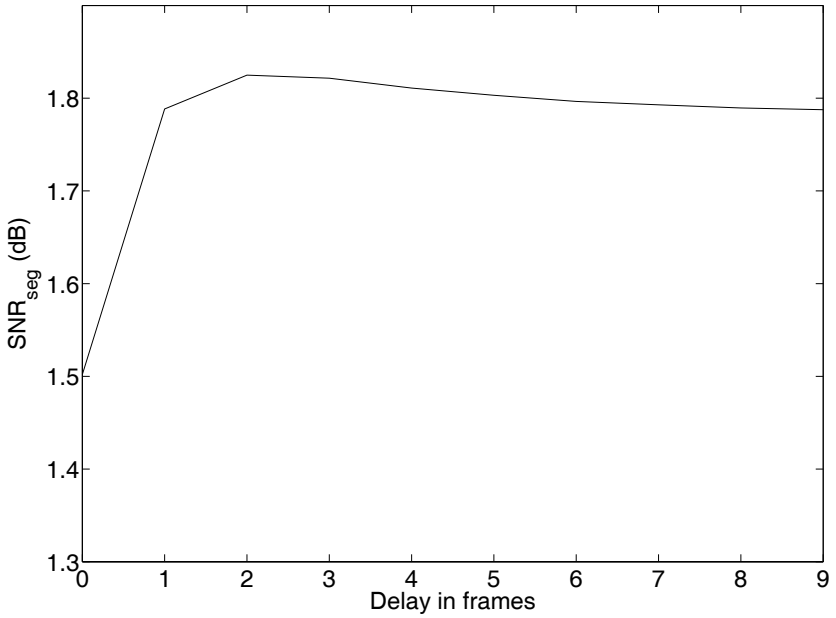

Figure 4.7: *Performance of the FBDD approach in terms of SNR_{seg} in dB versus adjusted delay in frames. Input SNR = 5 dB.*

IFBDD approach and the FDD approach, that is

$$\text{SNR}_{\text{diff}}(k,i) = \mathcal{T}\left\{\text{SNR}_{\text{IFBDD}}(k,i) - \text{SNR}_{\text{FDD}}(k,i)\right\},$$

where $\mathcal{T}\{\cdot\} = \min(\max(\cdot,-5),5)$ is a clipping function, which limits the dynamic range and is used for ease of visualization. In Fig. 4.9a we show the clean speech spectrogram and in Figs. 4.9b and 4.9c we show $\text{SNR}_{\text{diff}}(k,i)$ where IFBDD uses 1 and 2 iterations, respectively. The time-frequency points in Figs. 4.9b and 4.9c that are indicated by the gray colors have a rather small difference between the IFBDD and FDD approach. The time-frequency points indicated by the white color indicate positive differences of 5 dB and more. We see that the time-frequency points indicated with the white color occur mostly at the start of stationary regions (at both silence and speech regions). Moreover, we see that the white areas get slightly larger when going from 1 to 2 iterations.

To evaluate the influence on the enhancement performance of applying more iterations, a comparison is made between IFBDD as a function of the number of iterations, and the FDD approach. Moreover, in order to obtain a performance bound we also consider a method where we select β per frame such that the enhancement performance in terms of SNR per frame is optimal, i.e. we use the clean speech signal to decide the value of β. To do so, we sample β between 0 and 1. The value for β in a frame with number i is then chosen in an analysis-by-synthesis approach, such that the SNR after enhancement is optimal per frame. Notice, that due to the dependencies

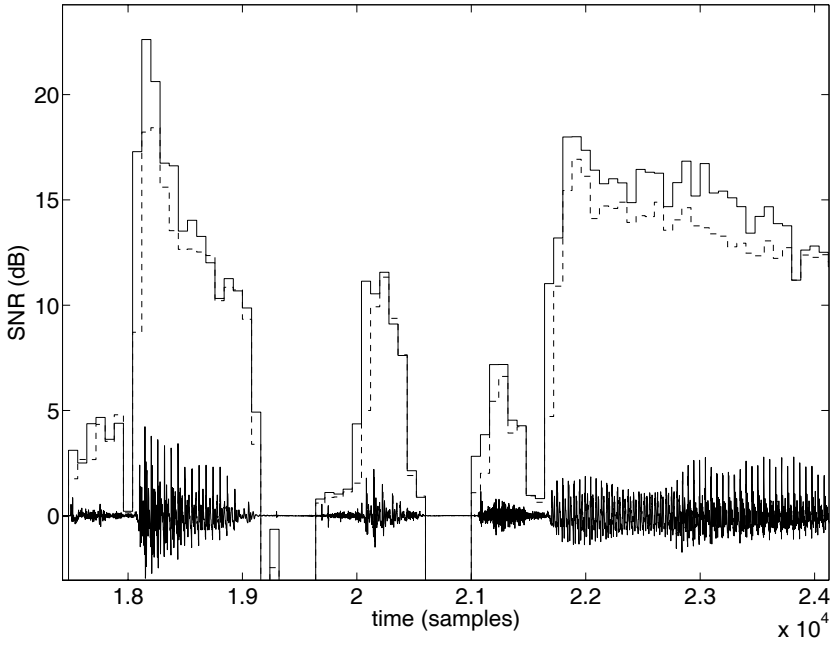

Figure 4.8: *Comparison between FDD (dashed) and IFBDD approach (solid) in terms of SNR per frame versus time with* $\alpha = 0.98$*, P=3, 1 iteration and input SNR of 15 dB.*

between frames that are introduced by the FDD and BDD approach this method is not guaranteed to be globally optimal.

For this comparison we degraded speech signals by white Gaussian noise at SNRs of 5 and 15 dB. The results, shown in Fig. 4.10, are expressed in terms of segmental SNR averaged over 8 speech signals. From Fig. 4.10 it follows that most improvement is gained when going from one to two iterations. Furthermore, when β is chosen optimally per frame by using the clean speech signal, an extra improvement of maximum 0.3 dB can be obtained. The improvement of IFBDD with two or more iterations over FDD is 0.75 dB and 1.1 dB for respectively input SNRs of 5 dB and 15 dB.

Similarly to Eq. (3.16) for the FDD approach, we can modify Eq. (4.3) such that instead of a periodogram estimate of the noisy PSD, an estimate $\hat{\sigma}^2_{Y,A}$ based on the adaptive time segmentation algorithm as discussed in Chapter 3 is used. With this combination an additional improvement of the segmental SNR of approximately 0.5 dB can be obtained. This improvement is mainly audible in term of more noise suppression. However, it results in somewhat more suppressed speech as well.

4.4.2 Subjective Evaluation

In this section we compare the perceptual quality difference between the FDD approach and the IFBDD approach. For this subjective performance evaluation an infor-

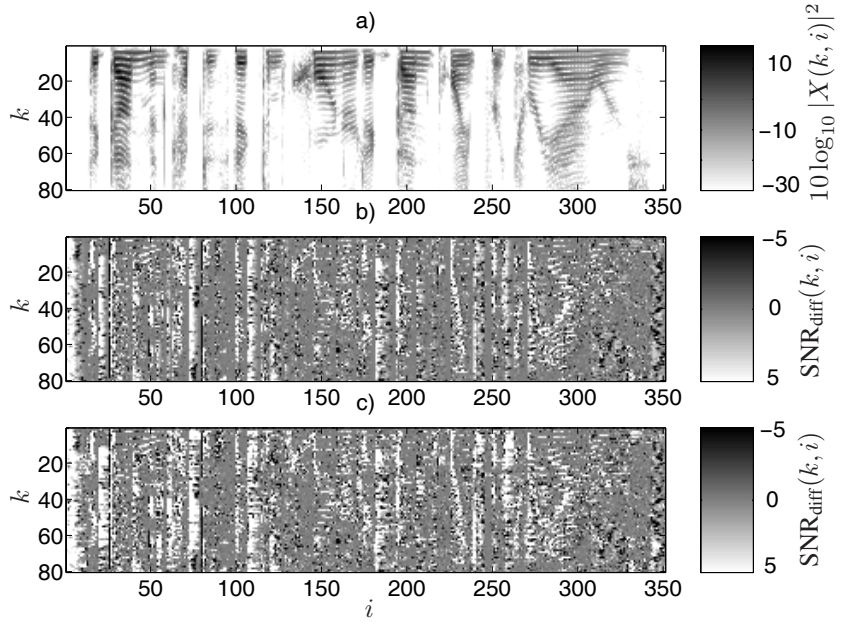

Figure 4.9: *a) Clean speech spectrogram. b) Improvement of IFBDD with 1 iteration over FDD in terms of SNR$_{diff}(k,i)$ c) Improvement of IFBDD with 2 iterations over FDD in terms of SNR$_{diff}(k,i)$.*

mal OAB listening test was performed with 6 participants, the authors not included. Here, O is the original clean speech signal and A and B are two noisy signals that are enhanced using the FDD approach and the IFBDD approach with 2 iterations. The listeners were presented first the original signal followed by the two different enhanced signals in randomized order. The participants were asked to choose the signal that sounds best in comparison to the original. Each series was repeated 4 times. In this listening test we used white noise at input SNRs of 15 dB and 5 dB. The relative preference of the IFBDD approach over FDD approach was 68% for both input SNRs of 5 dB and 15 dB. A statistical Wilcoxon signed rank test revealed that for both input SNRs the difference between the two methods was significant at a significance level of $p = 0.025$. The preference for IFBDD is mainly due to the better estimation of the *a priori* SNR, resulting in less suppressed speech at the start of stationary regions. Also, echo-like artifacts at the start of noise-only regions that are introduced when using the FDD approach are very much reduced due to a better estimation of the *a priori* SNR.

4.5 Conclusions

In this chapter a backward decision-directed (BDD) approach has been presented. This approach overcomes the introduction of distortions at the start of stationary regions and is based on a time-reversed processing order of frames. Consequently, estimation

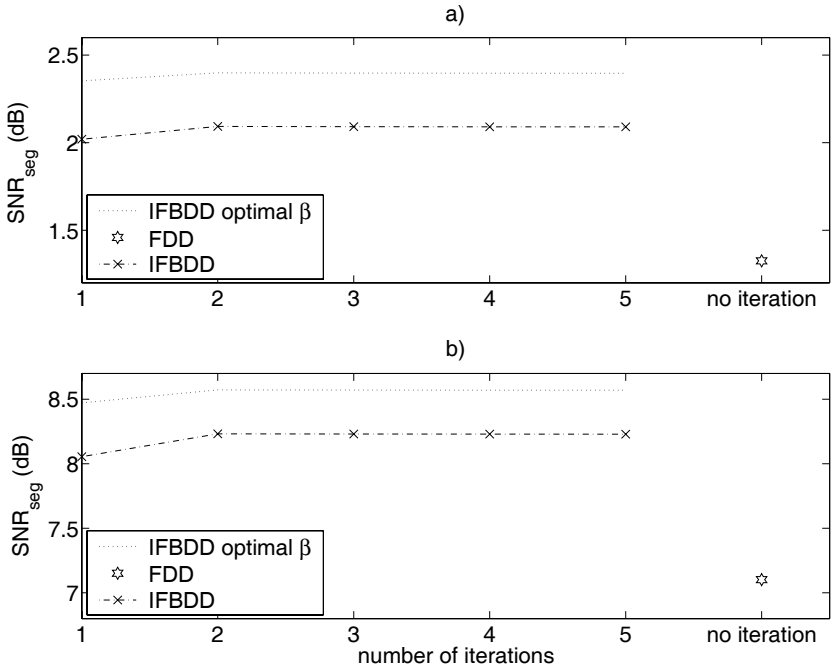

Figure 4.10: *Comparison in terms of segmental SNR between IFBDD with $L = 3$ frames delay versus the number of iterations and FDD. a) Input SNR = 5 dB. b) Input SNR = 15 dB.*

of the *a priori* SNR with BDD is dependent on future frames. A limited delay BDD approach is presented, which makes it possible to reduce the delay to a few frames. Using a soft-decision framework, the forward decision-directed (FDD) and BDD approach can be combined. This leads to less biased estimates of ξ at the beginning of stationary regions. Objective experiments where the proposed approach is compared with the FDD approach demonstrated improvements of more than 7 dB of local SNR and improvements of more than 0.75 dB and 1.1 dB average segmental SNR for input SNRs of 5 dB and 15 dB respectively. Informal listening tests show a statistical significant preference for the proposed method over the standard decision-directed approach.

References

[1] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[2] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[3] 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 Trans. Audio Speech and Language Processing*, 6:1741 – 1752, August 2007.

[4] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[5] O. Cappé. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Processing*, 2(2):345–349, April 1994.

[6] P. Scalarti and J. V. Filho. Speech enhancement based on a priori signal to noise estimation. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 2, pages 629–633, 1996.

[7] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

[8] R. C. Hendriks, R. Heusdens, and J. Jensen. Forward-backward decision directed approach for speech enhancement. In *Int. Workshop Acoustic Echo and Noise Control (IWAENC)*, pages 109–112, September 2005.

[9] DARPA. Timit, Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc 1-1.1, October 1990.

[10] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

Chapter 5

Speech Enhancement under a Stochastic-Deterministic Speech Model

This chapter is based on the article published as "An MMSE Estimator for Speech Enhancement under a Combined Stochastic-Deterministic Speech Model", by R. C. Hendriks, R. Heusdens and J. Jensen in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 15, no. 2, pages 406-415, Feb. 2007.

5.1 Introduction

As discussed in Chapter 2, most DFT-domain enhancement algorithms rely on stochastic signal models. However, it can be observed that certain speech sounds have a more deterministic character. For example, it is well-known that voiced speech segments may be represented well by a linear combination of sinusoidal functions with constant frequency and exponentially decaying amplitude or, as a special case, a constant amplitude [1, Ch. 4]. With this signal representation, the sequence of DFT coefficients seen across time for one particular frequency bin constitutes a completely deterministic time series. In [2] a maximum likelihood based spectral amplitude estimator was derived under a deterministic speech model. Here, the clean speech DFT coefficients are characterized by deterministic, but unknown amplitude and phase values, while the noise DFT coefficients are assumed to follow a zero-mean Gaussian pdf. This estimator leads to less suppression as compared to the case where speech DFT coefficients are assumed stochastic, see e.g. [3]. Obviously, a deterministic speech model is not always appropriate. For example, for noise-like speech sounds, such as /s/, /f/, etc. the DFT coefficients should rather be represented by a stochastic model.

Assuming that speech can not necessarily be modelled as either strictly stochastic or deterministic, we present in this chapter an MMSE clean speech estimator where the speech DFT coefficients are modelled as a mixture of a deterministic and a stochastic speech model.

The remainder of this chapter is organized as follows. In Section 5.2 we consider the individual deterministic and stochastic speech models and present their corresponding MMSE estimators. In Section 5.3 we specify the deterministic model and explain how to estimate its parameters. In Section 5.4 we derive the MMSE estimator under the combined stochastic-deterministic (SD) speech model. In Section 5.5 we present experimental results and finally in Section 5.6 we draw some conclusions.

5.2 The Stochastic and Deterministic Speech Model

In this section we introduce the stochastic and the deterministic speech model. We assume the noise process to be additive, i.e.

$$Y(k,i) = X(k,i) + D(k,i),$$

with $Y(k,i)$, $X(k,i)$ and $D(k,i)$ the noisy speech, clean speech and noise DFT coefficient, respectively at frequency bin k and time frame i. Further we assume that $X(k,i)$ and $D(k,i)$ are uncorrelated (for the stochastic model) and that the noise DFT coefficients have a zero-mean complex Gaussian distribution, as is argued for in Section 2.3.

By deriving an MMSE estimator under an SD speech model we exploit the idea that certain speech DFT coefficients can be better modelled with a deterministic model while others can be better modelled with a stochastic model. In the following derivations we use the complex zero-mean Gaussian distribution as stochastic representation for the clean speech DFT coefficients. However, we note that this work is general and

can also be extended to other distributions like the ones that are proposed in Chapters 6 and 7.

5.2.1 Probability Density Function of Noisy DFT Coefficients

In this subsection we consider the probability density functions of the noisy DFT coefficients under both the stochastic and the deterministic model, respectively. We introduce a random variable M that can take on the realizations $m \in \{sm,\, dm,\, am\}$. Here $m = am$, $m = dm$ and $m = sm$ indicate speech absence, that speech was generated with a deterministic model and that speech was generated with a stochastic model, respectively.

Stochastic Model

Assuming that clean speech DFT coefficients have a complex zero-mean Gaussian distribution, the noisy speech DFT coefficients have the following zero-mean complex Gaussian distribution

$$f_{Y|M}(y(k,i)|sm) = \frac{1}{\pi {\sigma_Y}^2(k,i)} \exp\left\{ -\frac{|y(k,i)|^2}{{\sigma_Y}^2(k,i)} \right\}, \tag{5.1}$$

where $\sigma_Y^2(k,i)$ is the variance of the noisy DFT coefficient $Y(k,i)$ which equals the sum of the noise variance and the clean speech variance, that is $\sigma_Y^2(k,i) = \sigma_X^2(k,i) + \sigma_D^2(k,i)$.

Deterministic Model

Under the deterministic speech model we assume that $Y(k,i)$ can be written as the sum of a deterministic variable (due to $X(k,i)$) and a stochastic variable (due to $D(k,i)$). Using the assumed (zero-mean) Gaussian distribution of the noise DFT coefficients this leads to a non-zero mean Gaussian distribution for the noisy DFT coefficients,

$$f_{Y|M}(y(k,i)|dm) = \frac{1}{\pi {\sigma_D}^2(k,i)} \exp\left\{ -\frac{|y(k,i) - E[Y(k,i)]|^2}{{\sigma_D}^2(k,i)} \right\}, \tag{5.2}$$

with $E[Y(k,i)] = x(k,i)$. Apart from having a non-zero mean, we note that the variance of $Y(k,i)$ under the deterministic model may be significantly smaller than that of $Y(k,i)$ under a stochastic model[1].

5.2.2 MMSE Estimators

In order to derive an MMSE estimator for the clean speech DFT coefficients under an SD speech model, we first consider the individual MMSE estimators for stochastic

[1]The variance of $Y(k,i) = X(k,i) + D(k,i)$ equals the sum of the individual variances of $X(k,i)$ and $D(k,i)$. Under the deterministic model the variance of $X(k,i)$ equals zero.

and deterministic representations.

Stochastic Model

Under the stochastic Gaussian speech model it is well known that the Wiener filter is the MMSE estimator, that is

$$\hat{x}(k,i) = E\left[X(k,i)|y(k,i)\right] = \frac{\xi(k,i)}{1+\xi(k,i)}y(k,i), \tag{5.3}$$

with $\xi(k,i) = \frac{E[X(k,i)^2]}{E[D(k,i)^2]} = \frac{\sigma_X^2(k,i)}{\sigma_D^2(k,i)}$ the *a priori* SNR.

Deterministic Model

Under the deterministic speech model, the clean speech DFT coefficients are assumed to be deterministic, but unknown. This means that $f_X(x(k,i)) = \delta(x(k,i) - x'(k,i))$ with $x'(k,i)$ the (unknown) value of the deterministic clean speech DFT coefficient itself and where $\delta(\cdot)$ is a delta function. The MMSE estimator then is

$$\hat{x}(k,i) = E[X(k,i)|y(k,i)] = x'(k,i), \tag{5.4}$$

where we observe that $x'(k,i) = E\left[Y(k,i)\right]$.

Notice that both estimators in (5.3) and (5.4) are expressed in terms of expected values. Since in practice these expected values are unknown, estimation is necessary. For estimation of $\xi(k,i)$, the decision-directed approach or maximum likelihood approach is often used [3]. Estimation of $E[Y(k,i)]$ will be considered in the next section.

5.3 Specification of the Deterministic Speech Model

So far we considered the use of a deterministic speech model, however we did not specify the exact model itself. A reasonable deterministic model for the clean speech signal is a representation by a sum of Q (exponentially damped) constant frequency sinusoids. Let $x_t(m)$ denote an arbitrary time-domain sample at time index m, a_q the amplitude, ϕ_q the phase, d_q the exponential decay factor and ν_q the frequency of component q. The clean speech signal can then be represented as

$$x_t(m) = \sum_{q=1}^{Q} a_q e^{j\phi_q} e^{(-d_q + j\nu_q)m}.$$

Using this model, the DFT coefficients at each frequency bin k can be described by a sum of Q complex exponentials seen across time. However, under the assumption of sufficiently long frame sizes, there will be no more than one dominant exponential, say component q, per frequency bin[2]. Let $w(m)$, $m = 0, \ldots, K-1$, be the analysis

[2] In principle, each frequency bin will have contributions due to smearing from complex exponentials at other frequency bins. However, increasing the frame-size will reduce the effect of smearing. Moreover, in speech signals the frequencies of the Q complex exponentials are in general spaced relatively far apart.

window (of length K) used to define the signal frame, P ($P \leq K$) the frame shift and $\omega_k = \frac{2\pi}{L}k$, where L is the DFT size ($L \geq K$) and n the index of a certain frame. Let us now assume that our deterministic model for a clean speech DFT coefficient $x(k,n)$ is indeed a single complex exponential, that is

$$x(k,n) = \sum_{m=0}^{K-1} a_q e^{j\phi_q} e^{(-d_q+j\nu_q)(m+nP)} w(m) e^{-j\omega_k m} \tag{5.5}$$

$$= e^{(-d_q+j\nu_q)nP} x(k,0). \tag{5.6}$$

We can write (5.6) in the form $x(k,n) = z^n x(k,0)$, with $z = e^{(-d_q+j\nu_q)P}$.

When the speech signal is degraded by noise that is wide sense stationary for $n = i-n_1, ..., i+n_2$ and if P is sufficiently large with respect to the time-span of dependency of the noise [4], then the noise that is observed in a sequence of DFT coefficients at frequency bin k and at time indices $n = i-n_1, ..., i+n_2$ is white, irrespective of the spectral color of the noise. Estimation of d_q and ν_q from a sequence of complex DFT coefficients is then known as a standard harmonic retrieval problem [5] and estimation of d_q and ν_q can be done from the noisy DFT coefficients using, for example, the ESPRIT algorithm [6][5].

Notice that for overlapping frames the observed noise in the sequence of DFT coefficients is not white. This might lead to wrong estimates of d_q and ν_q for very low SNRs. However, this is not necessarily a problem, since the deterministic model is mainly used at harmonics, which have in general quite a high SNR. Alternatively, one could use a whitening transform as discussed in Section 8.5.1.

When $n = i-n_1, ..., i+n_2$ is the time span across which we assume the deterministic model to be valid, then we can approximate Eq. (5.4) using the relation in Eq. (5.6). With Eq. (5.6) we correct for the exponential decay in amplitude and for the phase shift. The estimate $\hat{x}(k,i)$ becomes

$$\hat{x}(k,i) = E\left[Y(k,i)\right] \tag{5.7}$$

$$\approx \frac{1}{n_2+n_1+1} \sum_{n=i-n_1}^{i+n_2} y(k,n) e^{(-d_q+j\nu_q)(i-n)P}. \tag{5.8}$$

The values for n_1 and n_2 should be chosen such that the deterministic model is valid. This could be done by using fixed values such that the deterministic model is valid over the interval $n = i-n_1, ..., i-n_2$ or adaptively, e.g. by using an adaptive segmentation as presented in Chapter 3. Note that for $d_q = 0$ we have a special case of the above presented model, namely, with constant amplitude. In that case the clean speech signal under the deterministic model is assumed to consist of a sum of sinusoids. Hence,

$$\hat{x}(k,i) = E\left[Y(k,i)\right] \tag{5.9}$$

$$\approx \frac{1}{n_1+n_2+1} \sum_{n=i-n_1}^{i+n_2} y(k,n) e^{j\nu_q(i-n)P}. \tag{5.10}$$

We see that Eq. (5.8) and (5.10) modify magnitude as well as phase of the noisy DFT coefficient $y(k,i)$. Further, notice that when $n_1 = n_2 = 0$, the estimate of $x(k,i)$ becomes $\hat{x}(k,i) = y(k,i)$.

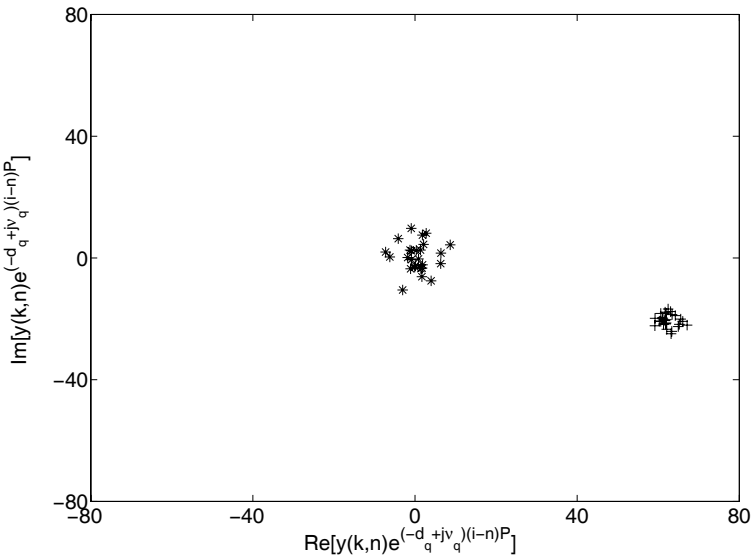

Figure 5.1: $y(k,n)e^{(-d_q+j\nu_q)(i-n)P}$ *at a frequency* k *containing a deterministic signal component* $(+)$ *and at a frequency* k *containing a stochastic signal component* $(*)$*, respectively.*

5.3.1 Simulation Examples

To illustrate the idea of using a deterministic speech model we conducted two simulation experiments. As a first experiment we generate a synthetic clean speech signal consisting of the sum of five (deterministic) sinusoidal components and a (stochastic) autoregressive process. Then we generate a noisy signal by adding white Gaussian noise at an SNR of 10 dB to the clean synthetic signal. We now compute DFT coefficients seen across time and plot in Fig. 5.1 the values of $y(k,n)e^{(-d_q+j\nu_q)(i-n)P}$ for $n = i - n_1, ..., i + n_2$ originating from a frequency bin containing only the stochastic noisy components (the cloud centered around the origin) and the values of $y(k,n)e^{(-d_q+j\nu_q)(i-n)P}$ originating from a frequency bin containing one of the deterministic components (the cloud with an offset from the origin). Notice that in both cases the plotted values of $y(k,n)$ are corrected for the exponential decay and phase shift. As expected, the variance of the latter is smaller than the variance of the first cloud and is only due to the noise variance. Notice, that for the cloud containing noisy deterministic components, it is sufficient to compute the mean of the cloud to estimate the clean deterministic signal component.

In Fig. 5.2 we present a second simulation example where the potential of distinguishing between a stochastic and a deterministic model on a natural speech signal is

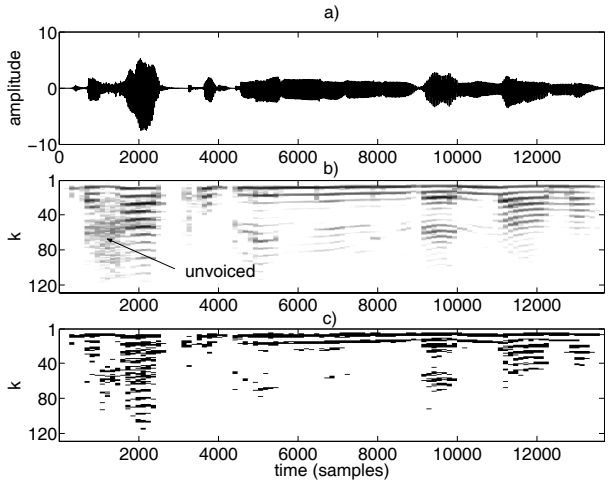

Figure 5.2: *a) Clean speech signal. b) Clean speech spectrogram. c) Black: deterministic model is optimal in terms of local SNR, White: stochastic model is optimal in terms of local SNR.*

demonstrated. In Fig. 5.2a and Fig. 5.2b an original clean speech time-domain signal and its spectrogram are shown, respectively. The signal was degraded by white noise at an SNR of 10 dB and enhanced using 2 different enhancement systems, one using the stochastic model and one using the deterministic model. We compute for each time-frequency point and for each method the resulting SNR and evaluate which of the two models lead to the highest SNR. This is shown in Fig. 5.2c; a preference for the deterministic model is expressed as a black dot and a preference for the stochastic model as a white dot. As expected, the deterministic model performs better at the spectral lines that are visible in the spectrogram (voiced regions), while in the unvoiced speech regions, the stochastic model is preferred. For this experiment we averaged the results over 100 different noise realizations and used the maximum likelihood (ML) approach [3] to estimate the *a priori* SNR ξ, where the number of frames that is used in the ML approach is set to 1. We use the ML approach instead of the often used decision-directed (DD) [3] approach to overcome a dependency on past frames, as will be the case with the DD approach. Such a dependency can lead to wrong, biased, estimates of the suppression gain under the stochastic speech model when speech sound changes take place and as a result can lead to too much suppression or even complete removal of low energy speech components that only last for a short time.

5.4 MMSE Estimator under Stochastic-Deterministic Speech Model

To find an MMSE estimator of the clean speech DFT coefficients under a combined SD speech model, we present in this section two different setups: a completely general model where a soft decision is made between the stochastic and deterministic model based on estimated probabilities and where speech presence uncertainty is taken into account. We will refer to this scheme as SOFT-SD-U. Secondly, a special case of the first model is taken where instead of a soft decision a hard decision between the stochastic and deterministic model is made without speech presence uncertainty, abbreviated as HARD-SD. Although all derivations in this section are per frequency bin k and frame index i, we leave out these indices for notational convenience. This means that $P_{M|Y}(dm|y(k,i))$ is written as $P_{M|Y}(dm|y)$.

5.4.1 SOFT-SD-U Estimator

To find the MMSE estimator SOFT-SD-U, we compute the conditional expectation $E[X|y]$. That is,

$$\begin{aligned}
\hat{x} &= E[X|y] \\
&= \int_x x f_{X|Y}(x|y)dx \\
&= \int_x x \sum_m f_{X|Y,M}\left(x|y,m\right) f_{M|Y}\left(m|y\right) dx \\
&= \int_x x \left\{ f_{X|Y,M}\left(x|y,dm\right) P_{M|Y}\left(dm|y\right) \right. \\
&\quad \left. + f_{X|Y,M}\left(x|y,sm\right) P_{M|Y}\left(sm|y\right) \right\} dx \qquad (5.11) \\
&= E[X|y,dm]P_{M|Y}(dm|y) + E[X|y,sm]P_{M|Y}(sm|y), \qquad (5.12)
\end{aligned}$$

where in (5.11) we used the fact that $x = 0$ when $m = am$. The conditional probabilities $P_{M|Y}(dm|y)$ and $P_{M|Y}(sm|y)$ can be computed using Bayes rule as

$$P_{M|Y}(dm|y) = \frac{\Lambda_{dm}}{\Lambda_{dm} + \Lambda_{sm} + 1}, \qquad (5.13)$$

$$P_{M|Y}(sm|y) = \frac{\Lambda_{sm}}{\Lambda_{dm} + \Lambda_{sm} + 1}, \qquad (5.14)$$

with

$$\Lambda_{dm} = \frac{f_{Y|M}(y|dm)P_M(dm)}{f_{Y|M}(y|am)P_M(am)}$$

and

$$\Lambda_{sm} = \frac{f_{Y|M}(y|sm)P_M(sm)}{f_{Y|M}(y|am)P_M(am)},$$

respectively. Here $P_M(am)$, $P_M(dm)$ and $P_M(sm)$ denote the prior probabilities speech is absent, deterministic speech is present and stochastic speech is present, respectively. The values chosen for these *a priori* probabilities will be discussed in Section 5.5. Further, $f_{Y|M}(y|am)$ is given by

$$f_{Y|M}(y|am) = \frac{1}{\pi\sigma_D^2} \exp\left\{-\frac{|y|^2}{\sigma_D^2}\right\}$$

and $f_{Y|M}(y|sm)$ and $f_{Y|M}(y|dm)$ are given by (5.1) and (5.2), respectively. Computation of (5.2) can be done by substitution of (5.10) in (5.2). Notice that Λ_{sm} can efficiently be written in terms of the *a priori* and *a posteriori* SNR $\xi(k,i)$ and $\zeta(k,i)$, respectively, as presented in [3].

5.4.2 HARD-SD Estimator

For the HARD-SD estimator we assume that speech is always present, i.e. $P_M(am) = 0$. The estimator HARD-SD follows from Eq. (5.12) by setting $P_{M|Y}(dm|y)$ either equal to 1 (deterministic speech model), or to 0 (stochastic speech model). This means that

$$\hat{x} = \begin{cases} E[X|y,dm] & \text{if speech DFT is classified as deterministic,} \\ E[X|y,sm] & \text{if speech DFT is classified as stochastic.} \end{cases} \tag{5.15}$$

The decision between the deterministic and stochastic speech model is made by the following hypothesis test,

$$\begin{aligned} H_0: &\quad E\left[Y(k,i)\right] = 0 \\ H_1: &\quad E\left[Y(k,i)\right] = x(k,i) \text{ and } \mathit{VAR}\left[Y(k,i)\right] = \sigma_D^2(k,i). \end{aligned}$$

Under the H_0 hypothesis the stochastic model is chosen ($P_{M|Y}(dm|y) = 0$) and under the H_1 hypothesis the deterministic model is chosen ($P_{M|Y}(dm|y) = 1$). We decide between H_0 and H_1 using the Bayes criterion [7], that is

$$T = \frac{f_{Y|M}(y|dm)}{f_{Y|M}(y|sm)} \underset{H_0}{\overset{H_1}{\gtrless}} \lambda_{th}, \tag{5.16}$$

where the threshold $\lambda_{th} = \frac{1-P_M(dm)}{P_M(dm)}$.

In Fig. 5.3 the hypothesis test to distinguish between a stochastic (Gaussian) and deterministic speech model in Eq. (5.16) is demonstrated using the same speech signal as used in Fig. 5.2, degraded by white noise at an SNR of 10 dB. For this experiment we used for $P_M\,(dm)$ the value that is specified in Section 5.5 and given in Table 5.1. The top figure shows the clean speech spectrogram. The bottom figure shows in the time-frequency plane the outcome of the hard decision of (5.16), where a black dot means that the speech component is classified as deterministic and a white dot that it is classified as stochastic. The hypothesis test appears to perform as expected: DFT coefficients representing harmonics are classified as deterministic, while e.g. the DFT coefficients in the region indicated with an arrow are classified as stochastic.

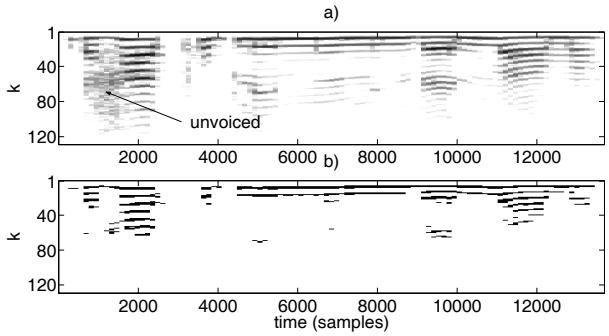

Figure 5.3: *a) clean speech signal spectrogram. b) Outcome of hypothesis test, Black: speech component is classified as deterministic, White: speech component is classified as stochastic.*

5.5 Experimental Results and Discussion

In this section we compare the proposed SOFT-SD-U and HARD-SD enhancement methods with a traditional enhancement method, similar to the one depicted in Fig. 1.2, where the speech estimator relies on a stochastic speech model alone with and without speech presence uncertainty, respectively. For evaluation we use the perceptual evaluation of speech quality (PESQ) measure [8] and segmental SNR defined as [9]

$$\text{SNR}_{\text{seg}} = \frac{1}{N}\sum_{i=0}^{N-1}\mathcal{T}\left\{10\log_{10}\frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i)-\hat{\mathbf{x}}_t(i)\|^2}\right\},$$

where $\mathbf{x}_t(i)$ and $\hat{\mathbf{x}}_t(i)$ denote frame i of the clean speech signal and the enhanced speech signal, respectively, N is the number of frames within the speech signal in question and $\mathcal{T}(x) = \min\{\max(x,-10),35\}$, which confines the SNR at each frame to a perceptually meaningful range between -10 dB and 35 dB. All results presented below are averaged over 24 different speech signals that originate from the TIMIT database [10].

In all experiments we use speech fragments sampled at 8 kHz and frame sizes of 256 samples taken with 50% overlap. To have good time resolution in the estimation of (5.4), the DFT samples $y(k,n)$, $n = i-n_1,...,i+n_2$, are computed from frames with an overlap of 84%. This overlap was chosen based on a trade off, where on one hand a small overlap is desirable to better satisfy the assumption made in Section 5.3, i.e. frame shift P is sufficiently large with respect to the time-span of the dependency of the noise. On the other hand, a large overlap is necessary when using multiple samples in (5.8), i.e. $n_1, n_2 > 0$, because approximation of (5.4) by (5.8) is only valid over relatively short time intervals. In all experiments, noise statistics are measured during silence regions preceding speech activity.

Initial experiments have shown that in terms of SNR_{seg}, the difference between the use of Eq. (5.8) and (5.10) for estimating $x(k,i)$ is negligible. Therefore, we use (5.10) in all our experiments. Furthermore, $n_1 = n_2 = 2$ is chosen based on initial

SOFT-SD-U		HARD-SD	
Probability	value	Probability	value
$P_M(dm)$	0.041	$P_M(dm)$	0.041
$P_M(sm)$	0.22	$P_M(sm)$	0.959
$P_M(am)$	0.739	$P_M(am)$	0

Table 5.1: *Probabilities used in experiments.*

experiments.

With respect to the SOFT-SD-U estimator, Eqs. (5.13) and (5.14) require knowledge of the prior probabilities $P_M(am)$, $P_M(dm)$, and $P_M(sm)$. To compute these probabilities we assume that English speech on average can be classified as voiced in 78% of the time [11], that the fundamental frequency of speech is between $f_0 = 50$ and $f_0 = 500$ Hz [9] and that for most voiced speech sounds, speech energy is dominantly present up to approximately $f_c = 2000$ Hz. We then can compute the prior probabilities as

$$\begin{aligned} P_M(dm) &= 0.78 * \frac{f_c}{f_0}\frac{2}{K} \\ P_M(sm) &= 0.22 \\ P_M(am) &= 1 - P_M(dm) - P_M(sm), \end{aligned}$$

where K is the window size. For a sample frequency $f_s = 8000$ Hz, a window size $K = 256$ samples and a typical fundamental frequency of $f_0 = 300$ Hz this leads to the values as listed in Table 5.1.

For the HARD-SD estimator it is assumed that speech is always present, that is $P_M(am) = 0$. To compute the threshold λ_{th} in Eq. (5.16), the prior probabilities $P_M(dm)$ and $P_M(sm)$ are required. These are determined as

$$\begin{aligned} P_M(dm) &= 0.78 * \frac{f_c}{f_0}\frac{2}{K} \\ P_M(sm) &= 1 - P_M(dm) \end{aligned}$$

The values that are used in the experiment are listed in Table 5.1.

Estimation of ν_p in (5.10) is done using the ESPRIT algorithm as mentioned in Section 5.3. Under very low SNRs, estimation of ν_p can lead to inaccurate estimates and, consequently, inaccurate values for (5.13) and (5.14). This in turn leads to a perceptually annoying switching between the deterministic and stochastic model. To overcome this we do not use the deterministic model when $\hat{\xi}(k,i) < -7$ dB and use a stochastic model alone instead. To estimate the *a priori* SNR $\xi(k,i)$ under the stochastic speech model, the decision-directed approach [3] is used with a smoothing factor $\alpha = 0.98$ with $\hat{\xi}(k,i) = \min(\hat{\xi}(k,i), -15\ dB)$.

To demonstrate that the proposed method is general and can also work with other distributions under the stochastic speech model as well, we present experimental results for both the Gaussian and Laplace distribution. The reference methods used in the experiments are named: Stoch-Gauss, which is when speech DFT coefficients

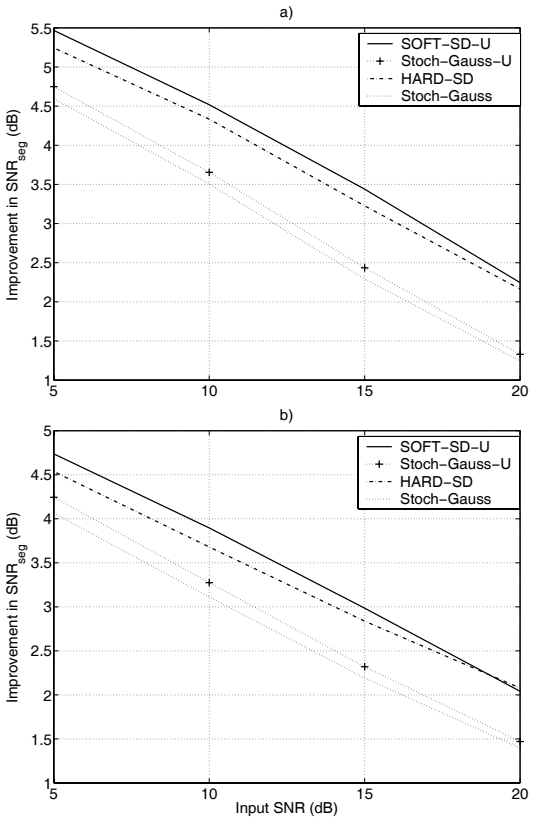

Figure 5.4: *Performance comparison for Gaussian stochastic model versus combined Gaussian stochastic/deterministic model in terms of input SNR versus improvement in SNR_{seg} for speech signals degraded by a) white noise b) F16-fighter cockpit noise.*

are always assumed to be Gaussian distributed and speech is always assumed to be present. When speech presence uncertainty is taken into account this is referred to as Stoch-Gauss-U. Similarly, when speech DFT coefficients are always assumed to be Laplace distributed and speech is always assumed to be present, this is referred to as Stoch-Lap. When speech presence uncertainty is taken into account this is referred to as Stoch-Lap-U.

5.5.1 Experimental Results under Gaussian Stochastic Model

In this section we present objective results for the proposed algorithms, where we model the clean speech DFT coefficients under the stochastic speech model with a Gaussian distribution.

In Fig. 5.4a we compare the performance of the proposed algorithms with the reference methods in terms of improvement in SNR_{seg} when speech signals are degraded

by white noise at an SNR in the range from 5 dB to 20 dB. Over the whole range of input SNRs the proposed methods improve the performance compared to the use of a stochastic model alone. In terms of SNR_{seg}, the performance improvement of HARD-SD over Stoch-Gauss is approximately 0.82 dB. Incorporating the soft decision model between speech absence, the deterministic speech model and the stochastic speech model, i.e. SOFT-SD-U over HARD-SD, leads to an additional 0.2 dB improvement. The improvement of SOFT-SD-U over Stoch-Gauss-U is approximately 0.87 dB.

In Fig. 5.4b objective results are shown for signals degraded by F16-fighter cockpit noise, where similar performance is shown as for the white noise case.

5.5.2 Experimental Results under Laplace Stochastic Model

In this section we present objective results for the proposed algorithms, for the case that clean speech DFT coefficients are modelled as Laplace distributed random variables under the stochastic model.

In Fig. 5.5a we compare the performance of the proposed algorithms with the reference methods in terms of improvement in SNR_{seg} for speech signals degraded by white noise in the range from 5 to 20 dB. Similarly as for the Gaussian stochastic model case also here SNR_{seg} is improved for the SD based approaches with respect to the use of Stoch-Lap and Stoch-Lap-U over the whole input range of SNRs. In general the performance differences are smaller than when a Gaussian distribution is assumed as in Section 5.5.1. We will comment on this in Section 5.5.5.

In terms of SNR_{seg}, the performance improvement of HARD-SD over the use of a stochastic Laplacian model alone is approximately 0.11 dB. Incorporating the soft decision model between speech absence, the deterministic speech model and the stochastic speech model, i.e. SOFT-SD-U over HARD-SD leads to an additional 0.21 dB improvement. The improvement of SOFT-SD-U over Stoch-Lap-U is approximately 0.22 dB.

In Fig. 5.5b similar objective results are shown, but now for signals degraded by F16-fighter cockpit noise. The comparison between SOFT-SD-U and Stoch-Lap-U shows improvements for SOFT-SD-U for input SNRs of 10 dB and larger. The performance difference between HARD-SD and Stoch-Lap is negligible.

5.5.3 PESQ Evaluation

For a further evaluation of the proposed algorithms, we use the perceptual evaluation of speech quality (PESQ) measure [8], which predicts the subjective quality of speech signals with high correlation between subjective and objective results and expresses the quality in a score from 1.0 (worst) up to 4.5 (best). In Fig. 5.6a and 5.6b we compare PESQ scores for speech signals degraded by white noise and F16-fighter cockpit noise, respectively, when it is assumed that speech is Gaussian distributed under the stochastic speech model. Both SOFT-SD-U and HARD-SD lead to improved PESQ scores with respect to Stoch-Gauss-U and Stoch-Gauss. For signals degraded by white noise SOFT-SD-U and HARD-SD lead to an improvement of approximately 0.2 and 0.1 over Stoch-Gauss-U and Stoch-Gauss, respectively. For signals degraded

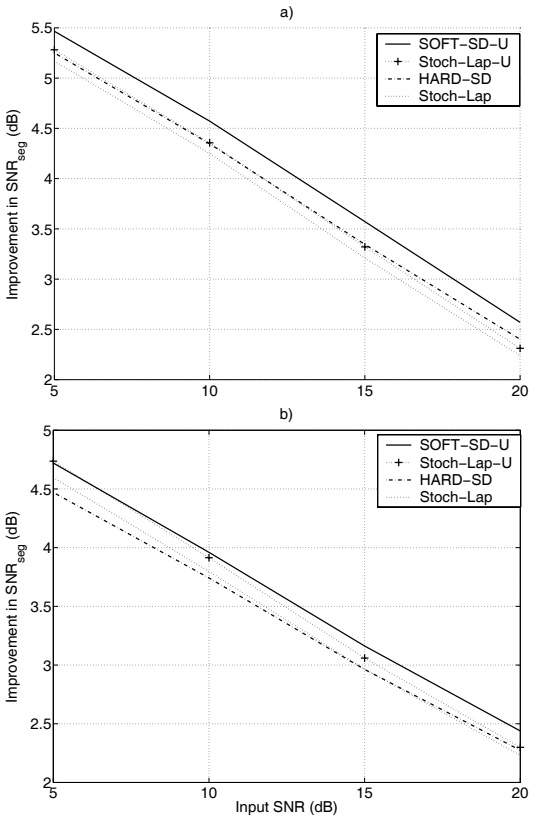

Figure 5.5: *Performance comparison for Laplacian stochastic model versus combined Laplacian stochastic/deterministic model in terms of input SNR versus improvement in SNR_{seg} for speech signals degraded by a) white noise b) F16-fighter cockpit noise.*

by F16-fighter cockpit noise, the improvement of Soft-SD-U and HARD-SD over Stoch-Gauss-U and Stoch-Gaus is 0.16 and 0.11, respectively.

In Fig. 5.7a and 5.7b we compare PESQ scores when it is assumed that speech DFT coefficients are Laplacian distributed under the stochastic speech model. For both white noise and F16-fighter cockpit noise the PESQ difference between HARD-SD and Stoch-Lap is more or less negligible. The PESQ improvement of SOFT-SD-U over Stoch-Lap-U is 0.08 and 0.05 for signals degraded by white noise and F16-fighter cockpit noise, respectively.

Notice that Figs. 5.6 and 5.7 show smaller differences in terms of PESQ score between the several enhancement methods at lower input SNR (e.g. at 5 dB) than at higher input SNR, while in Section 5.5.1 and 5.5.2 it is shown that over the whole range of input SNRs the improvement in terms of SNR_{seg} is approximately equal. Although PESQ and SNR_{seg} are both quality measures, we cannot expect them to measure the same kind of improvement, since they measure different aspects of quality.

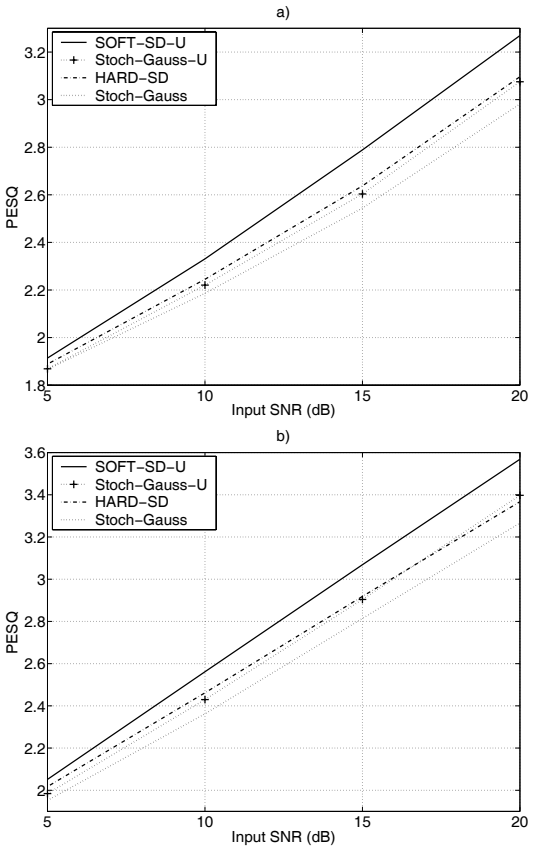

Figure 5.6: *Performance comparison in terms of PESQ under a Gaussian stochastic model for a) input signals degraded by white noise b) input signals degraded by F16-fighter cockpit noise.*

5.5.4 Subjective Evaluation

Informal listening was performed on the presented methods. In these listening experiments we compared the proposed methods to the use of a stochastic model alone. As densities under the stochastic model both the Gaussian and the Laplace density were used.

When using a Gaussian density under the stochastic model, the difference between SOFT-SD-U and HARD-SD is mainly reflected in a lower broadband noise floor and less reverberant speech for SOFT-SD-U than for HARD-SD. Comparing SOFT-SD-U using a Gaussian density under the stochastic model with Stoch-Gauss-U, it turns out that SOFT-SD-U leads to less suppressed and better understandable speech. Further,

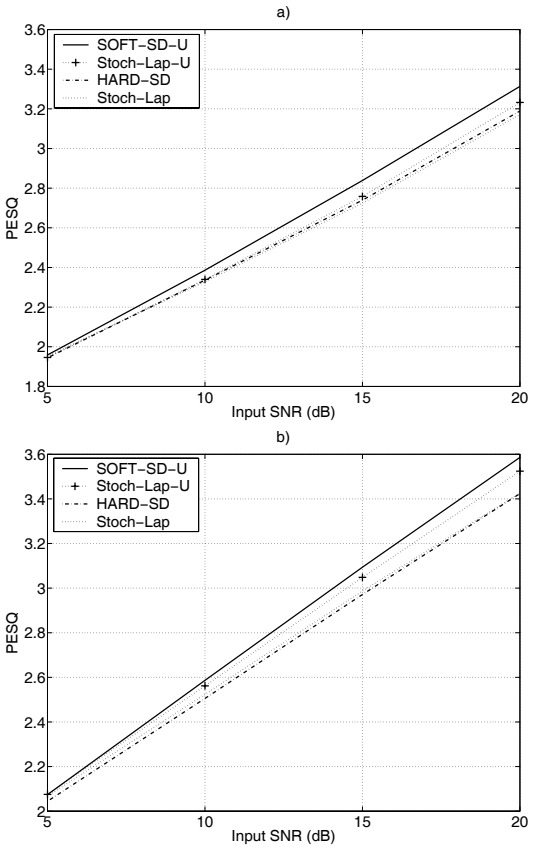

Figure 5.7: *Performance comparison in terms of PESQ under a Laplace stochastic model for a) input signals degraded by white noise b) input signals degraded by F16-fighter cockpit noise.*

the residual noise is lower in level, but slightly more musical when using SOFT-SD-U than when using Stoch-Gauss-U. From the comparison between HARD-SD and Stoch-Gauss it follows that the speech sounds less suppressed when using HARD-SD and also less reverberant than when using Stoch-Gauss.

Using a Laplace density under the stochastic model instead of a Gaussian model leads in general to smaller differences between SOFT-SD-U and HARD-SD, but also between SOFT-SD-U and Stoch-Lap-U, and between HARD-SD and Stoch-Lap.

Using SOFT-SD-U with a Laplace density under the stochastic model leads to a somewhat better speech quality than Stoch-Lap-U. However, the performance difference is smaller than under the Gaussian stochastic model. Both Soft-SD-U and Stoch-Lap-U lead to some musical tones, but SOFT-SD-U introduces slightly more musical tones than Stoch-Lap-U.

Also between Hard-SD with a Laplace density under the stochastic model and

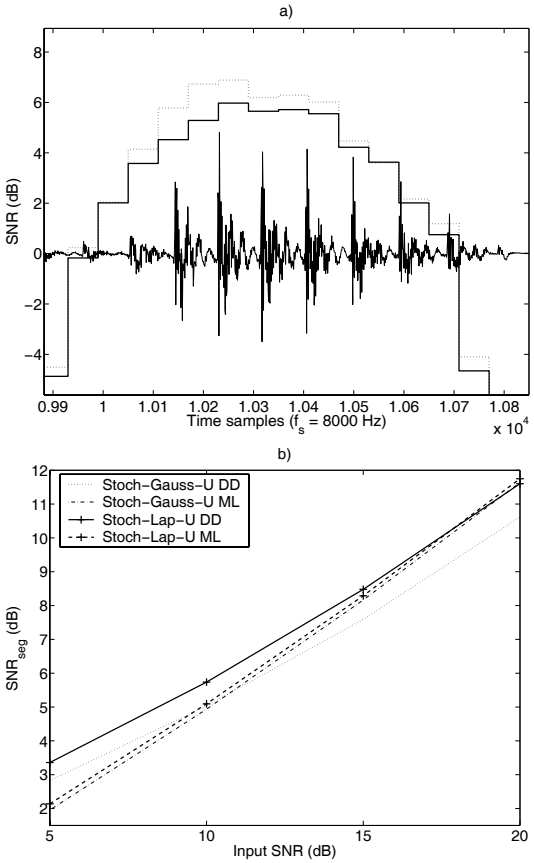

Figure 5.8: *Performance comparison in terms of SNR over time between a) Stoch-Lap-U (dotted) and Stoch-Gauss-U (solid) b) Stoch-Lap-U versus Stoch-Gauss-U when ξ is estimated with both the decision-directed and the maximum likelihood approach.*

Stoch-Lap the difference is reflected in terms of a better speech quality for Hard-SD, but less musical artifacts for Stoch-Lap.

5.5.5 Gaussian versus Laplace Stochastic Model

In this section we study the difference in performance between the Gaussian and Laplace stochastic speech model as demonstrated in the experimental results in the previous sections. We explain why there is a smaller performance difference between SOFT-SD-U and Stoch-Lap-U than between SOFT-SD-U and Stoch-Gauss-U. To do so, we compare in Table 5.2 the average SNR_{seg} after enhancement of speech signals that were originally degraded by white noise at an SNR of 10 dB. We see from Table 5.2 that the use of a Laplace distribution (Stoch-Lap-U) instead of a Gaussian distribution (Stoch-Gauss-U) for the speech DFT coefficients leads to improved SNR_{seg}. This

Speech model	Gaussian model SNR_{seg} (dB)	Laplacian model SNR_{seg} (dB)
Stoch-Gaus/Lap-U	5.0	5.7
SOFT-SD-U	5.9	6.0

Table 5.2: *Comparison between the use of a Gaussian and Laplace distribution.*

is in accordance with the results in [12] where an improvement of approximately 0.5 dB was reported. Moreover, we see from Table 5.2 that also the proposed SOFT-SD-U method with the Laplace distribution as a stochastic model performs slightly better in terms of SNR_{seg} as compared to SOFT-SD-U when using a Gaussian model. However, comparing the results for SOFT-SD-U in Table 5.2 we see that the difference between SOFT-SD-U under the two different stochastic models is decreased to approximately 0.1 dB.

Investigation of the Laplace gain function as presented in [13] and experimental analysis given in this section reveal that the 0.7 dB performance improvement of Stoch-Lap over Stoch-Gauss is only partly due to a better speech model, but that there are other beneficial side-effects of using the Laplace distribution that lead to performance improvement. More specifically, it can be observed that the better performance is partly related to the use of the decision-directed approach for estimating the *a priori* SNR. From [13] we know that the gain function under the Laplace distribution applies less suppression than the Wiener gain when the *a posteriori* SNR $\zeta(k,i)$ is high and the *a priori* SNR $\xi(k,i)$ low, a situation that typically arises for speech onsets. The Wiener gain, using the Gaussian distribution, on the other hand does not have this mechanism and will always apply high suppression when $\xi(k,i)$ is low independent of the *a posteriori* SNR. Because the decision-directed approach leads to an underestimated *a priori* SNR at speech onsets [14] due to a dependency on previous frames, the Wiener filter will apply too much suppression on the onsets. The Laplace based gain function, on the other hand, applies less suppression, due to the above described mechanism, and will thus lead to less distorted speech. This effect is visualized in Fig. 5.8a, where the SNR per frame after enhancement of a speech signal degraded by white noise at an SNR of 5 dB is shown, together with the original clean speech signal. It is clearly visible that especially at the first half of the speech sound the use of the Laplace distribution leads to improved SNR. This is where the DD approach leads to an underestimation of the *a priori* SNR. In the second half of the speech sound there is still some improvement, although much smaller, because the influence on the *a priori* SNR estimation of the noise only frames preceding the current speech sound decreases as time evolves.

To support our discussion of the above described mechanism we compare enhancement using Stoch-Lap with Stoch-Gauss in terms of SNR_{seg} averaged over 24 different speech signals degraded by white noise at an SNR of 10 dB. In this comparison we use two different *a priori* SNR estimators, namely the DD approach and the maximum likelihood approach [3][15]. With the maximum likelihood approach the *a priori* SNR is computed based on noisy periodograms that are averaged over the current and the two last frames. The latter approach leads in general to more musical noise than the

DD approach, however, it has a smaller dependency on previous frames. Fig. 5.8b shows that the Laplace distribution still leads to somewhat better performance, but that by elimination of the dependency and consequently the above described mechanism the performance gain of the Laplace distribution over the Gaussian distribution is decreased from 0.7 dB to 0.15 dB. Moreover, this mechanism also explains why the improvements of the SD methods are relatively smaller when the Laplace distribution is used. Specifically, one advantage of the deterministic model is the independence of the *a priori* SNR estimation and therefore it is independent of the use of a decision-directed approach. This overcomes, similarly as with the Laplace gain function, an oversuppression at the start of stationary speech sounds. It explains why combining the Laplace model with the deterministic model leads to a relatively smaller improvement than combining the Gaussian distribution as a stochastic model with the deterministic speech model.

5.6 Conclusions

In this chapter we proposed the use of a combined stochastic-deterministic speech model for DFT-domain based speech enhancement. Under the deterministic speech model, clean speech DFT coefficients are modelled as a complex exponential across time. Using the combined speech model we derived an MMSE estimator for clean speech where speech presence uncertainty can be taken into account. We demonstrated the use of the combined stochastic-deterministic speech model using the Gaussian and Laplace distributions, however, the presented method is general and can be extended to be used with other distributions under the stochastic representation. Experiments showed that the use of the proposed MMSE estimator leads to improvements in terms of segmental SNR over the use of a stochastic speech model alone. Moreover, evaluation with PESQ demonstrated an improvement in speech quality. However, performance differences tend to get smaller when using the Laplace density as a stochastic model instead of a Gaussian density. A discussion was presented to explain this performance difference.

References

[1] W. B. Kleijn and K. K. Paliwal. *Speech coding and synthesis*. Elsevier, 1995.

[2] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(2):137–145, April 1980.

[3] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[4] D. R. Brillinger. *Time Series: Data Analysis and Theory*. SIAM, Philadelphia, 2001.

[5] B. D. Rao and K. S. Arun. Model based processing of signals: a state space approach. *Proc. Of the IEEE*, 80(2):283–307, Feb. 1992.

[6] R. Roy, A. Paulraj, and T. Kailath. Esprit–a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoust., Speech, Signal Processing*, 34(5):1340 – 1342, Oct. 1986.

[7] S. K. Kay. *Fundamentals of Statistical signal processing*, volume 2. Prentice Hall, Upper Saddle River, NJ, 1998.

[8] J. G. Beerends. Extending p.862 PESQ for assessing speech intelligibility. *White contribution COM 12-C2 to ITU-T Study Group 12*, October 2004.

[9] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

[10] DARPA. Timit, Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc 1-1.1, October 1990.

[11] G. Dewey. *Relative Frequency of English Speech Sounds*. Harvard Univ. Press, 1923.

[12] R. Martin and C. Breithaupt. Speech enhancement in the DFT domain using Laplacian speech priors. In *Int. Workshop on Acoustic, Echo and Noise Control*, pages 87–90, September 2003.

[13] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[14] R. C. Hendriks, R. Heusdens, and J. Jensen. Forward-backward decision directed approach for speech enhancement. In *Int. Workshop Acoustic Echo and Noise Control (IWAENC)*, pages 109–112, September 2005.

[15] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-27(2):113–120, April 1979.

Chapter 6

Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors

---

This chapter is based on work performed jointly by J. S. Erkelens, R. C. Hendriks, R. Heusdens and J. Jensen and has been published as "Minimum Mean-Square Error Estimation of Discrete Fourier Coefficients with Generalized Gamma Priors", in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 15, no. 6, pages 1741 - 1752, August. 2007.

6.1 Introduction

As mentioned in Section 2.3, measured histograms of speech DFT coefficients show under certain conditions a super-Gaussian shape. In this chapter we focus on MMSE estimators of the clean speech DFT coefficient magnitudes, as well as of the complex-valued DFT coefficients, under a density that can model this observed super-Gaussian behavior. We assume that the noise DFT coefficients obey a (complex) Gaussian distribution as argued in Section 2.3 and investigate the use of the generalized Gamma density to model the speech DFT magnitudes and DFT coefficients.

6.1.1 Modelling Speech DFT Magnitudes

For estimation of the speech DFT magnitudes, we assume that the speech DFT magnitudes are distributed according to a one-sided prior of the general form

$$f_A(a) = \frac{\gamma\beta^\nu}{\Gamma(\nu)} a^{\gamma\nu-1} \exp(-\beta a^\gamma), \qquad \beta > 0,\ \gamma > 0,\ \nu > 0,\ a \geq 0, \tag{6.1}$$

where $\Gamma(\cdot)$ is the gamma function and the random variable A represents the DFT magnitude. This density is known as the generalized Gamma density and is able to model heavy-tailed densities, depending on the parameter settings in Eq. (6.1). Fig. 6.1 shows example densities for $\gamma = 1$ and $\gamma = 2$, respectively. For $\gamma = 2$ and $\nu = 1$, the Rayleigh distribution occurs as a special case, for which an MMSE amplitude estimator and a MAP amplitude estimator have been derived in [1] and [2], respectively. For $\gamma = 2$, a generalized MMSE amplitude estimator was derived in [3] as a function of ν. Furthermore, [4] presented a generalized MAP estimator and an adaptive algorithm for estimating the parameters of the generalized prior. For $\gamma = 1$, no generalized MMSE amplitude estimator is known in closed form, but a numerical approximation was presented in [3]. An approximate generalized MAP estimator for the case $\gamma = 1$ was derived in [5]. The first column of Table 6.1 summarizes the special cases of Eq. (6.1) for which estimators (MAP or MMSE) have been documented in the literature.

The MMSE DFT magnitude estimators that we present in this chapter are derived assuming that these random variables have a single-sided generalized gamma prior as in Eq. (6.1). In contrast to the estimators presented in [3], we present analytical expressions for both $\gamma = 1$ and $\gamma = 2$ and do not make use of numerical solutions.

6.1.2 Modelling Speech DFT Coefficients

For estimation of the complex-valued speech DFT coefficients we assume that the real and imaginary parts of these coefficients are statistically independent. The validity of this assumption will be discussed in Secion 6.2. We will derive MMSE estimators for a two-sided generalized gamma prior density of the following form

$$f_{X_\Re}(x_\Re) = \frac{\gamma\beta^\nu}{2\Gamma(\nu)} |x_\Re|^{\gamma\nu-1} \exp(-\beta|x_\Re|^\gamma),\ \beta > 0,\ \gamma > 0,\ \nu > 0, -\infty < x_\Re < \infty \tag{6.2}$$

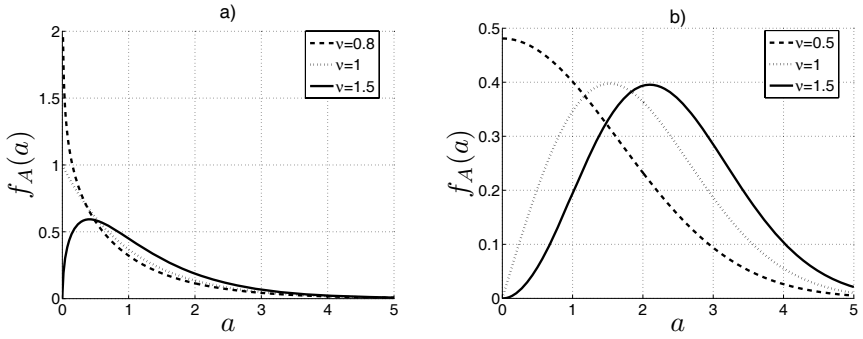

Figure 6.1: *Prior densities $f_A(a)$ for a) $\gamma = 1$ with $\nu = \{0.8, 1, 1.5\}$, and for b) $\gamma = 2$ with $\nu = \{0.5, 1, 1.5\}$. The densities have been normalized to unit variance.*

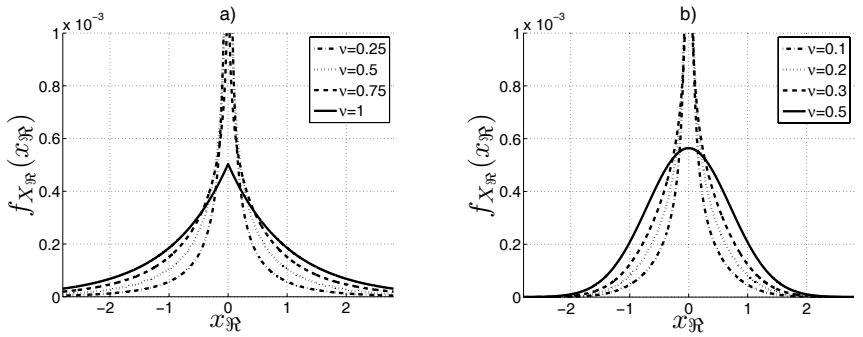

Figure 6.2: *Prior densities $f_{X_\Re}(x_\Re)$ for a) $\gamma = 1$ with $\nu = \{0.25, 0.5, 0.75, 1\}$, and for b) $\gamma = 2$ with $\nu = \{0.1, 0.2, 0.3, 0.5\}$. The densities have been normalized to unit variance.*

where the random variable $X_\Re$ represents the real part of a complex-valued DFT coefficient. A similar equation holds for the imaginary part. We consider the cases of $\gamma = 1$ and $\gamma = 2$. Examples of the resulting prior densities are shown in Fig. 6.2. These densities (parameterized by β and ν) contain a number of special cases for which estimators are already known. Specifically, for $\gamma = 2$, the prior parameterizes the Gaussian density ($\nu = 1/2$) for which the Wiener estimator is the MMSE estimator [6]. For $\gamma = 1$, Eq. (6.2) has the Gamma and the Laplacian density as special cases ($\nu = 1/2$ and $\nu = 1$, respectively). MMSE estimators under these densities are derived in [7]. Choosing $\nu > 1.0$ with $\gamma = 1$ or $\nu > 0.5$ with $\gamma = 2$ leads to bimodal priors. Although the estimators derived below remain valid for bimodal priors, we have chosen to restrict ν in our evaluations to the range $0 < \nu \leq 1.0$ for $\gamma = 1$ and $0 < \nu \leq 0.5$ for $\gamma = 2$ to have unimodal priors and thus be better in line with observed speech data, see e.g. [7]. Table 6.1 summarizes the special cases of Eq (6.2) for which estimators (MMSE) have been documented in the literature.

The MMSE estimators of the complex clean speech DFT coefficients that we

	DFT magnitudes ($f_A(a)$)	Complex DFTs ($f_{X_\Re}(x_\Re)$)
$\gamma = 1$	Generalized gamma (MMSE) [3]* Generalized gamma (MAP) [5]*, [3]*	Laplacian (MMSE) [7] Gamma (MMSE) [7]
$\gamma = 2$	Rayleigh (MMSE) [1] Rayleigh (MAP) [2] Generalized gamma (MMSE / MAP) [3][4]	Gaussian (MMSE) [6]

Table 6.1: *Special cases of the generalized priors in Eqs. (6.1) and (6.2) for which estimators are known. For all estimators the noise is assumed to be additive and the noise DFT coefficients are assumed to be Gaussian distributed. The * indicates estimators for which no exact closed-form solutions exist.*

present in this chapter are derived assuming that the real and imaginary parts of clean speech DFT coefficients have a two-sided generalized gamma distribution, as in Eq. (6.2). As mentioned, specific choices of ν and β lead to special cases for which MMSE estimators already exist. The derived estimators are more general and cover all possible MMSE estimators (including the ones shown) in each of the quadrants of Table 6.1.

The remaining sections in this chapter are organized as follows. In Section 6.2 we discuss the validity of the assumptions made in the introduction and study the consistency between the models of complex DFT coefficients and magnitudes. In Section 6.3, we introduce the signal model and the notation used throughout this chapter. Section 6.4 treats MMSE estimation of DFT coefficient magnitudes, while Section 6.5 considers MMSE estimators of complex DFT coefficients. Filter characteristics corresponding to the derived estimators are shown in Section 6.6. In Section 6.7 we present experimental results. In Section 6.8 concluding remarks are given.

6.2 Discussion of the Modelling Assumptions

As outlined in the previous section, existing DFT coefficient estimators, as well as the ones derived here, rely on a number of assumptions with respect to speech DFT coefficients. In this section we discuss the consistency and validity of these assumptions. Let us first discuss the pdfs in Eqs. (6.1) and (6.2). The pdfs are dependent on the parameter β which is related to the speech spectral variance σ_X^2 (See Appendix A). Since in practice σ_X^2 is unknown, it is estimated from the noisy data, e.g. using the decision-directed approach introduced by Ephraim and Malah in [1]. Consequently, the pdfs in Eqs. (6.1) and (6.2) are actually models for the priors faced in practice that are *conditioned* on the *estimated* speech spectral variance, rather than on the true underlying (but unknown) value of σ_X^2. Martin [7] and Lotter and Vary [5] showed that super-Gaussian models of the real and imaginary parts as well as magnitudes of DFT coefficients, conditioned on speech spectral variances estimated by the decision-directed approach, offer a better fit than a gaussian model. Hence it is important to notice that the appropriate distributional assumption is related to the speech vari-

ance estimator that is used. For example, Cohen [8] suggests that for an *a priori* SNR estimator based on GARCH models, the Gaussian speech model is superior. A slight preference for complex Gaussian distributions has also been found for the DFT-coefficients from short analysis frames of individual speech sound classes (vowels, plosives, fricatives, etc.) using maximum likelihood estimates of the speech spectral variance [9].

The second point concerns the consistency between the models in the complex domain (real and imaginary parts) and the polar domain (amplitude and phase). It is well-known that independent and identically distributed (i.i.d.) Gaussian real and imaginary parts correspond to a Rayleigh distribution for the amplitudes which are independent of the uniformly distributed phase. We investigated whether i.i.d. generalized-Gamma distributed real and imaginary parts lead to generalized-Gamma distributed amplitudes. It is not difficult to show that this is indeed the case for the $\gamma = 2$ class of distributions. The value of the parameter ν in the polar domain is then twice as large as the corresponding value in the complex domain. Furthermore, amplitude and phase are also independent, but the phase is not uniformly distributed (except for the Gaussian case, of course). For the $\gamma = 1$ case, these results do not hold in an exact mathematical sense. However, an accurate fit can be made to the resulting amplitude distribution. As in the $\gamma = 2$ case, the resulting phase distribution is generally non-uniform. When we start with a generalized-gamma model in the polar domain, and assume uniformly distributed phase, the corresponding real and imaginary parts are not independent, except for the Gaussian case. However, simulations showed that a fairly accurate fit of the pdf in Eq. (6.2) can still be made to their marginal distributions.

This bring us to perhaps the most important issue; how well do the assumed distributions match real speech data? Martin [7], and Lotter and Vary [5] have measured the distributions of speech DFT coefficients conditioned on a certain narrow range of estimated *a priori* SNR values. Contour lines of the measured joint pdf of real an imaginary parts of the DFT coefficients are nearly circular. A circularly symmetric joint pdf means that the real and imaginary parts are uncorrelated (but, as we shall see, they are not independent), and that the phase distribution is uniform and independent from the amplitude distribution.

In order to gain further insights, we performed a similar experiment as in [5][7] leading to the contour plots of measured histograms of real and imaginary parts shown in Fig. 6.3. As in [5], only DFT coefficients have been taken into account for which the *a priori* SNR estimated using the decision-directed approach was between 19 and 21 dB. The entire TIMIT-TRAIN database provided the speech material, limited to telephone bandwidth, to which white Gaussian noise at an SNR of 30 dB was added. The noise variance was estimated for each sentence from a noise-only segment of 0.64 seconds, preceding each sentence. Fig. 6.3a shows the contours for the joint distribution; this distribution is very similar to the ones in [7, 5]. Fig. 6.3b shows the contours for the product of the marginal distributions. This plot is different from Fig. 6.3a, and therefore, even though real and imaginary parts may be uncorrelated, there is clearly some higher-order dependency between them.

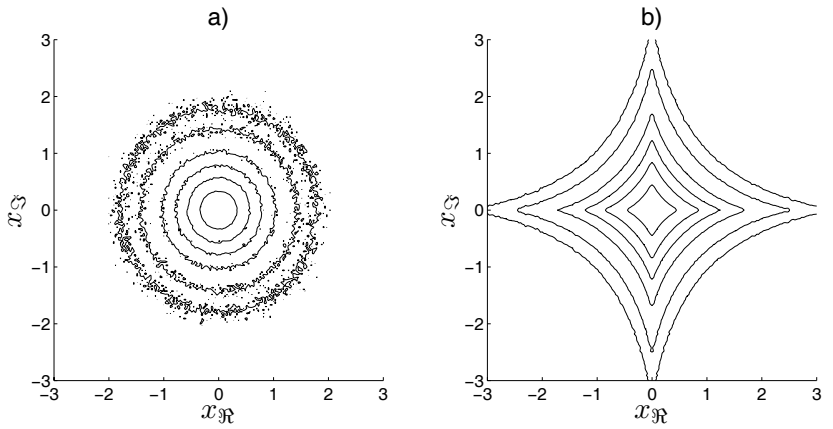

Figure 6.3: *Contour lines of measured distributions of real and imaginary parts normalized to unit variance: a) joint distribution; b) product of marginal distributions.*

In the derivation of the complex DFT estimators, the real and imaginary parts are assumed independent for mathematical tractability, but, as shown in Fig. 6.3b, this is not entirely in line with measured speech data. Still, we cannot predict beforehand whether the magnitude or complex DFT estimators lead to the best speech enhancement performance, because the parametric distributions of Eqs. (6.1) and (6.2) are only *models* of the actual conditional speech distributions. The fits to measured data are not perfect in either domain and the derived estimators in each domain may not be equally sensitive to the modelling errors. In this chapter we will investigate the performance of the generalized estimators, and will show experimentally that the amplitude estimators perform slightly better than the complex DFT estimators under the assumed models.

6.3 Signal Model and Notation

We assume that $X(k,i)$ and $D(k,i)$ are statistically independent across time and frequency, which leads to estimators that are independent of time and frequency as well. This allows us to increase readability by dropping the time/frequency indices, i.e. we write Eq. (2.1) as

$$Y = X + D. \tag{6.3}$$

We assume that the noise DFT coefficients D obey a Gaussian distribution, as argued for in Section 2.3, with independent and identically distributed real and imaginary parts, i.e.

$$\sigma_D^2 = \sigma_{D_\Re}^2 + \sigma_{D_\Im}^2, \text{ and } \sigma_{D_\Re}^2 = \sigma_{D_\Im}^2. \tag{6.4}$$

Next to the *a priori* SNR ξ and the *a posteriori* SNR ζ, as defined in Chapter 2, we also define the *a posteriori* SNR with respect to the real and imaginary part of Y, that is $\zeta_\Re = y_\Re^2/\sigma_{D_\Re}^2$, and $\zeta_\Im = y_\Im^2/\sigma_{D_\Im}^2$. From Eq. (6.4) we have $\zeta = (\zeta_\Re + \zeta_\Im)/2$.

6.4 MMSE Estimation of Magnitudes of DFT Coefficients

In this section we derive MMSE estimators of the magnitude of the clean speech DFT coefficients. The MMSE estimator is identical to the conditional mean [10] given by

$$E\{A|r\} = \frac{\int_0^\infty a f_{R|A}(r|a) f_A(a) da}{\int_0^\infty f_{R|A}(r|a) f_A(a) da}. \tag{6.5}$$

Further, since the noise DFT coefficients are assumed to be Gaussian distributed, $f_{R|A}(r|a)$ can be written as shown in Section 2.3 as

$$f_{R|A}(r|a) = \frac{2r}{\sigma_D^2} \exp\left(-\frac{r^2+a^2}{\sigma_D^2}\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right), \tag{6.6}$$

where $\mathcal{I}_0$ is the 0th order modified Bessel function of the first kind. No assumption about the clean speech phase distribution has to be made to derive this expression.

We will consider the case $\gamma = 2$ first, because the corresponding estimator can be derived without any approximations.

6.4.1 DFT Magnitudes, γ=2

Let the superscript (2) indicate that $\gamma = 2$. Inserting Eqs. (6.1) with $\gamma = 2$ and (6.6) into Eq. (6.5) gives

$$\hat{A}^{(2)} = \frac{\int_0^\infty a^{2\nu} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a^2\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right) da}{\int_0^\infty a^{2\nu-1} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a^2\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right) da}. \tag{6.7}$$

Let ${}_1\mathcal{F}_1(a;b;x)$ denote the confluent hypergeometric function [11, Ch. 13]. Using [12, Eqs. 6.643.2, 9.210.1, and 9.220.2] we can solve the integrals for $\nu > 0$ and find

$$\hat{A}^{(2)} = \frac{\Gamma(\nu+0.5)}{\Gamma(\nu)} \sqrt{\frac{\xi}{\zeta(\nu+\xi)}} \frac{{}_1\mathcal{F}_1\left(\nu+0.5;1;\frac{\zeta\xi}{\nu+\xi}\right)}{{}_1\mathcal{F}_1\left(\nu;1;\frac{\zeta\xi}{\nu+\xi}\right)} r, \tag{6.8}$$

where we have inserted in Eq. (6.8) the relation between β and $E\{A^2\}$ given in Eq. (A.2). This result has also recently been derived in [3].

6.4.2 DFT Magnitudes, γ=1

Let the superscript (1) denote that $\gamma = 1$. Substitution of Eqs. (6.1) with $\gamma = 1$ and (6.6) into Eq. (6.5) then gives the following expression for the amplitude estimator

$$\hat{A}^{(1)} = \frac{\int_0^\infty a^{\nu} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right) da}{\int_0^\infty a^{\nu-1} \exp\left(-\frac{a^2}{\sigma_D^2} - \beta a\right) \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right) da}. \tag{6.9}$$

Unfortunately, no closed-form solutions are known. Therefore, we introduce two approximations of the Bessel function in Eq. (6.9). One of these approximations is most accurate at low SNRs , while the other is most accurate at high SNRs. With these approximations, the integrals can be solved in closed-form. Before discussing the approximations we introduce a change of variable that makes it more clear to see under which conditions the various approximations are expected to be accurate.

Change of Variable

For convenience the following transformation is made: $w = 2ar/\sigma_D^2$. In addition, we make use of the relation in Eq. (A.1) between β and the second moment of A, $\beta = \sqrt{\nu(\nu+1)}/\sigma_X$. The expression for $\hat{A}^{(1)}$ now becomes

$$\hat{A}^{(1)} = \frac{\sigma_D^2}{2r} \frac{\int_0^\infty w^\nu \exp\left[-\frac{w^2}{4\zeta} - \frac{\mu w}{2\sqrt{\zeta\xi}}\right] \mathcal{I}_0(w)\, dw}{\int_0^\infty w^{\nu-1} \exp\left[-\frac{w^2}{4\zeta} - \frac{\mu w}{2\sqrt{\zeta\xi}}\right] \mathcal{I}_0(w)\, dw}, \tag{6.10}$$

where we have introduced $\mu = \sqrt{\nu(\nu+1)}$. The approximations of Eq. (6.10) discussed below concern the Bessel function $\mathcal{I}_0(w)$. The function $w^\nu \exp[-\frac{w^2}{4\zeta} - \frac{\mu w}{2\sqrt{\zeta\xi}}]$ attains its maximum at a small value of w when the exponentials decay fast and w^ν rises slowly. In this case it is especially important to approximate the Bessel function well at small arguments. This happens when ζ and/or $\sqrt{\zeta\xi}$ are small, i.e., at low SNRs. Note that ζ is the more dominant parameter of ζ and ξ, because ξ is not present in the quadratic term in the exponentials. In other cases, namely high SNR conditions, the Bessel function should be accurately approximated for large arguments.

Approximation of $E\{A|r\}$, Low SNR Conditions

For low SNR conditions we approximate $\mathcal{I}_0$ by a Taylor series expansion around $w = 0$. The Taylor series of $\mathcal{I}_0$, truncated after N terms, is given by [11, Eq. 9.6.10]

$$\mathcal{I}_0(w; N) = \sum_{n=0}^{N-1} \left(\frac{w}{2}\right)^{2n} \frac{1}{(n!)^2}. \tag{6.11}$$

Fig. 6.4a shows $\mathcal{I}_0$ and several truncated Taylor series expansions. We see that for small arguments, $\mathcal{I}_0$ is approximated well by only a few terms. Let $\mathcal{D}_\nu(\cdot)$ denote the parabolic cylinder function of order ν [11, Ch. 19]. Substituting Eq. (6.11) into Eq. (6.10) and using [12, Eq.3.462.1] gives us an estimator, $\hat{A}^{(1)}_{\ll,N}$, which is most accurate for low SNRs, that is,

$$\hat{A}^{(1)}_{\ll,N} = \frac{1}{\sqrt{2\zeta}} \frac{\sum_{n=0}^{N-1} \left(\frac{1}{n!}\right)^2 \left(\frac{\zeta}{2}\right)^n \Gamma(\nu+2n+1) \mathcal{D}_{-(\nu+2n+1)}\left(\frac{\mu}{\sqrt{2\xi}}\right)}{\sum_{n=0}^{N-1} \left(\frac{1}{n!}\right)^2 \left(\frac{\zeta}{2}\right)^n \Gamma(\nu+2n) \mathcal{D}_{-(\nu+2n)}\left(\frac{\mu}{\sqrt{2\xi}}\right)} r. \tag{6.12}$$

The subscript $\ll,N$ indicates that the approximation from Eq. (6.11) uses N terms and is valid at low SNRs.

For $N \to \infty$, $\hat{A}^{(1)}_{\ll,N}$ converges to $\hat{A}^{(1)}$. This is because the Taylor expansion in Eq. (6.11) converges for all w and because changing the order of integration and summation as is used in the derivation of Eq. (6.12) is allowed for $N \to \infty$ according to Fubini's theorem [13].

Approximation of $E\{A|r\}$, High SNR Conditions

Using the approximate estimator $\hat{A}^{(1)}_{\ll,N}$ under high SNR conditions, requires the number of terms N in the Taylor expansion to be large for an accurate result. Large N leads to a high computational load and numerical problems may result when evaluating Eq. (6.12). In order to avoid these complications, we investigate an approximation of Eq. (6.10) that is more accurate under high SNR conditions. We apply the following well-known large-argument approximation of $\mathcal{I}_0$ [11, Eq. 9.7.1]

$$\mathcal{I}_0(w) \sim \frac{1}{\sqrt{2\pi w}} \exp(w). \tag{6.13}$$

Fig. 6.4b shows $\mathcal{I}_0$ and its approximation for large arguments. Substituting this approximation in Eq. (6.10) and using [12, Eq. 3.462.1] we find for $\nu > 0.5$:

$$\hat{A}^{(1)}_{\gg} = \frac{(\nu - 1/2)}{\sqrt{2\zeta}} \frac{D_{-(\nu+1/2)}\left(\frac{\mu}{\sqrt{2\xi}} - \sqrt{2\zeta}\right)}{D_{-(\nu-1/2)}\left(\frac{\mu}{\sqrt{2\xi}} - \sqrt{2\zeta}\right)} r. \tag{6.14}$$

The approximation in Eq. (6.14) is most accurate when ζ and $\sqrt{\zeta\xi}$ are large and ν is large too.

6.4.3 Combining the Estimators

In this section we present a procedure that can be used to decide which of the two approximations to use under which circumstances.

From Eq. (6.10) it is clear that the faster the exponential term in the integrals decreases, the less important it becomes how well the Bessel function is approximated for large values of w. So, generally speaking, the approximation for small arguments is most accurate for low SNRs. The approximation Eq. (6.14) is more accurate for high SNRs and large ν. Fortunately, the behavior of the approximations is such that a simple binary decision strategy can be found that leads to good results. Specifically, it turns out that taking the maximum of Eq. (6.12) and Eq. (6.14) is generally a good approximation of $\hat{A}^{(1)}$ for, say, $N > 4$. This procedure is motivated as follows. In [14] it was shown that the approximation for low SNRs, $\hat{A}^{(1)}_{\ll,N}$, is always smaller than $\hat{A}^{(1)}$, for all N. The approximation for high SNRs, $\hat{A}^{(1)}_{\gg}$, can be both smaller and larger than $\hat{A}^{(1)}$, depending on the values of the parameters. As we will show by simulation experiments in Section 6.4.4, however, $\hat{A}^{(1)}_{\gg}$ can be only slightly larger than $\hat{A}^{(1)}$. It turns out that the combined estimator $\hat{A}^{(1)}_{C,N} = \max\left[\hat{A}^{(1)}_{\ll,N}, \hat{A}^{(1)}_{\gg}\right]$ obtained from a

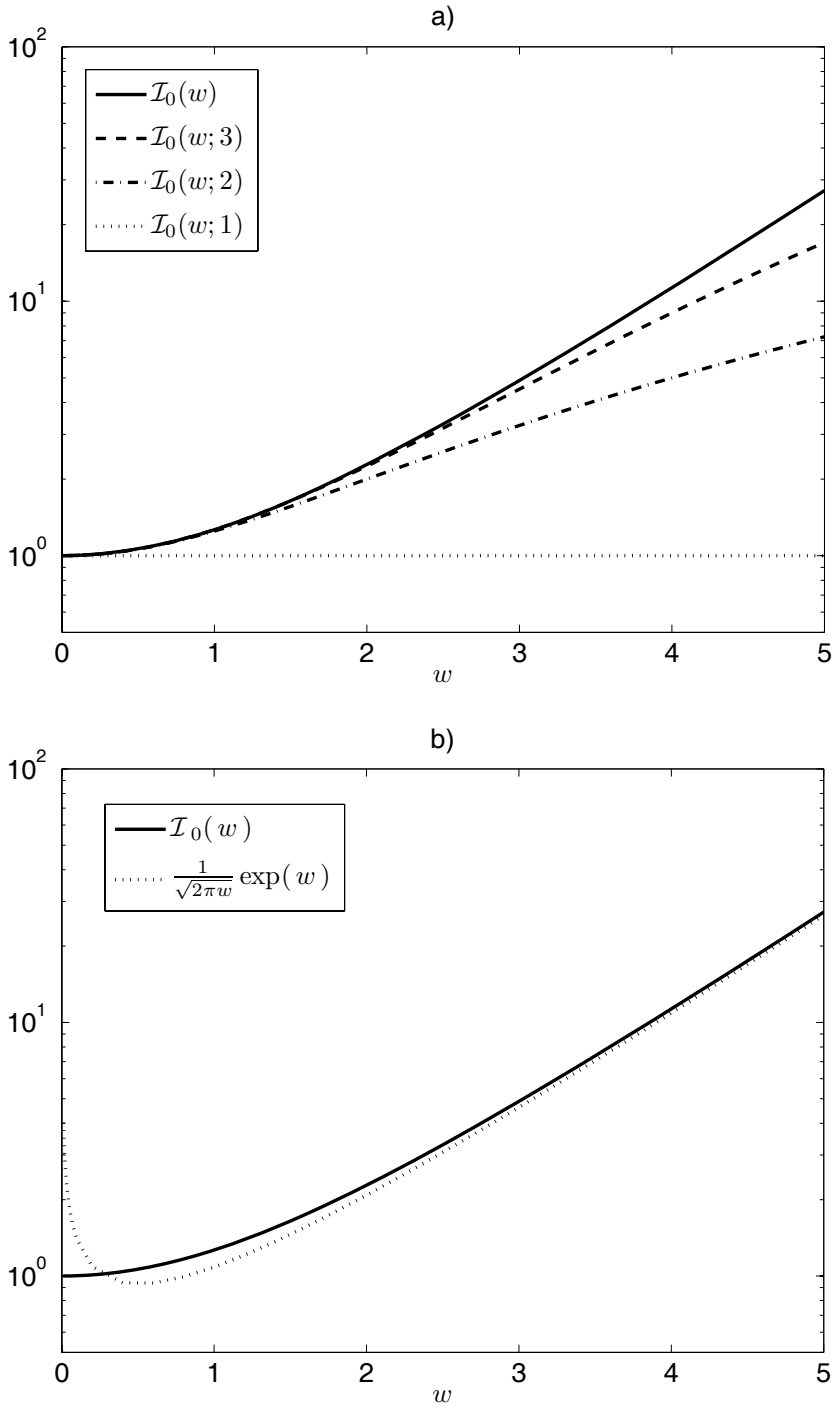

Figure 6.4: *$\mathcal{I}_0$ and its approximations. a) Taylor series expansion for small arguments, b) approximation for large arguments.*

simple binary decision leads to an accurate approximation of $\hat{A}^{(1)}$ for all values of ζ, ξ, and ν. To illustrate the practical use of the decision rule $\hat{A}^{(1)}_{C,N} = \max\left[\hat{A}^{(1)}_{\ll,N}, \hat{A}^{(1)}_{\gg}\right]$ we computed Eq. (6.10) by numerical integration, denoted by $\hat{A}^{(1)}_{MMSE}$. Let the gain G be defined as $G = \hat{A}/r$. In Fig. 6.5 gain curves versus the *a posteriori* SNR ζ are shown for $\nu = 0.6$ and for several values of ξ. In each plot we show $G^{(1)}_{\ll,5} = \hat{A}^{(1)}_{\ll,5}/r$, $G^{(1)}_{\gg} = \hat{A}^{(1)}_{\gg}/r$ and $G^{(1)}_{MMSE} = \hat{A}^{(1)}_{MMSE}/r$ and see that taking the maximum of $G^{(1)}_{\ll,5}$ and $G^{(1)}_{\gg}$ leads to a gain $G^{(1)}_{C,5}$ close to $G^{(1)}_{MMSE}$.

6.4.4 Experimental Analysis of Errors Due to Approximations

The errors in the combined estimator have been investigated for the range $0.6 \leqslant \nu \leqslant 3.2$, $-20\ dB \leqslant \xi \leqslant +20\ dB$, $-20\ dB \leqslant \zeta \leqslant +14\ dB$. For this range, $\hat{A}^{(1)}_{MMSE}$ could be evaluated numerically. For larger ζ, the accuracy of $\hat{A}^{(1)}_{\gg}$ only increases, so its error will be smaller. For the binary decision $\max[\hat{A}^{(1)}_{\ll,5}, \hat{A}^{(1)}_{\gg}]$, the maximum positive error was $+3.7$ dB, and the maximum negative error was -0.2 dB. A positive error means that $\hat{A}^{(1)}_{MMSE}$ was larger than the approximate gain function. Using $\max[\hat{A}^{(1)}_{\ll,20}, \hat{A}^{(1)}_{\gg}]$ the maximum positive error decreases to $+0.1$ dB. The largest positive errors occur for the lowest values of ν, and, for a given value of *a priori* SNR, only for a small range of *a posteriori* SNRs, as can be seen in Fig. 6.5. These results show that the simple binary decision procedure generally works well.

6.4.5 Computational Complexity

MATLAB implementations of the algorithms in [15] (implementations available from [16]) have been used to evaluate the parabolic cylinder and confluent hypergeometric functions. We have adapted these programs so that they can handle vector arguments. The estimator for $\gamma = 2$ can be evaluated in real-time on a PC with a Pentium 4 processor. The combined estimator for $\gamma = 1$ is more complex, because two estimators have to be evaluated and because of the sums in Eq. (6.12). The sums can be efficiently computed by making use of recursive relations [11, Eq. 19.6.4]. In this way, the combined estimator can be evaluated in 2-3 times real-time. In a practical system, for a fixed value of ν, all gain functions can be evaluated off-line for the relevant range of the parameters and stored in a table. Computational complexity is not an issue then.

6.5 MMSE Estimation of Complex DFT Coefficients

In this section we derive the MMSE estimator of the clean speech DFT coefficient X. Assuming that the real and imaginary parts of X, $X_{\Re}$ and $X_{\Im}$, are statistically independent, it follows that [7]

$$E\{X|y\} = E\{X_{\Re}|y_{\Re}\} + jE\{X_{\Im}|y_{\Im}\}. \tag{6.15}$$

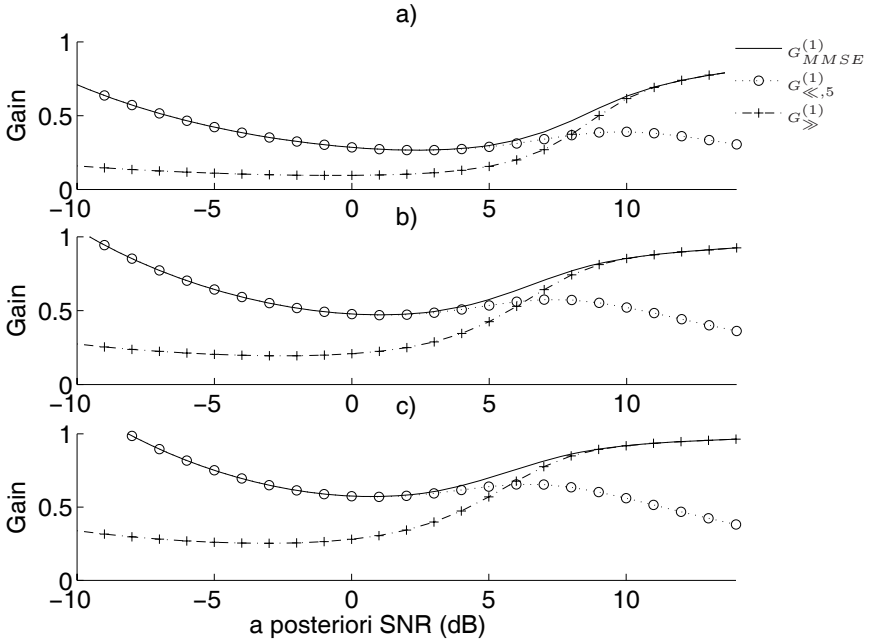

Figure 6.5: *Comparison of gain functions for amplitude estimators $G^{(1)}_{\ll,5}$ Eq. (6.12), $G^{(1)}_{\gg}$ Eq. (6.14), and $G^{(1)}_{MMSE}$ for $\nu = 0.6$ and a) $\xi = -5$ dB, b) $\xi = +5$ dB, and c) $\xi = +15$ dB.*

We now consider estimation of $X_{\Re}$. A similar procedure can be followed for $X_{\Im}$. We have

$$E\{X_{\Re}|y_{\Re}\} = \frac{\int_{x_{\Re}} x_{\Re} f_{Y_{\Re}|x_{\Re}}(y_{\Re}|x_{\Re}) f_{X_{\Re}}(x_{\Re}) dx_{\Re}}{\int_{x_{\Re}} f_{Y_{\Re}|x_{\Re}}(y_{\Re}|x_{\Re}) f_{X_{\Re}}(x_{\Re}) dx_{\Re}}. \tag{6.16}$$

Using the Gaussian noise assumption it follows

$$f_{Y_{\Re}|x_{\Re}}(y_{\Re}|x_{\Re}) = (2\pi\sigma^2_{D_{\Re}})^{-\frac{1}{2}} \exp(-\frac{1}{2\sigma^2_{D_{\Re}}}(y^2_{\Re} + x^2_{\Re} - 2y_{\Re}x_{\Re})). \tag{6.17}$$

6.5.1 Complex DFTs, $\gamma = 1$

Using Eq. (6.2) with $\gamma = 1$, [12, Eq. 3.462.1] and the relation between β and $\sigma^2_{X_{\Re}}$ given by Eq. (A.3) in Appendix A, we find the following expression for the conditional mean [17][18]

$$E\{X_{\Re}|y_{\Re}\} = \sigma_{D_{\Re}} \nu \frac{\exp(\frac{1}{4}y^2_-)\mathcal{D}_{-(\nu+1)}(y_-) - \exp(\frac{1}{4}y^2_+)\mathcal{D}_{-(\nu+1)}(y_+)}{\exp(\frac{1}{4}y^2_-)\mathcal{D}_{-\nu}(y_-) + \exp(\frac{1}{4}y^2_+)\mathcal{D}_{-\nu}(y_+)}, \tag{6.18}$$

where

$$y_{\pm} = \sqrt{\frac{\nu(\nu+1)}{\xi}} \pm \frac{y_{\Re}}{\sigma_{D_{\Re}}}. \tag{6.19}$$

6.5.2 Complex DFTs, γ=2

We now consider the MMSE estimator for $\gamma = 2$. Maintaining the Gaussian noise assumption, the conditional density $f_{Y_{\Re}|x_{\Re}}(y_{\Re}|x_{\Re})$ given in Eq. (6.17) remains valid. Using Eq. (6.2) with $\gamma = 2$ and [12, Eq. 3.462.1], it can be shown that the conditional mean estimator can be written as

$$E\{X_{\Re}|y_{\Re}\} = 2\nu \frac{\sigma_{D_{\Re}}}{\sqrt{1+2\nu\xi^{-1}}} \frac{\mathcal{D}_{-(2\nu+1)}(y_-) - \mathcal{D}_{-(2\nu+1)}(-y_-)}{D_{-2\nu}(y_-) + D_{-2\nu}(-y_-)}, \tag{6.20}$$

where

$$y_- = -\frac{y_{\Re}}{\sigma_{D_{\Re}}} \left(1 + 2\nu\xi^{-1}\right)^{-1/2}. \tag{6.21}$$

It is easy to see that when ν/ξ is small and $\zeta_{\Re}$ is not, the estimators for $\gamma = 1$ (Eq. (6.18)) and $\gamma = 2$ (Eq. (6.20)) are approximately equal when the quantity $\gamma\nu$ has the same value for both estimators.

6.6 Filter Characteristics

In this section we study the input-output characteristics for the DFT magnitude estimators as well as for the complex DFT coefficient estimators.

6.6.1 Magnitudes of DFT Coefficients

Fig. 6.6 shows examples of input-output characteristics for the magnitude estimators. In Fig. 6.6a we consider the case $\gamma = 1$ and $\nu \in \{0.8, 1, 1.5\}$ for the combined estimator $\hat{A}_C^{(1)} = \max[\,\hat{A}_{\ll,5}^{(1)}, \hat{A}_{\gg}^{(1)}]$. In Fig. 6.6b we consider the case $\gamma = 2$ for $\nu \in \{0.5, 1, 1.5\}$. Further, the constraint $\sigma_X^2 + \sigma_D^2 = 2$ is used, and we consider the *a priori* SNRs $\xi = -5$ dB and $\xi = 5$ dB. The input-output characteristics are more sensitive to ν values for the $\gamma = 2$ case than for the $\gamma = 1$ case, and a smaller ν value clearly leads to less suppression at higher input values and to more suppression for lower input values.

6.6.2 Complex DFT Coefficients

Fig. 6.7a shows examples of input-output characteristics of the complex DFT estimators for the case of $\gamma = 1$ and $\nu \in \{0.25, 0.50, 0.75, 1\}$. For $\nu = 1.0$ we recognize the input-output characteristic of the Laplacian (two-sided exponential) prior and for $\nu = 0.5$ we get the input-output characteristics of the two-sided Gamma distribution. For high *a priori* SNRs, the relation between $y_{\Re}$ and the estimator $E\{X_{\Re}|y_{\Re}\}$ is

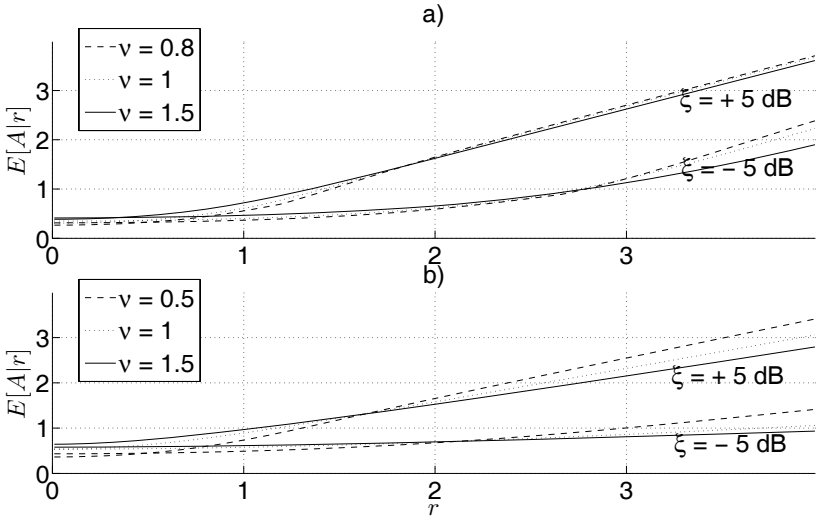

Figure 6.6: *Input-output characteristics for DFT magnitude estimators ($\sigma_X^2 + \sigma_D^2 = 2$). a) $\gamma = 1$ b) $\gamma = 2$.*

almost linear. At low *a priori* SNRs, the relation is non-linear, especially for small values of ν, i.e., more peaked priors.

For the $\gamma = 2$ case we consider $\nu \in \{0.1, 0.2, 0.3, 0.5\}$. Choosing $\nu = 0.5$ gives a Gaussian prior, while lower values of ν correspond to more peaked distributions. Fig. 6.7b shows input-output characteristics for the resulting MMSE estimators. For $\nu = 0.5$ the Wiener estimator occurs (solid line in Fig. 6.7b). For all other choices of ν, the estimators are non-linear in the noisy observation $y_{\Re}$.

6.7 Experimental Results

In this section we present experimental results obtained with the complex DFT and magnitude estimators. For the experiments we use the Noizeus database [19], which consists of 30 IRS-filtered speech signals sampled at 8 kHz, contaminated by various additive noise sources. We added computer-generated telephone-bandwidth white Gaussian noise as an extra noise source, since it is not present in the data base. The frames have a size of 256 samples and are taken with an overlap of 50 %. The decision-directed approach with a smoothing factor $\alpha = 0.98$ was used to estimate ξ [1]. The noise variance was estimated using the minimum statistics approach [20]. Further, for perceptual reasons, in all experiments the maximum suppression was limited to 0.1 [7]. Experiments with a lower limit did not change the conclusions.

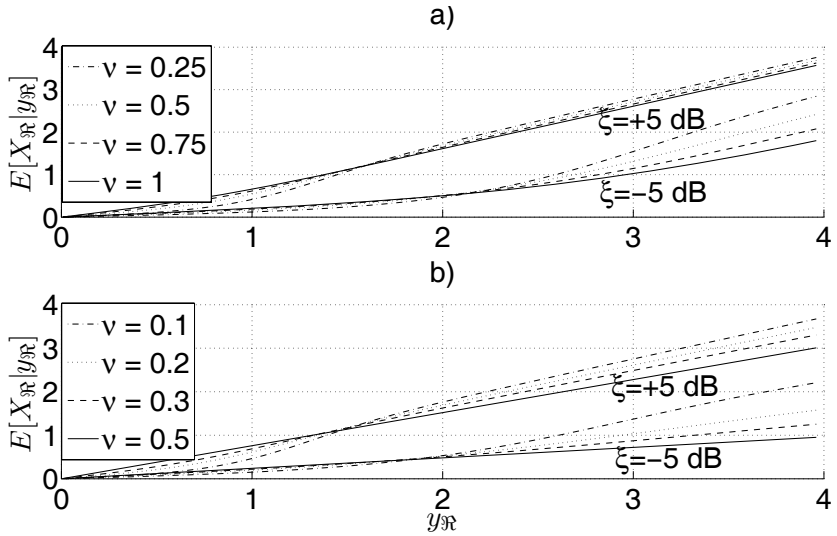

Figure 6.7: *Input-output characteristics for complex DFT estimators ($\sigma_X^2 + \sigma_D^2 = 2$). a) $\gamma = 1$, b) $\gamma = 2$.*

6.7.1 Objective Quality Measures

We measure the performance of the proposed estimators using several objective speech quality measures. First, we introduce the squared error distortion measures

$$D_{\text{ampl}} = \sum_{(k,i)\in\mathcal{Q}} (a(k,i) - \hat{a}(k,i))^2 \tag{6.22}$$

and

$$D_{\text{DFT}} = \sum_{(k,i)\in\mathcal{Q}} |x(k,i) - \hat{x}(k,i)|^2 \tag{6.23}$$

for the magnitude and complex DFT estimators, respectively. Our estimators assume speech presence. In order to avoid contamination of our experimental results by noise-only regions, we discard non-speech frequency bins by using an index set $\mathcal{Q}$ denoting the DFT bins with energy no less than 50 dB below the maximum bin energy in the particular speech signal. These distortion measures evaluate the quantities for which the estimators are optimized.

In an attempt to express the objective performance of the estimators in terms of speech distortion and noise reduction separately, we follow the approach in [5] and measure speech attenuation as

$$\text{SATT}_{\text{seg}} = \frac{1}{|\mathcal{P}|} \sum_{i\in\mathcal{P}} 10\log_{10}\left(\frac{\|\mathbf{x}_t(i)\|_2^2}{\|\mathbf{x}_t(i) - \tilde{\mathbf{x}}_t(i)\|_2^2}\right), \tag{6.24}$$

where the vector $\mathbf{x}_t(i)$ represents a clean speech (time-domain) frame and $\tilde{\mathbf{x}}_t(i)$ is the result of applying the gain functions to the clean speech frame. To discard non-speech

frames, an index set $\mathcal{P}$ is used of all clean speech frames with energy within 30 dB of the maximum frame energy in a particular speech signal. $|\mathcal{P}|$ denotes the cardinality of $\mathcal{P}$. Similarly, noise attenuation is measured as

$$\mathrm{NATT}_{\mathrm{seg}} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} 10 \log_{10} \left(\frac{\|\mathbf{d}_t(i)\|_2^2}{\|\tilde{\mathbf{d}}_t(i)\|_2^2} \right), \tag{6.25}$$

where $\mathbf{d}_t(i)$ is a noise frame, and $\tilde{\mathbf{d}}_t(i)$ is the residual noise frame resulting from applying the noise suppression filter to $\mathbf{d}_t(i)$.

Further, we use segmental SNR defined as

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{T} \left(10 \log_{10} \frac{\|\mathbf{x}_t(i)\|_2^2}{\|\mathbf{x}_t(i) - \hat{\mathbf{x}}_t(i)\|_2^2} \right), \tag{6.26}$$

where $\hat{\mathbf{x}}_t(i)$ denotes an enhanced signal frame, M is the total number of frames, and $\mathcal{T}[y] = \max(\min(y, 35), -10)$, confining the local SNR to a perceptual meaningful range [21]. Finally, we apply the PESQ speech quality measure [22].

6.7.2 Magnitude Estimators

We evaluate the performance of the MMSE amplitude estimator $\hat{A}^{(2)}$ and two approximations of $\hat{A}^{(1)}$, namely $\hat{A}^{(1)}_{\gg}$ and $\hat{A}^{(1)}_{C,5} = \max[\, \hat{A}^{(1)}_{\ll,5}, \hat{A}^{(1)}_{\gg}]$. We included $\hat{A}^{(1)}_{\gg}$ in this comparison to show that the combined estimator $\hat{A}^{(1)}_{C,N}$ has clear advantages over using just the well-known high-SNR approximation used for $\hat{A}^{(1)}_{\gg}$. Further, we make a comparison to a modification of the MAP amplitude estimator as presented in [5], which is in fact a MAP estimator under the generalized gamma distribution Eq. (6.1) with $\gamma = 1$. Details on the modified MAP estimator, which we refer to as $\hat{A}^{(1)}_{MAP}$, can be found in Appendix A.2.

Fig. 6.8 plots D_{ampl} versus ν for speech signals degraded by white noise at SNRs of 0 and 10 dB. We see that $\hat{A}^{(1)}_{C,5}$ improves over $\hat{A}^{(1)}_{\gg}$ and $\hat{A}^{(1)}_{MAP}$, and that $\hat{A}^{(2)}$ does very well for $\nu \approx 0.1$.

Fig. 6.9 shows performance in terms of $\mathrm{SATT}_{\mathrm{seg}}$ versus $\mathrm{NATT}_{\mathrm{seg}}$ for several values of ν and speech signals degraded by white noise at SNRs of 0, 5, 10 and 15 dB. It is shown that for a fixed $\mathrm{NATT}_{\mathrm{seg}}$ performance, $\hat{A}^{(1)}_{C,5}$ often leads to the best speech quality. Furthermore, we see that $\hat{A}^{(2)}$ has the worst $\mathrm{SATT}_{\mathrm{seg}}$ versus $\mathrm{NATT}_{\mathrm{seg}}$ trade-off.

In Fig. 6.10 an evaluation in terms of segmental SNR versus ν is shown for the input SNRs of 5 and 15 dB and speech signals degraded by street noise and white noise. The estimators $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$ and $\hat{A}^{(1)}_{MAP}$ have a comparable performance and are relatively insensitive to ν. The estimator $\hat{A}^{(2)}$ is much more sensitive to ν and shows maximum performance at $\nu \approx 0.1$. The maximum performance of all four estimators $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(1)}_{MAP}$ and $\hat{A}^{(2)}$ is approximately the same.

Fig. 6.11 plots PESQ versus ν for the input SNRs of 5 and 15 dB and speech signals degraded by street noise and white noise. We see that the shape of the graphs

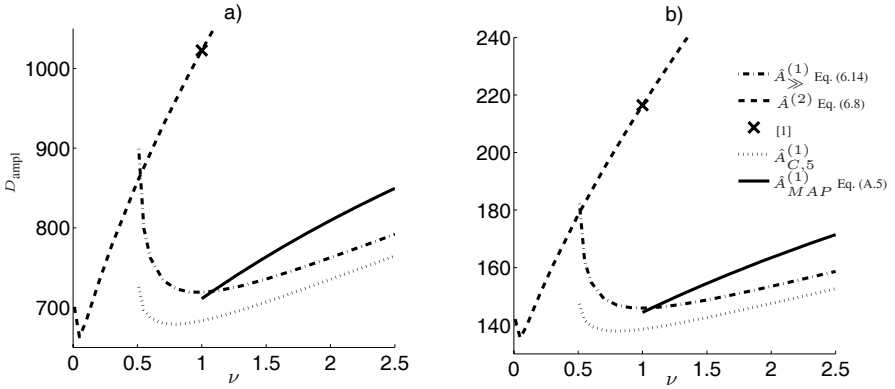

Figure 6.8: *Measured squared error D_{ampl} for white noise with a) SNR = 0 dB. b) SNR = 10 dB.*

representing the performance in terms of PESQ are very similar to the shape of the graphs representing the performance in terms of segmental SNR in Fig. 6.10.

6.7.3 Complex DFT Estimators

We first consider the squared error distortion measure D_{DFT}. Fig. 6.12 plots D_{DFT} versus ν. The estimator based on a Gamma prior [7], i.e. $\gamma = 1$ and $\nu = 0.5$, is indicated by (+) and performs well. Choosing $\nu \approx 0.3$ leads to a slightly better performance. In Fig. 6.12 the amplitude estimators $\hat{A}^{(1)}_{C,5}$ and $\hat{A}^{(2)}$ are evaluated with D_{DFT} as well. Although counterintuitive, it shows that the amplitude estimators perform better as measured by the D_{DFT} distortion measure than the complex DFT estimators. This indicates that the underlying model assumptions for the complex DFT estimators are less valid for natural speech than those of the amplitude estimators.

Fig. 6.13 shows performance in terms of $\mathrm{SATT}_{\text{seg}}$ versus $\mathrm{NATT}_{\text{seg}}$ as a function of ν for the complex DFT estimators and speech signals degraded by white noise. Clearly, the estimator based on the two-sided gamma prior (+) gives relatively low speech distortions (high $\mathrm{SATT}_{\text{seg}}$) for a given residual noise level. Further, the Wiener estimator (x) provides the weakest $\mathrm{SATT}_{\text{seg}}$ vs. $\mathrm{NATT}_{\text{seg}}$ tradeoff; as discussed in Section 6.2 this suggests that the speech distribution *conditional* on the estimated *a priori* SNR is not well described by a Gaussian model. The Gaussian model and thus the Wiener estimator may perform better for a different *a priori* SNR estimator, see e.g. [8]. As expected (see comment after Eq. (6.21)), choosing low ν values leads to similar performance of both classes $\gamma = 1$ and $\gamma = 2$. Comparing with Fig. 6.9, we see that the maximum achievable speech quality in terms of $\mathrm{SATT}_{\text{seg}}$ is lower than for the amplitude estimators.

6.7.4 Subjective Evaluation

Andrianakis and White [3] report that the MMSE amplitude estimators generally perform better than their MAP counterparts, in the sense that weaker speech spectral

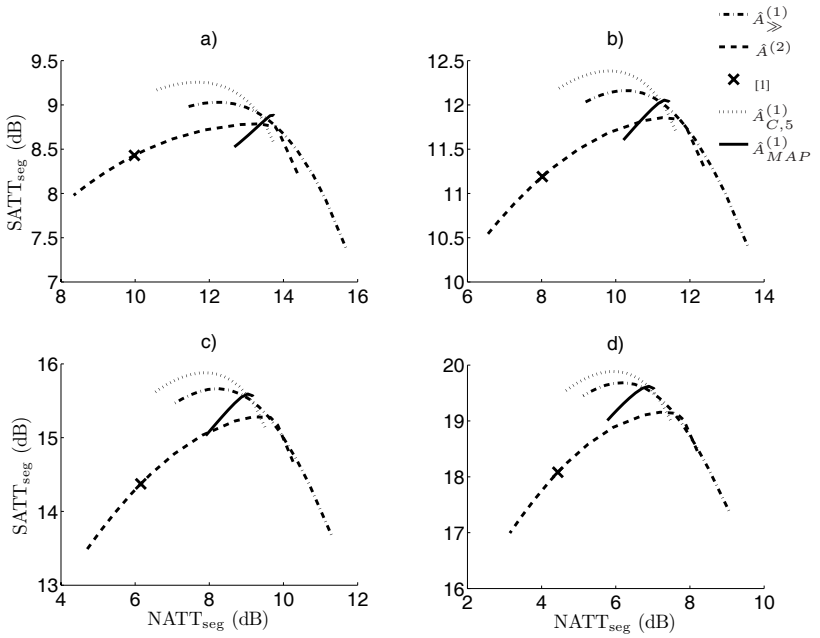

Figure 6.9: $\mathrm{SATT}_{\mathrm{seg}}$ *plotted vs.* $\mathrm{NATT}_{\mathrm{seg}}$ *while varying* ν *for* $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, *the MAP estimator and the estimator of [1] for: a) Input SNR = 0 dB, b) SNR = 5 dB, c) SNR = 10 dB, d) SNR = 15 dB.* ν *decreases along the curves from the left to the right. The range of* ν *values is the same as for Fig. 6.8.*

components are better preserved, while the residual noise has a much more broadband character. The better preservation of weak speech components is confirmed by our informal listening, although the differences between the various estimators are generally small, partly because the maximum suppression was limited to 0.1 for all methods. The combined MMSE estimator $\hat{A}^{(1)}_{C,5}$ introduces slightly less speech distortions than the MAP estimator, $\hat{A}^{(1)}_{\gg}$, and $\hat{A}^{(2)}$. The complex DFT estimators appear to give better noise suppression and seem to introduce slightly less noise artifacts than the amplitude estimators, although at the cost of somewhat higher speech distortions (see also Figs. 6.9 and 6.13). Informal listening tests are in line with the objective results presented above concerning the influence of the parameter ν. For $\gamma = 1$, the perceived quality was rather insensitive to adjustments of ν, while for $\gamma = 2$ changes in ν had a bigger effect: for large values of ν, more residual noise, but less musical noise was present as compared to the smallest values of ν.

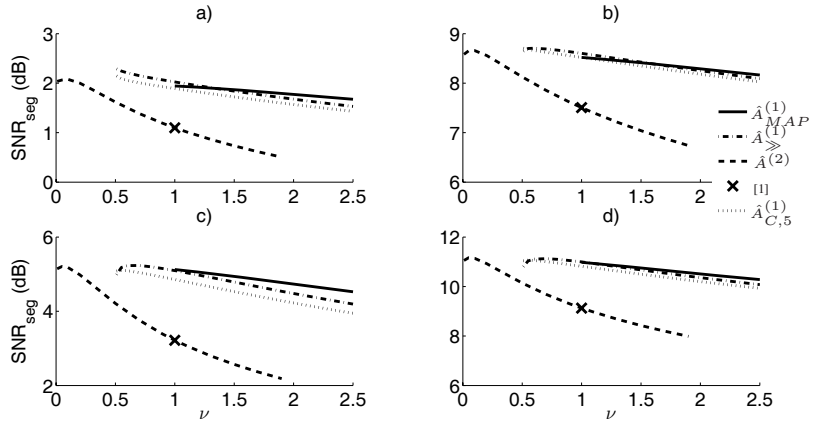

Figure 6.10: *Performance in terms of SNR_{seg} vs. ν for the MAP estimator, $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, and the estimator of [1] for: a) Street noise at input SNR=5 dB. b) Street noise, SNR=15 dB. c) White noise, SNR=5 dB. d) White noise, SNR=15 dB.*

6.8 Concluding Remarks

In this chapter we considered DFT based techniques for single-channel speech enhancement. In the first part, we extended existing MMSE estimators of the *magnitude* estimators for DFT-based noise suppression. The optimal estimators are found under a one-sided generalized Gamma distribution, which takes as special cases (different parameter settings) all priors used in known noise suppression schemes so far. Deriving the MMSE estimators involves integration of (weighted) Bessel functions. In order to find analytical solutions, approximations were necessary for some parameter settings. Ultimately, we combined two types of Bessel function approximations using a simple binary decision between the two. We showed by computer simulations that the estimator thus obtained is very close to the exact MMSE estimator for all SNR conditions. The presented estimators lead to improved performance compared to the suppression rule proposed by Ephraim and Malah [1].

The second part of this chapter dealt with MMSE estimators of *complex* DFT coefficients by deriving two classes of estimators based on generalized gamma prior pdfs. Estimators from the class $\gamma = 1$ typically perform better than the $\gamma = 2$ class, except for small values of the parameter ν, where the estimators are very similar. Applying a complex Gaussian model assumption for the complex speech DFT coefficients clearly leads to suboptimal results. The amplitude estimators performed better than the complex DFT estimators, even under the DFT distortion measure because the modelling assumptions in the complex domain are less accurate than those in the polar domain.

Super-Gaussian priors have been proposed in the literature because they fit better to measured distributions than the Gaussian/Rayleigh priors. These measured distributions are conditional on the estimated spectral variance parameters. However, the super-Gaussian priors still do not perfectly match the measured distributions. The

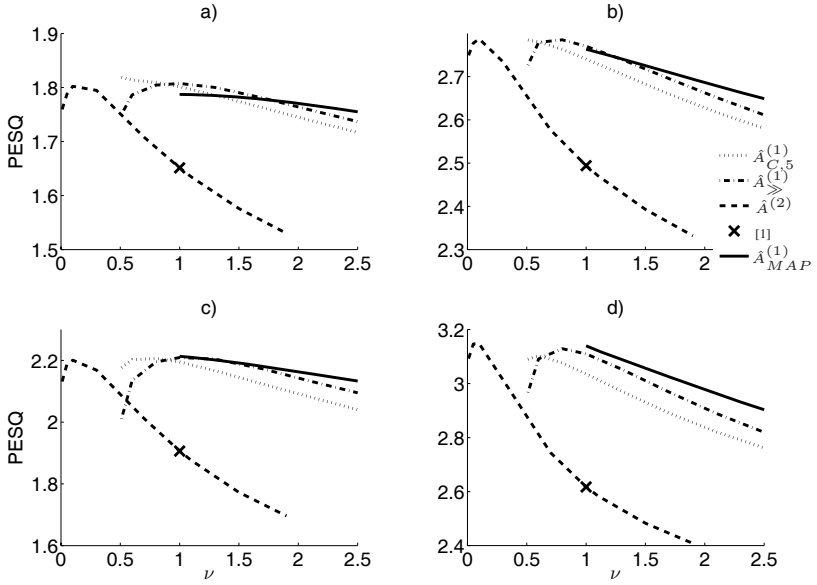

Figure 6.11: *Performance in terms of PESQ (MOS) vs. ν for the MAP estimator, $\hat{A}^{(1)}_{C,5}$, $\hat{A}^{(1)}_{\gg}$, $\hat{A}^{(2)}$, and the estimator of [1] for: a) Street noise at input SNR=5 dB. b) Street noise, SNR=15 dB. c) White noise, SNR=5 dB. d) white noise, SNR=15 dB.*

independency assumption in the complex domain is also inconsistent with the data.

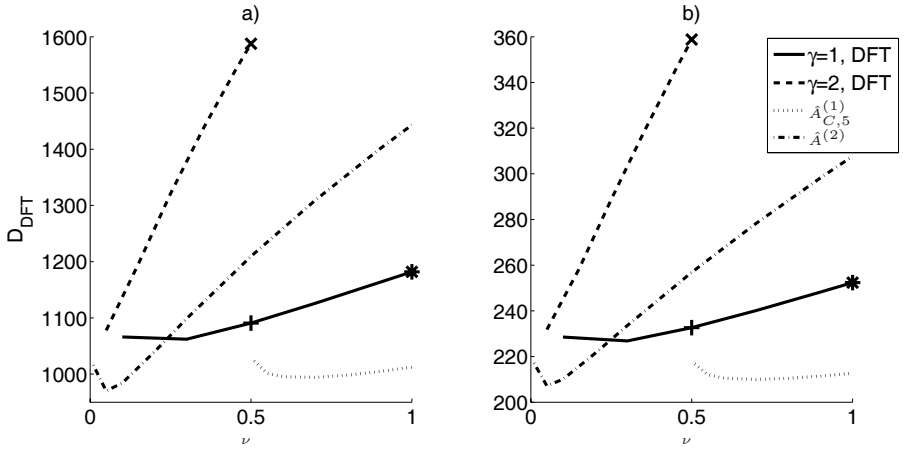

Figure 6.12: *D_{DFT} evaluated on complex DFT estimators with $\gamma = 1$ and $\gamma = 2$ and the amplitude estimators $\hat{A}^{(1)}_{C,5}$ and $\hat{A}^{(2)}$ for white noise with a) SNR = 0 dB. b) SNR = 10 dB. The special cases that correspond to the Gamma, Laplace and Gaussian priors are indicated by $+$, $*$ and $\times$, respectively.*

References

[1] Y. Ephraim and D. Malah. Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[2] P. J. Wolfe and S. J. Godsill. Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement. *EURASIP Journal on Applied Signal Processing*, (10):1043–1051, 2003.

[3] I. Andrianakis and P. R. White. MMSE speech spectral amplitude estimators with chi and gamma speech priors. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, III:1068 – 1071, May 2006.

[4] T. H. Dat, K. Takeda, and F. Itakura. Generalized Gamma modeling of speech and its online estimation for speech enhancement. *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, IV:181 – 184, 2005.

[5] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, 2005.

[6] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*. Principles of Electrical Engineering Series. MIT Press, 1949.

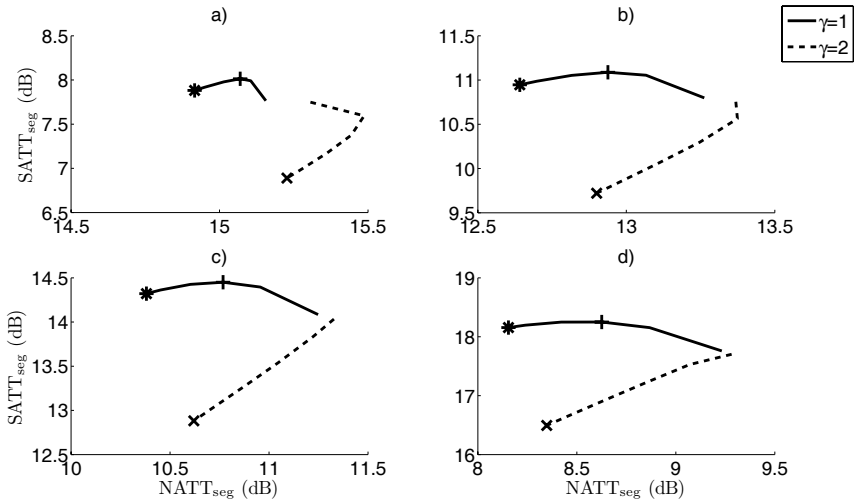

Figure 6.13: $\mathrm{SATT}_{\mathrm{seg}}$ *plotted vs.* $\mathrm{NATT}_{\mathrm{seg}}$ *while varying* ν *for the complex DFT estimators for* $\gamma = 1$ *and* $\gamma = 2$ *for white noise. a) Input SNR = 0 dB, b) SNR = 5 dB, c) SNR = 10 dB, d) SNR = 15 dB.* ν *decreases along the curves from the left to the right. The range of* ν *values is the same as for Fig. 6.12.*

[7] R. Martin. Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors. *IEEE Trans. Speech, Audio Processing*, 13(5):845–856, September 2005.

[8] I. Cohen. Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. *Signal Processing*, 86(4):698–709, 2006.

[9] J. Jensen, I. Batina, R. C. Hendriks, and R. Heusdens. A study of the distribution of time-domain speech samples and discrete fourier coefficients. In *Proc. IEEE First BENELUX/DSP Valley Signal Processing Symposium*, pages 155–158, April 2005.

[10] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall International, Inc., 1992.

[11] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New-York, ninth dover printing, tenth gpo printing edition, 1964.

[12] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, Inc., 6 edition, 2000.

[13] W. Rudin. *Real and complex analysis*. McGraw-Hill, New-York, 3rd ed edition, 1987.

[14] 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 Trans. Audio Speech and Language Processing*, 15(6):1741 – 1752, August 2007.

[15] Shanjie Zhang and Jianming Jin. *Computation of Special Functions*. John Wiley, New York, 1996.

[16] Matlab routines for computation of special functions. `http://ceta.mit.edu/comp_spec_func/`.

[17] J. Jensen, R. C. Hendriks, J. S. Erkelens, and R. Heusdens. MMSE estimation of complex-valued discrete fourier coefficients with generalized gamma priors. In P*roc. Intern. Conf. on Spoken Lang. Proc. – Interspeech 2006*, September 2006.

[18] J. Jensen, R. C. Hendriks, and J. S. Erkelens. MMSE estimation of discrete fourier coefficients with a generalized gamma prior. Technical report, Delft University of Technology, 2006.

[19] Y. Hu and P. Loizou. Subjective comparison of speech enhancement algorithms. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 153–156, May 2006.

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

[21] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

[22] J. G. Beerends. Extending p.862 PESQ for assessing speech intelligibility. *White contribution COM 12-C2 to ITU-T Study Group 12*, October 2004.

Chapter 7

MAP Estimators for Speech Enhancement under Normal and Rayleigh Inverse Gaussian Distributions

This chapter is based on the article published as "MAP Estimators for Speech Enhancement under Normal and Rayleigh Inverse Gaussian Distributions", by R. C. Hendriks and R. Martin in the *IEEE Trans. Speech, Audio and Language Processing*, vol. 15, no. 3, pages 918 - 927, March 2007.

7.1 Introduction

In the previous chapter, estimators were presented under the generalized Gamma density. In order to derive estimators for the complex DFT coefficients under this density we assumed that the real and imaginary parts of DFT coefficients are independent. However, as shown in Fig. 6.3 this assumption is not completely in line with measured speech data. Further, as discussed in Section 6.2 there is no consistency between the models in the complex domain and the polar domain for all parameter settings of the density.

In this chapter we present a class of clean speech estimators based on scale mixtures of normals [1][2] that do not have the above mentioned inconsistencies. Scale mixtures of normals have received an increased amount of interest in several signal processing applications, because they can be used to model non-Gaussian heavy-tailed processes [3][4]. A random variable X is distributed as a scale mixture of normals if it can be expressed as a product of two random variables, that is

$$X = \sqrt{\Lambda_X} Z, \tag{7.1}$$

with $Z \sim N(0,1)$, i.e. a standard normal density, and Λ_X the mixing or scaling random variable, which is drawn from an arbitrary nonnegative distribution. Notice that the random variable $X|\Lambda_X$ is Gaussian with variance Λ_X and that the variance is a random variable. This is different from the generalized Gamma densities used in Chapter 6, where $X|\sigma_X^2$ was assumed to be generalized Gamma distributed and where the variance σ_X^2 was assumed to be deterministic. However, the variance of speech DFT coefficients is unknown in practice and cannot directly be observed, but must be estimated. Moreover, the variance of speech DFT coefficients is not entirely fixed in practice, but shows some variations over time. A reasonable alternative is to take these variations of the variance into account and assume the variance of speech DFT coefficients be to a random variable. This is in line with [5], where Cohen proposed to model the speech variance as a (stochastic) GARCH process. Further, notice that the definition in Eq. (7.1) can be extended to model multivariate processes having non-zero mean and skewness [3]. An example of a scale mixture of normals that obtained increased interest recently is the so-called normal inverse Gaussian (NIG) density that was presented in [6] by Barndorff-Nielsen to model stochastic volatility of heavy-tailed data for financial data modelling applications. More recently, the NIG distribution and its multivariate extension, known as the multivariate NIG (MNIG) [3], have shown to be very suitable to model a large class of heavy-tailed processes [3].

We assume that the complex speech DFT coefficients follow an MNIG distribution. Under this assumption, the speech DFT amplitudes can be shown to be Rayleigh inverse Gaussian (RIG) distributed [7]. Under the MNIG and RIG distribution we derive clean speech MAP estimators for the complex DFT coefficients and speech amplitudes, respectively. The MNIG and RIG distribution parameters of the resulting gain functions can be adapted to the speech signal using the expectation-maximization procedure presented in [3]. Hence, the shape of the assumed density, and consequently the suppression characteristics, can be adapted to the observed distribution of the speech DFT coefficients over time and frequency. Moreover, in contrast to the generalized

Gamma based estimators presented in Chapter 6, the MNIG density is suitable for modelling vector processes as well. As such, the 2-dimensional version of the MNIG based estimator can take the dependency between the real and imaginary part of DFT coefficients into account.

The remaining sections of this chapter are organized as follows. In Section 7.2 the MNIG distribution and its most relevant properties are briefly reviewed. In Section 7.3 we present the assumed speech model and derive MAP estimators for complex DFT coefficients and amplitudes. Further, in Section 7.6 experimental results are presented. Finally, in Section 7.7 conclusions are drawn.

7.2 The Normal and Rayleigh Inverse Gaussian Distribution

In order to facilitate our discussion on the clean speech estimators that we will derive in this chapter, we summarize in this section the relevant properties of the MNIG distribution and derive the RIG distribution. For more detailed information we refer the reader to [6][3].

A d-dimensional MNIG distributed random variable $\mathbf{X}$ is defined as

$$\mathbf{X} = \boldsymbol{\mu} + \Lambda_X \boldsymbol{\Gamma}\boldsymbol{\beta} + \sqrt{\Lambda_X}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{Z}, \tag{7.2}$$

where $\mathbf{Z} \sim N_d(\mathbf{0}, \mathbf{I})$, so that $\mathbf{X}$ given Λ_X has a Gaussian distribution, i.e. $\mathbf{X}|\Lambda_X \sim N_d(\mu + \Lambda_X \boldsymbol{\Gamma}\beta, \Lambda_X \boldsymbol{\Gamma})$, with $\Lambda_X \sim IG(\delta^2, \alpha^2 - \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{\Gamma}\boldsymbol{\beta})$. IG denotes the inverse Gaussian distribution with scalar parameters $\alpha > 0$ and $\delta > 0$, vector parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ and a correlation matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}$, which is assumed to be positive definite. The IG distribution is defined for $\lambda_X > 0$ as

$$\begin{aligned} f_{\Lambda_X}(\lambda_X) = \left(\frac{\delta^2}{2\pi {\lambda_X}^3}\right)^{1/2} \exp\left[\sqrt{\delta^2(\alpha^2 - \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{\Gamma}\boldsymbol{\beta})}\right] \\ \times \exp\left[-\frac{1}{2}\left(\delta^2\lambda_X^{-1} + (\alpha^2 - \boldsymbol{\beta}^{\mathbf{T}}\boldsymbol{\Gamma}\boldsymbol{\beta})\lambda_X\right)\right]. \end{aligned} \tag{7.3}$$

The name inverse Gaussian was introduced by Tweedie [8] who noted an inverse relationship between cumulant generating functions of IG distributions and those of Gaussian distributions.

Let $\mathcal{K}_{(d+1)/2}$ denote the modified Bessel function of the second kind of order $(d+1)/2$. Further, let

$$p(\mathbf{x}) = \delta\sqrt{\alpha^2 - \boldsymbol{\beta}^T\boldsymbol{\Gamma}\boldsymbol{\beta}} + \boldsymbol{\beta}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{7.4}$$

and

$$q(\mathbf{x}) = \sqrt{\delta^2 + [(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Gamma}^{-1}(\mathbf{x} - \boldsymbol{\mu})]}. \tag{7.5}$$

The MNIG distribution of $\mathbf{X}$ can then be computed using [9, Th. 3.471,9] and substitution of Eq. (7.3) into

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_{\mathbf{X}}) f_{\Lambda_X}(\lambda_X) d\lambda_X \tag{7.6}$$

leading to

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= \frac{\delta}{2^{(d-1)/2}} \left(\frac{\alpha}{\pi q(\mathbf{x})} \right)^{(d+1)/2} \exp\left[p(\mathbf{x})\right] \\ &\times \mathcal{K}_{(d+1)/2}\left[\alpha q(\mathbf{x})\right]. \end{aligned} \tag{7.7}$$

Notice that when d is even, $\mathcal{K}_{(d+1)/2}$ can be written in a closed form expression [3]. The density $f_{\mathbf{X}}(\mathbf{x})$ is parameterized by α, $\boldsymbol{\beta}$, δ and $\boldsymbol{\mu}$. The shape of the density is determined by α such that the smaller α is, the heavier the tails become. Parameter $\boldsymbol{\beta}$ determines the skewness of the density; for $\boldsymbol{\beta} \neq 0$ the density will be asymmetrical. Further, δ is the scale parameter and $\boldsymbol{\mu}$ a translation parameter. The MNIG distribution provides a very flexible model for the clean speech DFT coefficients, where the distribution can be adapted to the speech signal by estimation of the parameters, e.g. using the expectation-maximization algorithm as presented in [3]. Based on the measured histograms of speech DFT coefficients in Chapter 6 we assume in this work that the distribution of $\mathbf{X}$ is symmetrical with zero mean, which means that $\boldsymbol{\mu} = 0$ and $\boldsymbol{\beta} = 0$. Those choices for $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$ will be used in the remainder of this chapter.

Although the MNIG probability density function (7.6) appears to be rather complicated, its cumulant generating function has a relatively simple form, that is,

$$\Psi_{\mathbf{X}}(\boldsymbol{\omega}) = \delta \left[\alpha - \sqrt{\alpha^2 - (j\boldsymbol{\omega})^T \boldsymbol{\Gamma} (j\boldsymbol{\omega})} \right]. \tag{7.8}$$

From the cumulant generating function it follows that the Gaussian distribution is a limiting distribution of $f_{\mathbf{X}}(\mathbf{x})$ when $\alpha \to \infty$. This also becomes clear from the MNIG pdf in (7.6) when observing the shape of the IG distribution for increasing α in Fig. 7.1. In Fig. 7.1 it is shown for $\delta = 1$ that when α becomes larger, the IG distribution becomes more and more peaked and will become a delta impulse for $\alpha \to \infty$. Therefore, (7.6) is asymptotically equivalent to a Gaussian distribution. Furthermore, from the cumulant generating function it follows that the covariance matrix of the MNIG distribution is given by [6][3] (for $\boldsymbol{\mu} = 0$ and $\boldsymbol{\beta} = 0$)

$$\boldsymbol{\Sigma} = \frac{\delta}{\alpha} \boldsymbol{\Gamma}. \tag{7.9}$$

In Fig. 7.2 we show some examples of an MNIG probability density function $f_{\mathbf{X}}(\mathbf{x})$ for $\boldsymbol{\beta} = \boldsymbol{\mu} = 0$, $\boldsymbol{\Sigma} = \mathbf{I}$, $d = 1$ and several combinations of δ and α. Comparing this with a zero-mean Gaussian distribution, we see that the NIG distribution approximates the Gaussian distribution as α gets larger. Further, the NIG distributions become more peaked and heavy-tailed as α becomes smaller.

The RIG distribution can be derived from the 2-dimensional MNIG distribution by a transformation of (7.6) into polar coordinates [7]. Consider a 2 dimensional

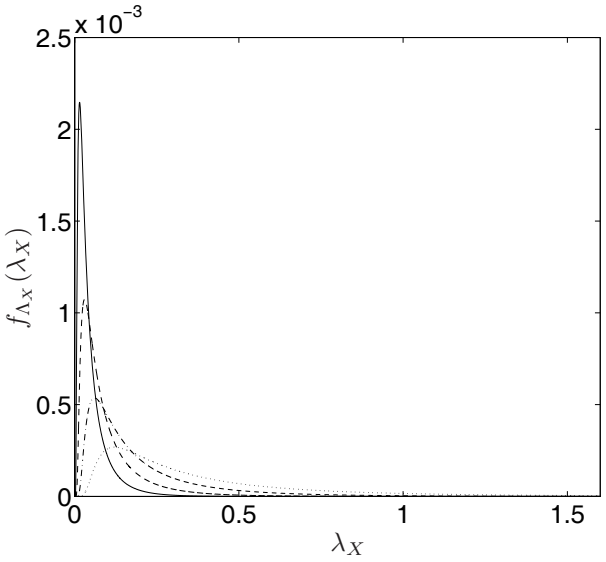

Figure 7.1: *IG distribution for* (δ, α)*=(1,20) (solid line),* (δ, α)*=(1,10) (dashed line),* (δ, α)*=(1,5) (dash-dotted line) and* (δ, α)*=(1,2.5) (dotted line).*

vector $\mathbf{X} = \mathbf{X} = [X_\Re, X_\Im]$, with $\mathbf{X} \sim MNIG(\delta, \alpha, \boldsymbol{\mu}, \boldsymbol{\beta}) = MNIG(\delta, \alpha, 0, 0)$ and $\boldsymbol{\Gamma} = \mathbf{I}$, i.e. $X_\Re$ and $X_\Im$ are assumed to be uncorrelated. Then, the distribution of the amplitude $A = \sqrt{X_\Re^2 + X_\Im^2}$ of $\mathbf{X}$ is a scale mixture of Rayleigh distributions. Indeed, consider the 2-dimensional case of (7.6), that is

$$f_{\mathbf{X}}(x_\Re, x_\Im) = \frac{\delta}{\sqrt{2}} \left(\frac{\alpha}{\pi \sqrt{\delta^2 + x_\Re^2 + x_\Im^2}} \right)^{3/2} \exp\left[\delta\alpha\right] \times \mathcal{K}_{3/2}\left[\alpha \sqrt{\delta^2 + x_\Re^2 + x_\Im^2}\right]. \tag{7.10}$$

Transformation of (7.6) into polar coordinates with $X_\Re = A\cos(\Phi)$ and $X_\Im = A\sin(\Phi)$, Jacobian A and integration over Φ then gives

$$f_A(a) = \int_0^{2\pi} f_{A,\Phi}(a, \phi) d\phi = \frac{a\sqrt{2}\alpha^{3/2}\delta}{\sqrt{\pi}(\delta^2 + a^2)^{3/4}} \exp\left[\delta\alpha\right] \times \mathcal{K}_{3/2}\left[\alpha\sqrt{\delta^2 + a^2}\right], \tag{7.11}$$

which, in analogy to (7.6), can be written as

$$f_A(a) = \int_{\lambda_X} f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X, \tag{7.12}$$

with $f_{A|\Lambda_X}(a|\lambda_X) = \frac{a}{\lambda_X} \exp\left[-\frac{a^2}{2\lambda_X}\right]$ the Rayleigh distribution and $f_{\Lambda_X}(\lambda_X)$ as in (7.3).

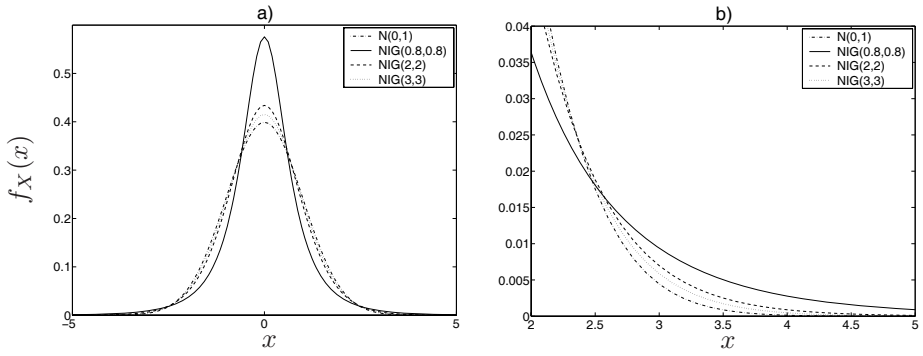

Figure 7.2: *The Normal inverse Gaussian distribution for several values of δ and α, and the Gaussian distribution. The abbreviations NIG and N in the legend indicate the normal inverse Gaussian and the Gaussian distribution, respectively.*

7.3 Speech Models and Distributions

In the remaining part of this chapter we consider DFT-domain speech enhancement where we assume an additive noise model, i.e. $Y(k,i) = X(k,i) + D(k,i)$, where Y is a noisy speech DFT coefficient, X a clean speech DFT coefficient, D a noise DFT coefficient, k the frequency index and i the time frame index. The DFT coefficients Y, X and D are assumed to be complex zero-mean random variables with X and D uncorrelated, i.e. $E[X(k,i)D(k,i)] = 0 \; \forall \, k,i$. We assume that $X(k,i)$ and $D(k,i)$ are statistically independent across time and frequency, which allows us to increase readability by leaving out the time/frequency indices. We assume the noise DFT coefficients to have a complex Gaussian distribution, as is argued for in Section 2.3. The real and imaginary part of the noise DFT coefficients are assumed to be independent and identically distributed with

$$\sigma_D^2 = \sigma_{D_\Re}^2 + \sigma_{D_\Im}^2, \text{ and } \sigma_{D_\Re}^2 = \sigma_{D_\Im}^2. \tag{7.13}$$

Here $\Re$ and $\Im$ denote the real and imaginary part of a DFT coefficient, respectively.

Experimental Data for Complex DFT Coefficients

In this section we study how well the MNIG density with preselected parameters α and δ fits measured histograms of speech in comparison to the Laplace density, which was reported to provide a much better fit to the histogram of speech DFT coefficients than the Gaussian density [10][11]. Histograms of the real part of speech DFT coefficients are obtained using a procedure similar to [11]. First only DFT coefficients with an estimated *a priori* SNR between 28 dB and 31 dB are selected. To do so, signal frames of 512 samples were taken with 50 % overlap and with a sample frequency of 16 kHz. For all frequency bins separate histograms were measured, normalized to unit variance. Finally, the histograms were averaged over frequencies. Both the Laplace and NIG distributions are fitted to the histogram with normalized variance.

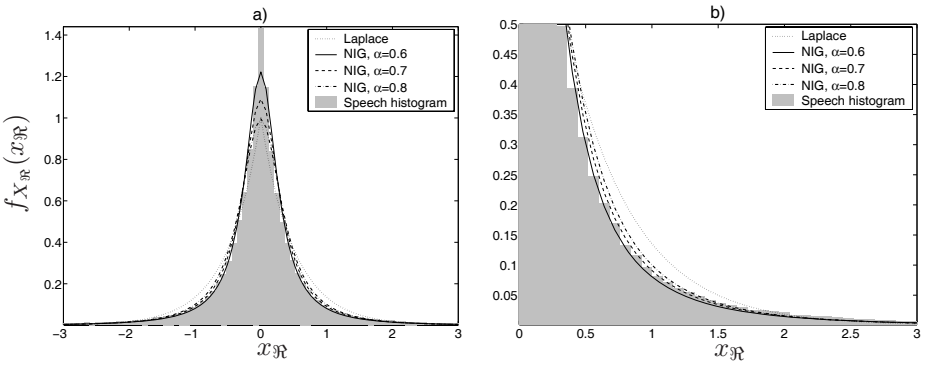

Figure 7.3: *Histogram of speech DFT coefficients and fitted distributions.*

The histogram is shown in Fig. 7.3 with the fitted densities. The figure demonstrates that the Laplace and NIG distribution have similar fit around the tails, but that the NIG has a better fit in between the top and the tail of the histogram.

The better fit is also reflected by the estimated Kullback-Leibler discrimination information [12]

$$I_{KB} = \sum_x f_H(x) log\left(\frac{f_H(x)}{f(x)}\right), \tag{7.14}$$

between the histogram $f_H(x)$ and one of the densities depicted in Fig. 7.3. It turns out that the Kullback-Leibler discrimination measure is about 4.4 times smaller for an NIG distribution with $\alpha = 0.6$ than for the Laplace distribution.

An interesting property of the MNIG distribution is that for a correlation matrix $\mathbf{\Gamma} = \mathbf{I}$, the density is spherically symmetric. This can be of special interest when jointly modelling the real and imaginary parts of DFT coefficients. In Chapter 6, complex DFT estimators were derived under a generalized Gamma density assuming that the real and imaginary parts of DFT coefficients were independent. However, the jointly measured histogram of real and imaginary parts of DFT coefficients in Fig. 6.3 showed that real and imaginary parts of speech DFT coefficients are uncorrelated, but not independent. To investigate the potential use of the 2-dimensional MNIG density for joint estimation of the real and imaginary parts of speech DFT coefficients we show in Fig. 7.4a the same measured histogram as in Fig. 6.3a. In Fig. 7.4b we show contour lines for the joint 2-dimensional MNIG density using correlation matrix $\mathbf{\Gamma} = \mathbf{I}$. We see that, similar as for the estimated density in Fig. 7.4a, indeed the MNIG density is spherical invariant. Furthermore, we see that the increase in height of the contour lines of the 2-dimensional MNIG density matches the measured histogram, as was also shown for the 1-dimensional case in Fig. 7.3. In Figs 7.5a and 7.5b contour lines of a complex Gaussian pdf and a complex Laplace pdf, respectively, are shown. Here it is assumed that real and imaginary part of speech DFT coefficients are independent, i.e. we show in Fig 7.5 not the joint distribution, but the product of the marginal densities. We see that the complex Gaussian pdf is spherical invariant, while the complex Laplace pdf is not. Further, we see that the height of the contour lines of

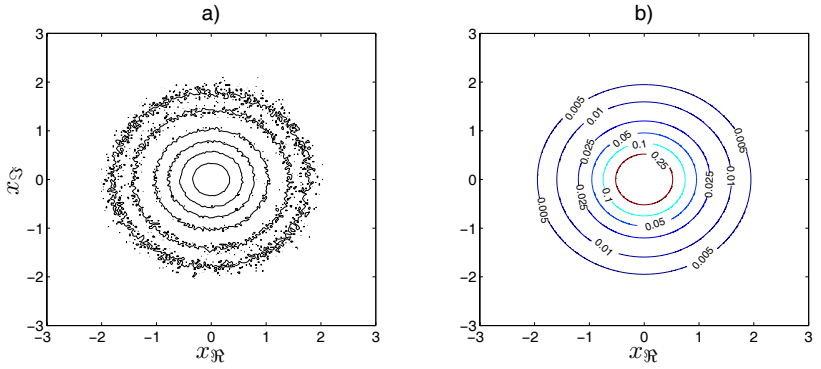

Figure 7.4: *a) Contour lines of jointly measured histogram of real and imaginary parts of speech DFT coefficients, normalized to unit variance. b) Contour lines of 2-dimensional MNIG pdf with* $\alpha = 0.6$ *and* $\mathbf{\Gamma} = \mathbf{I}$*, normalized to unit variance.*

the Gaussian pdf shows a weaker match with the contours of the histogram in Fig. 7.4a than the MNIG density.

Experimental Data for DFT Amplitudes

Fig. 7.6 shows a histogram of measured amplitudes. The data for this histogram is obtained in a similar way as for Fig. 7.3. To this histogram we fit in Fig. 7.6 the RIG distribution for several α values. Moreover, we show the super-Gaussian approximations defined in [13] as

$$f_A(a) = \frac{\mu^{\nu+1} a^\nu}{\mathbf{\Gamma}(\nu+1)\sigma_X^{\nu+1}} \exp\left[-\mu \frac{a}{\sigma_X}\right], \tag{7.15}$$

with $\sigma_X^2 = E\left[A^2\right]$ and parameters $(\nu, \mu) = (1, 2.5)$ and $(\nu, \mu) = (0.126, 1.74)$ that are based on choices made in [13]. The latter set was chosen in [13] to optimize for the dataset used in [13]. From Fig. 7.6 we conclude that especially for amplitudes in the range of $1 \leq A \leq 2$ the RIG distribution shows a better fit than the two super-Gaussian approximations.

The Kullback-Leibler discrimination measure (7.14) is about the same for the RIG distribution with $\alpha = 0.6$ and the super-Gaussian distribution with parameter settings $(\nu, \mu) = (0.126, 1.74)$. Compared to the super-Gaussian distribution with parameter settings $(\nu, \mu) = (1, 2.5)$ the RIG distribution with $\alpha = 0.6$ has a Kullback-Leibler discrimination measure that is more than 7 times smaller.

7.4 MAP Estimator of Complex DFT Coefficients

The complex DFT MAP estimator that we present is rather general and allows to model speech vector processes. Let $\mathbf{Y} \in \mathbb{R}^d$, $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{D} \in \mathbb{R}^d$, whose elements

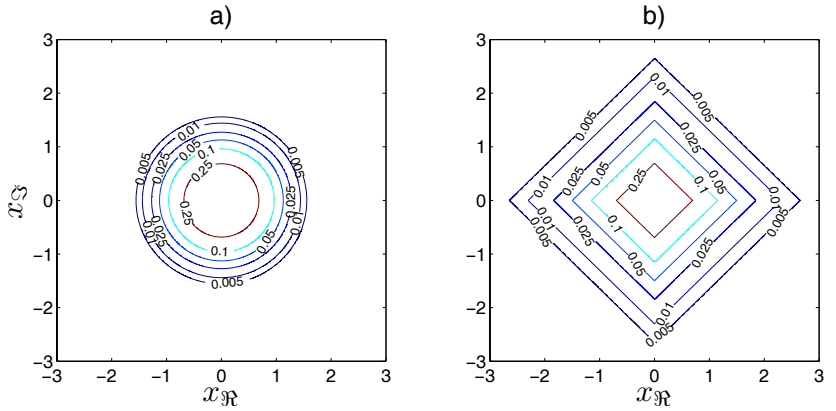

Figure 7.5: *Contour lines using independent cartesian coordinates and normalized to unit variance for a) complex Gaussian pdf b) complex Laplace pdf.*

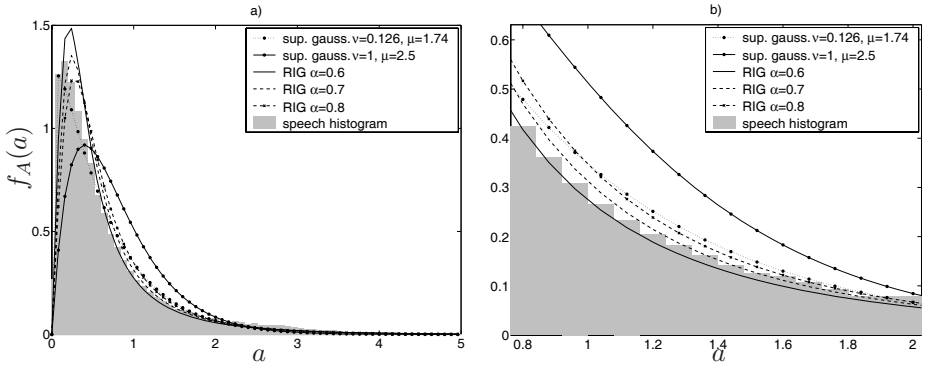

Figure 7.6: *Histogram of speech DFT amplitudes and fitted distributions.*

for example can be the real or imaginary part of a DFT coefficient, respectively, at a frequency bin k ($d = 1$), both the real and imaginary part of DFT coefficients at a frequency bin k ($d = 2$) or even at a series of frequency bins $k = k_1, \ldots, k_2$ ($d = 2(k_2 - k_1 + 1)$). We define the correlation matrix of $\mathbf{D}$ as

$$\boldsymbol{\lambda}_D = E\left[\mathbf{D}\mathbf{D}^H\right] = \text{diag}\left(\sigma^2_{D_1}, \sigma^2_{D_2}, ..., \sigma^2_{D_d}\right).$$

Notice, that when we model the real or imaginary part of a DFT coefficient (i.e., $d = 1$), $\boldsymbol{\lambda}_D = \sigma^2_{D_\Re}$ or $\boldsymbol{\lambda}_D = \sigma^2_{D_\Im}$, respectively. In the case that we model the real and imaginary part jointly, i.e. $d = 2$, then

$$\boldsymbol{\lambda}_D = \begin{pmatrix} \sigma^2_{D_\Re} & 0 \\ 0 & \sigma^2_{D_\Im} \end{pmatrix}. \tag{7.16}$$

The MAP estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ is found, see Section 2.2, by computing

$$\begin{aligned}\hat{\mathbf{x}} &= \arg\max_{\mathbf{x}} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) \\ &= \arg\max_{\mathbf{x}} \frac{f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{Y}}(\mathbf{y})}. \end{aligned} \tag{7.17}$$

Because $f_{\mathbf{Y}}(\mathbf{y})$ is independent of $\mathbf{x}$ and the natural logarithm is a monotonic increasing function, it is sufficient to maximize

$$\ln f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}). \tag{7.18}$$

A Posteriori Distributions for Complex DFT Coefficients

Under the assumption that $\mathbf{D} \sim N_d(0, \boldsymbol{\lambda}_D)$, we can write the distribution of $\mathbf{Y}$ conditioned on $\mathbf{X}$ as

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\lambda}_D|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{x})^T \boldsymbol{\lambda}_D^{-1}(\mathbf{y}-\mathbf{x})\right]. \tag{7.19}$$

We assume that $\mathbf{X}$ is a d-dimensional scale mixture of normals with an MNIG distribution with $\boldsymbol{\mu} = \boldsymbol{\beta} = 0$. This simplifies (7.2) to $\mathbf{X} = \sqrt{\Lambda_X}\boldsymbol{\Gamma}^{\frac{1}{2}}\mathbf{Z}$ with distribution

$$f_{\mathbf{X}}(\mathbf{x}) = \int f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X, \tag{7.20}$$

where

$$f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) = \frac{1}{(2\pi)^{d/2}|\lambda_X\boldsymbol{\Gamma}|^{\frac{1}{2}}} \exp\left[-\frac{1}{2\lambda_X}\mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x}\right], \tag{7.21}$$

and with the mixing distribution $f_{\Lambda_X}(\lambda_X)$ as in (7.3).

Map Estimator

In order to compute a MAP estimate $\hat{\mathbf{x}}$ of $\mathbf{x}$ we first substitute (7.3) and (7.19)-(7.21) in (7.18). Then the derivative of (7.18) is taken with respect to $\mathbf{x}$ using rules for matrix calculation and using [9, Th. 3.471,9]. This gives the following derivative f' (see Appendix B.1 for more details)

$$\begin{aligned} f' &= \boldsymbol{\lambda}_D^{-1}(\mathbf{y}-\mathbf{x}) \\ &\quad - \frac{\int_{\lambda_X} \lambda_X^{-1} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int_{\lambda_X} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \boldsymbol{\Gamma}^{-1}\mathbf{x} \end{aligned} \tag{7.22}$$

$$\begin{aligned} &= \boldsymbol{\lambda}_D^{-1}(\mathbf{y}-\mathbf{x}) - \left(\frac{\alpha^2}{\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x}}\right)^{\frac{1}{2}} \\ &\quad \times \frac{\mathcal{K}_{\frac{3+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})}\right)}{\mathcal{K}_{\frac{1+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})}\right)} \boldsymbol{\Gamma}^{-1}\mathbf{x}. \end{aligned} \tag{7.23}$$

To find the MAP estimate of $\mathbf{x}$, the derivative f' is equated to 0 and solved for $\mathbf{x}$. Unfortunately, it is, to our knowledge, not possible to do this analytically. However, we can compute an approximate solution. Specifically, the ratio of integrals in (7.22) constitutes an MMSE estimate of the inverse first moment of Λ_X, that is

$$E\left[\Lambda_X^{-1}|\mathbf{x}\right] = \frac{\int_{\lambda_X} \lambda_X^{-1} f_{\mathbf{x}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int_{\lambda_X} f_{\mathbf{x}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}. \tag{7.24}$$

Assuming that we are given a pre-estimate of $\mathbf{x}$, denoted by $\tilde{\mathbf{x}}$, we can compute (7.24), that is $E\left[\Lambda_X^{-1}|\tilde{\mathbf{x}}\right]$, subsequently this is substituted in Eq. (7.22) after which we can solve $f' = 0$, leading to

$$\hat{\mathbf{x}} = (\boldsymbol{\lambda}_D^{-1} + E\left[\Lambda_X^{-1}|\tilde{\mathbf{x}}\right] \boldsymbol{\Gamma}^{-1})^{-1} \boldsymbol{\lambda}_D^{-1} \mathbf{y} \tag{7.25}$$

with

$$E\left[\Lambda_X^{-1}|\tilde{\mathbf{x}}\right] = \left(\frac{\alpha^2}{\delta^2 + \tilde{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \tilde{\mathbf{x}}}\right)^{\frac{1}{2}} \frac{\mathcal{K}_{\frac{3+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \tilde{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \tilde{\mathbf{x}})}\right)}{\mathcal{K}_{\frac{1+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \tilde{\mathbf{x}}^T \boldsymbol{\Gamma}^{-1} \tilde{\mathbf{x}})}\right)}. \tag{7.26}$$

For now we assume $\tilde{\mathbf{x}}$ to be known. In section 7.6 we will specify how we obtain $\tilde{\mathbf{x}}$ in practice. Further, notice that the Wiener filter is a special case of (7.25), namely when there is no uncertainty in Λ_X so that $f_{\Lambda_X}(\lambda_X)$ becomes a delta function.

For $\hat{\mathbf{x}}$ in (7.25) to constitute the maximum, it is necessary that the second derivative f'' evaluated at $\hat{\mathbf{x}}$ is negative. Using [9, Th. 8.486,11], the second derivative is given by

$$\begin{aligned} f'' \;=\; & -\boldsymbol{\lambda}_D^{-1} - \frac{\mathcal{K}_{\frac{3+d}{2}}(z)}{\mathcal{K}_{\frac{1+d}{2}}(z)} \\ & \times \left(\frac{\boldsymbol{\Gamma}^{-1}\alpha}{\sqrt{\delta^2 + \mathbf{x}^T \boldsymbol{\Gamma}^{-1} \mathbf{x}}} - \frac{\alpha \mathbf{x}^T \boldsymbol{\Gamma}^{-2} \mathbf{x}}{(\delta^2 + \mathbf{x}^T \boldsymbol{\Gamma}^{-1} \mathbf{x})^{1\frac{1}{2}}}\right) \\ & - \frac{\mathbf{x}^T \boldsymbol{\Gamma}^{-2} \mathbf{x} \alpha^2}{\delta^2 + \mathbf{x}^T \boldsymbol{\Gamma}^{-1} \mathbf{x}} \frac{\mathcal{K}^2_{\frac{3+d}{2}}(z) + \mathcal{K}_{\frac{3+d}{2}}(z) \mathcal{K}_{\frac{-1+d}{2}}(z) - \mathcal{K}^2_{\frac{1+d}{2}}(z)}{2\mathcal{K}^2_{\frac{1+d}{2}}(z)} \\ & + \frac{\mathbf{x}^T \boldsymbol{\Gamma}^{-2} \mathbf{x} \alpha^2}{\delta^2 + \mathbf{x}^T \boldsymbol{\Gamma}^{-1} \mathbf{X}} \frac{\mathcal{K}_{\frac{5+d}{2}}(z)}{2\mathcal{K}_{\frac{1+d}{2}}(z)}, \end{aligned} \tag{7.27}$$

with $z = \alpha\sqrt{\delta^2 + \mathbf{x}^T \boldsymbol{\Gamma}^{-1} \mathbf{x}}$. Unfortunately, the last term of f'' can become positive, which means that f'' can become positive. In practice this does only happen for very low SNR, i.e. when no speech energy is present. In order to overcome this problem, we detect in practice whether f'' is positive. When this occasionally is the case we assume that there is no uncertainty on Λ_X and replace (7.3) in (7.6) by a delta function and make use of (7.9), that is $\Lambda_X = \frac{\delta}{\alpha}$.

Fig. 7.7a shows the input-output characteristics of (7.25) for $d = 1$ applied on the real part of the noisy DFT coefficients $Y_\Re$ for $0 \leq Y_\Re \leq 5$, with $E\left[X_\Re^2\right] + E\left[D_\Re^2\right] =$

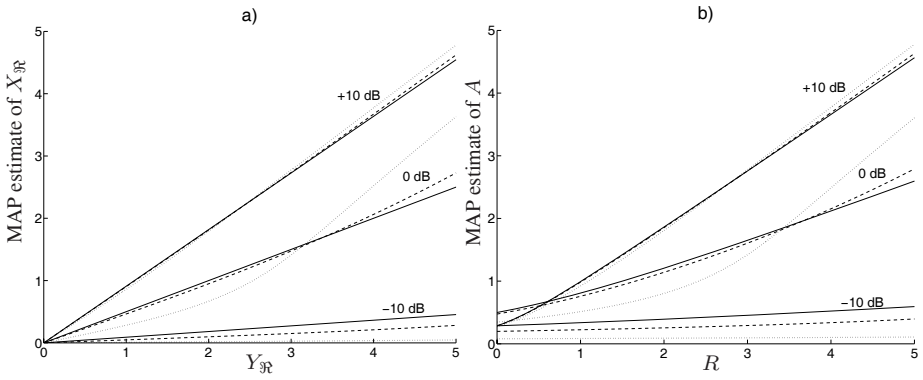

Figure 7.7: *Input versus Output characteristics for a priori SNR values 10 dB, 0 dB and -10 dB and* $\alpha = 3$ *(dashed) and* $\alpha = 1$ *(dotted) for a) the NIG based estimator compared to the Wiener gain (solid) and b) the RIG based estimator compared to the Rayleigh based amplitude estimator (solid).*

2, for the *a priori* SNR values of 10, 0 and -10 dB and several values of α. This is compared with the input-output characteristic of the Wiener filter which assumes that the DFT coefficients are Gaussian distributed. Compared to the Wiener filter the NIG based estimator shows similar characteristics for an *a priori* SNR of 10 dB and high α values, while for lower α values the NIG based estimator leads to less suppression for the higher input values. For an *a priori* SNR value of 0 dB the NIG estimator shows a more pronounced non-linear characteristics. Compared to the Wiener filter there is more suppression for smaller input values and less suppression for larger input values. For an *a priori* SNR of -10 dB the NIG MAP estimator leads to more suppression. Notice that the behavior of the NIG estimator is in principle similar to the super-Gaussian distribution based estimators as presented in [11].

7.5 Map Estimator of DFT Amplitudes

For the amplitude MAP estimator we consider 2-dimensional MNIG distributed vectors $\mathbf{X} = [X_{\Re}, X_{\Im}]^T$ such that $X = X_{\Re} + jX_{\Im} = Ae^{j\Phi}$ with $j = \sqrt{-1}$. Further, $\mathbf{Y} = [Y_{\Re}, Y_{\Im}]^T$ such that $Y = Y_{\Re} + jY_{\Im} = Re^{j\Theta}$ and $\mathbf{D} = [D_{\Re}, D_{\Im}]^T$ with $D = D_{\Re} + jD_{\Im}$.

An amplitude MAP estimator $\hat{a}$ is found, see Section 2.2, by computing

$$\begin{aligned} \hat{a} &= \arg\max_a f_{A|R}(a|r) \\ &= \arg\max_a \frac{f_{R|A}(r|a) f_A(a)}{f_R(r)}. \end{aligned} \tag{7.28}$$

Again, because of the monotonic property of the natural logarithm and the independence of $f_R(r)$ from a it is sufficient to compute,

$$\hat{a} = \arg\max_a \ln\left[f_{R|A}(r|a) f_A(a)\right]. \tag{7.29}$$

A Posteriori Distributions for Amplitudes of Complex DFT Coefficients

The distribution of R given A can be derived from transformation of Eq. (7.19) into polar coordinates as shown in Sec. 2.2, such that

$$f_{R|A}(r|a) = \frac{2r}{\sigma_D^2} \exp\left[-\frac{r^2+a^2}{\sigma_D^2}\right] \mathcal{I}_0\left(\frac{2ar}{\sigma_D^2}\right). \tag{7.30}$$

For $\frac{2ar}{\sigma_D^2} \geq 3$, it is reasonable to approximate $\mathcal{I}_0(x)$ by $\mathcal{I}_0 \approx \frac{1}{\sqrt{2\pi x}} \exp[x]$ [14], such that

$$f_{R|A}(r|a) = \frac{2r}{\sigma_D^2} \exp\left[-\frac{r^2-2ar+a^2}{\sigma_D^2}\right] \sqrt{\frac{\sigma_D^2}{4\pi ar}}. \tag{7.31}$$

We assume that the speech amplitudes A are RIG distributed and its distribution is given by (7.11) and (7.12).

MAP Estimator for Amplitudes

Substitution of (7.11) and (7.31) in (7.29) and taking the derivative with respect to a using [9, Th. 3.471,9] gives the derivative f' (see Appendix B.2 for more details)

$$f' = 2\frac{-a+r}{\sigma_D^2} + \frac{1}{2a} - a\frac{\int \lambda_X^{-1} f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \tag{7.32}$$

$$= 2\frac{-a+r}{\sigma_D^2} + \frac{1}{2a} - a\left(\frac{\alpha^2}{\delta^2+a^2}\right)^{\frac{1}{2}} \frac{\mathcal{K}_{2\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}{\mathcal{K}_{1\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}. \tag{7.33}$$

The amplitude MAP estimator is then given by solving $f' = 0$ for a. As for the complex DFT MAP estimator it is not possible to solve this equation analytically for a. Therefore, we use the approximate solution where the ratio of integrals in (7.32) constitute an MMSE estimate of the inverse first moment of Λ_X, that is

$$E\left[\Lambda_X^{-1}|a\right] = \left(\frac{\alpha^2}{\delta^2+a^2}\right)^{\frac{1}{2}} \frac{\mathcal{K}_{2\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}{\mathcal{K}_{1\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}. \tag{7.34}$$

Given a pre-estimate of a, denoted by $\tilde{a}$, we can compute (7.34) as $E[\Lambda_X^{-1}|\tilde{a}]$. $E[\Lambda_X^{-1}|\tilde{a}]$ is then substituted in (7.32) and subsequently $f' = 0$ is solved for a using elementary calculus leading to

$$\hat{a} = \frac{\frac{2}{\sigma_D^2} + \sqrt{\frac{4}{\sigma_D^4} + 2\left(\frac{2}{r^2\sigma_D^2} + \frac{E[\Lambda_X^{-1}|\tilde{a}]}{r^2}\right)}}{2\left(\frac{2}{\sigma_D^2} + E[\Lambda_X^{-1}|\tilde{a}]\right)} r. \tag{7.35}$$

The second solution that follows from solving $f' = 0$ for a is neglected since that leads to $a < 0$. Notice, that the MAP amplitude estimator proposed in [15] is a special case of (7.35), namely when there is no uncertainty on λ_X and $f_{\Lambda_X}(\lambda_X)$ becomes a delta function.

Again, as with the complex DFT estimator, for $\hat{a}$ in (7.35) to constitute the maximum, it is necessary that the second derivative f'' evaluated at $\hat{a}$ is negative. Using [9, Th. 8.486,11], the second derivative of the amplitude estimator is given by

$$\begin{aligned} f'' &= -\frac{2}{\sigma_D^2} - \frac{1}{2a^2} - \frac{\mathcal{K}_{2\frac{1}{2}}(\alpha\sqrt{\delta^2+a^2})}{\mathcal{K}_{1\frac{1}{2}}(\alpha\sqrt{\delta^2+a^2})} \\ &\quad \times \left(\frac{\alpha}{\sqrt{\delta^2+a^2}} - \frac{\alpha a^2}{\sqrt{\delta^2+a^2}\,(\delta^2+a^2)} \right) - \frac{\alpha^2 a^2}{\delta^2+a^2} \\ &\quad \times \frac{\mathcal{K}_{2\frac{1}{2}}(z)^2 + \mathcal{K}_{2\frac{1}{2}}(z)\mathcal{K}_{\frac{1}{2}}(z) - \mathcal{K}_{1\frac{1}{2}}(z)^2 - \mathcal{K}_{1\frac{1}{2}}(z)\mathcal{K}_{3\frac{1}{2}}(z)}{2\mathcal{K}_{1\frac{1}{2}}(z)^2} \end{aligned} \tag{7.36}$$

with $z = \alpha\sqrt{\delta^2+a^2}$. As for the complex DFT estimator, f'' can become positive when z is very small and the noise level very high. To overcome this problem, we take the same measures as mentioned in Section 7.4.

In Fig. 7.7b the input-output characteristics of the RIG amplitude estimator are shown and compared with the characteristics of the MAP estimator under the Rayleigh distribution as presented in [15]. The characteristics are normalized to $E\left[X_{\Re}^2\right] + E\left[D_{\Re}^2\right] = 2$. The exact characteristics of the RIG based estimator depend on the α parameter. When α gets larger, the characteristics are close to that of the Rayleigh distribution based estimator, while for smaller α values, the characteristics show less suppression for the larger input values, which will preserve speech components, and more suppression for the lower input values which will lead to more noise reduction.

7.6 Experimental Results

In this section we evaluate the performance of the presented clean speech estimators. For evaluation we use segmental SNR defined as [16]

$$\mathrm{SNR}_{\mathrm{seg}} = \frac{1}{N}\sum_{i=0}^{N-1} \mathcal{T}\left\{ 10\log_{10} \frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i) - \hat{\mathbf{x}}_{\mathbf{t}}(i)\|^2} \right\}, \tag{7.37}$$

where $\mathbf{x}_t(i)$ and $\hat{\mathbf{x}}_{\mathbf{t}}(i)$ are time-domain vectors and denote frame i of the clean speech signal $\mathbf{x}_t$ and the enhanced speech signal $\hat{\mathbf{x}}_t$, respectively. N is the number of frames within the speech signal in question and $\mathcal{T}(z) = \min\{\max(z,-10),35\}$ a function which limits the SNR to a perceptually meaningful range.

In addition to $\mathrm{SNR}_{\mathrm{seg}}$ we use intelligibility weighted segmental SNR [17], defined as

$$\mathrm{IWSNR}_{\mathrm{seg}} = \frac{1}{N}\sum_{i=0}^{N-1}\sum_{b=1}^{B} w(b)\mathcal{T}\left\{ 10\log_{10} \frac{\|\mathbf{x}_t(i,b)\|^2}{\|\mathbf{x}_t(i,b) - \hat{\mathbf{x}}_{\mathbf{t}}(i,b)\|^2} \right\}, \tag{7.38}$$

where the weight $w(b)$, emphasizes the importance of the bth frequency band and where $\mathbf{x}_t(i,b)$ is a time-domain vector in band b.

Let $\mathbf{G}(i)$ denote the speech enhancement filter that is used to enhance the ith noisy signal frame. To get an indication whether a difference in $\mathrm{SNR}_{\mathrm{seg}}$ is due to more noise reduction or less speech distortion we process the clean signal and the noise signal with the same filters $\mathbf{G}$ and measure noise attenuation $\mathrm{NATT}_{\mathrm{seg}}$, defined as

$$\mathrm{NATT}_{\mathrm{seg}} = \frac{1}{N}\sum_{i=0}^{N-1} 10\log_{10}\frac{\|\mathbf{d}_t(i)\|^2}{\|\mathbf{G}(i)\mathbf{d}_t(i)\|^2}, \tag{7.39}$$

where $\mathbf{d}(i)$ is a vector and denotes frame i of the noise sequence. Speech attenuation is defined as

$$\mathrm{SATT}_{\mathrm{seg}} = \frac{1}{N}\sum_{i=0}^{N-1} 10\log_{10}\frac{\|\mathbf{x}_t(i)\|^2}{\|\mathbf{x}_t(i)-\mathbf{G}(i)\mathbf{x}_t(i)\|^2}. \tag{7.40}$$

For both $\mathrm{NATT}_{\mathrm{seg}}$ and $\mathrm{SATT}_{\mathrm{seg}}$ only those frame are taken into account where the SNR of the noisy frame i is larger than -10 dB.

Experiments are done with speech signals degraded by white noise, F16 noise, car noise and factory noise, at the input SNRs of 5, 10 and 15 dB. The speech and noise signals originate from the Timit-database [18] and Noisex-92-database [19], respectively. All results are averaged over 32 different speech signals that sampled at 16 kHz. We use a frame size of 512 samples with an overlap of 50% between adjacent frames. Noise statistics are measured using the minimum statistics approach [20]. For perceptual reasons we set the lower bound of the enhancement gain at 0.0316. Increasing this bound will lead to a smaller dynamic range of the estimated gain functions. Decreasing this bound will in general lead to somewhat more suppression. To compute (7.26) and (7.34) we make use of a preliminary estimate of the clean signal by applying a Wiener filter to the noisy speech signal. The parameters α and δ in (7.26) and (7.34) are computed per frame and per frequency bin using the expectation-maximization procedure presented in [3].

7.6.1 Evaluation of 1d-MNIG and RIG Based MAP Estimators

In this section we evaluate the performance of MAP estimators under the MNIG density with $d = 1$, i.e. under the NIG density, and the RIG density and compare that with several other existing estimators.

To evaluate the performance of the MNIG MAP estimator we use (7.25) with $d = 1$ and assume that the real and imaginary part of speech DFT coefficients are independent. We form the estimate $\hat{X}(k,i)$ of the clean speech DFT coefficients by estimation of the real and imaginary parts $\hat{X}_{\Re}$ and $\hat{X}_{\Im}$ of the clean speech DFT coefficients, respectively, and combine them by $\hat{X}(k,i) = \hat{X}_{\Re}(k,i) + j\hat{X}_{\Im}(k,i)$. The NIG based MAP estimator is compared with the Laplace based MMSE estimator [11] and the Wiener filter (notice that these are special cases of the generalized Gamma based MMSE estimator presented in Chapter 6). The results in Table 7.1 show that the improvement of the NIG MAP estimator over the Laplace based MMSE estimator in terms of $\mathrm{SNR}_{\mathrm{seg}}$ varies, dependent on noise source and noise level, from 0.2 to 0.6

dB. Compared to the Wiener filter the improvement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ varies from 0.6 to 1.4 dB.

The performance of the RIG based amplitude MAP estimator in (7.35) is compared with the Rayleigh distribution based MAP amplitude estimator proposed in [15], the super-Gaussian MAP amplitude estimator and the joint MAP amplitude and phase estimator as proposed in [13] with the distribution as in (7.15) with $(\nu, \mu) = (1, 2.5)$ and $(\nu, \mu) = (0.126, 1.74)$, abbreviated with supergauss^1 and supergauss^2, respectively and with the MMSE estimator under the generalized Gamma density as proposed in Chapter 6 with $\gamma = 1$ and $\nu = 0.6$ denoted by $\hat{A}^{(1)}_{C,5}$. Table 7.1 shows that the RIG based MAP estimator has an improvement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ of 0.2 to 0.5 dB over the supergauss^1 estimator and an improvement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ of 0.2 to 0.4 dB over the supergauss^2 estimator. In comparison to the MMSE estimator $\hat{A}^{(1)}_{C,5}$ the improvement is 0.1 to 0.7 dB. The performance difference between the MMSE estimator $\hat{A}^{(1)}_{C,5}$ and the RIG based estimator is more sensitive for the type of noise source than for the supergauss^1 and supergauss^2 estimators. Compared to the estimator under the Rayleigh distribution, the improvement in terms of $\mathrm{SNR}_{\mathrm{seg}}$ is in the order of 0.8 to 1.5 dB.

In addition to segmental SNR we show in Table 7.2 the performance improvement in terms of intelligibility weighted segmental SNR. The MNIG MAP estimator with $d = 1$ has an improvement in terms of $\mathrm{IWSNR}_{\mathrm{seg}}$ of 0.3 to 0.5 dB compared to the Laplace based estimator and 0.6 to 1.2 dB compared to the Wiener filter. The RIG based amplitude MAP estimator has an improvement in terms of $\mathrm{IWSNR}_{\mathrm{seg}}$ of 0.1 to 0.7 dB and 0.1 to 0.5 dB compared to supergauss^1 and supergauss^2, respectively, and an improvement of 1.0 to 1.3 dB compared to the estimator under the Rayleigh distribution. In comparison to the MMSE estimator $\hat{A}^{(1)}_{C,5}$, the improvement is 0.2 to 0.8 dB.

In Table 7.3 and 7.4 we show the $\mathrm{SATT}_{\mathrm{seg}}$ and $\mathrm{NATT}_{\mathrm{seg}}$ scores, respectively. It reveals that the proposed MNIG and RIG MAP estimators in general have a better speech quality in terms of $\mathrm{SATT}_{\mathrm{seg}}$, but a somewhat smaller noise reduction than the estimators based on pre-selected super-Gaussian densities, i.e. the Laplace, supergauss^1 and supergauss^2 based estimators. In terms of $\mathrm{SATT}_{\mathrm{seg}}$ both the RIG based MAP estimator and the MMSE estimator $\hat{A}^{(1)}_{C,5}$ have more or less the same speech quality. However, the RIG based MAP estimator leads to a better performance in terms of $\mathrm{NATT}_{\mathrm{seg}}$.

7.6.2 Evaluation of 2d-MNIG estimator

The experiments with the complex DFT estimator in Section 7.6.1 (MNIG with $d = 1$) were performed using the assumption that the real and imaginary part of DFT coefficients are independent. However, from the discussion in Section 6.2 it followed that it is reasonable to assume the real and imaginary parts of DFT coefficients to be uncorrelated, but not independent. The MNIG based MAP estimator offers a possibility to estimate the real and imaginary part jointly, i.e. take the dependency between real and imaginary parts into account.

input		$\mathrm{SNR}_{\mathrm{seg}}$ (dB) DFT est.			$\mathrm{SNR}_{\mathrm{seg}}$ (dB) Amplitude est.				
noise source	SNR (dB)	NIG	Wiener	Laplace	RIG	Rayleigh	Supergauss^1	Supergauss^2	$\hat{A}_{C,5}^{(1)}$
white	5	6.2	5.2	5.7	6.1	4.9	5.7	5.9	5.6
	10	5.3	4.1	4.8	5.2	3.8	4.7	4.9	4.8
	15	4.3	2.9	3.7	4.2	2.7	3.7	3.9	3.9
F16	5	5.3	4.3	4.8	5.2	4.0	4.8	5.0	4.8
	10	4.5	3.4	4.0	4.4	3.1	4.0	4.1	4.1
	15	3.7	2.4	3.2	3.6	2.2	3.2	3.3	3.4
Car	5	6.9	6.1	6.5	6.9	5.6	6.6	6.6	6.0
	10	5.5	4.5	5.0	5.4	4.2	5.1	5.1	4.7
	15	3.7	2.6	3.2	3.7	2.5	3.2	3.3	3.1
Factory	5	3.8	3.2	3.6	3.7	2.9	3.5	3.6	3.5
	10	3.3	2.4	3.0	3.2	2.2	2.9	3.0	3.1
	15	2.8	1.7	2.4	2.8	1.5	2.3	2.5	2.5

Table 7.1: *Improvement in* $\mathrm{SNR}_{\mathrm{seg}}$ *(dB).*

input		$\mathrm{IWSNR}_{\mathrm{seg}}$ (dB) DFT est.			$\mathrm{IWSNR}_{\mathrm{seg}}$ (dB) Amplitude est.				
noise source	SNR (dB)	NIG	Wiener	Laplace	RIG	Rayleigh	Supergauss^1	Supergauss^2	$\hat{A}_{C,5}^{(1)}$
white	5	4.8	4.1	4.4	4.7	3.4	4.5	4.6	3.9
	10	4.2	3.5	3.9	4.2	2.9	3.9	4.0	3.5
	15	3.5	2.6	3.1	3.4	2.1	3.1	3.2	2.9
F16	5	4.6	3.9	4.2	4.5	3.2	4.3	4.4	3.7
	10	4.0	3.1	3.6	3.9	2.6	3.7	3.8	3.3
	15	3.3	2.3	2.9	3.2	1.9	2.9	3.0	2.8
Car	5	-0.59	-1.8	-1.1	-0.53	-1.7	-1.1	-0.90	-0.85
	10	-1.8	-3.0	-2.3	-1.7	-2.8	-2.3	-2.1	-2.0
	15	-3.1	-4.3	-3.6	-2.9	-3.9	-3.6	-3.4	-3.3
Factory	5	3.6	3.0	3.3	3.5	2.4	3.4	3.3	2.9
	10	3.2	2.4	2.8	3.1	1.9	2.8	2.9	2.6
	15	2.6	1.6	2.2	2.6	1.3	2.2	2.3	2.1

Table 7.2: *Improvement in* $\mathrm{IWSNR}_{\mathrm{seg}}$ *(dB).*

To investigate the influence on the enhancement performance of taking dependencies into account we can consider the real and imaginary parts of speech DFT coefficients as a vector process, that is $\mathbf{X} = [X_{\Re}, X_{\Im}]^T$ and use the MNIG based MAP estimator with $d = 2$ and $\mathbf{\Gamma} = \mathbf{I}$, i.e. $X_{\Re}$ and $X_{\Im}$ are assumed uncorrelated, but not independent. Experiments were performed with speech signals degraded by white noise at SNRs of 5, 10 and 15 dB. It turned out that there was no significant performance difference in terms of segmental SNR between an MNIG based estimator that assumes independent real and imaginary parts ($d = 1$) and an MNIG based estimator that assumes dependent real and imaginary parts ($d = 2$).

The reason for no significant difference between these two estimators may be that the impact of the independence assumption is relatively small compared to other imperfections in the speech enhancement system, e.g. estimation of the noise variance, estimation of the distribution parameters α and δ and the fact that the MNIG density is not identical to the true, but unknown, density of speech DFT coefficients. Moreover, it could be that the difference between the two estimators is so small that it is

input		$\mathrm{SATT}_{\mathrm{seg}}$ (dB) DFT est.			$\mathrm{SATT}_{\mathrm{seg}}$ (dB) Amplitude est.				
noise source	SNR (dB)	NIG	Wiener	Laplace	RIG	Rayleigh	Supergauss^1	Supergauss^2	$\hat{A}_{C,5}^{(1)}$
white	5	10.1	8.2	9.3	10.3	9.0	9.1	9.7	10.6
	10	12.8	10.6	11.8	13.1	11.5	11.7	12.4	13.2
	15	15.8	13.2	14.6	16.1	14.1	14.5	15.2	16.1
F16	5	10.2	8.0	9.2	10.4	9.0	9.1	9.8	10.7
	10	13.5	10.8	12.2	13.7	11.9	12.1	12.9	13.8
	15	17.0	13.8	15.4	17.3	15.0	15.3	16.2	17.0
Car	5	23.8	20.7	21.9	24.2	22.2	22.0	22.8	23.3
	10	26.6	23.1	24.4	27.1	24.6	24.5	25.3	25.7
	15	28.7	25.2	26.5	29.4	26.5	26.6	27.4	27.7
Factory	5	10.5	8.2	9.5	10.8	9.4	9.3	10.1	11.0
	10	13.9	11.0	12.5	14.2	12.2	12.4	13.3	14.1
	15	17.6	14.1	15.7	18.0	15.3	15.7	16.6	17.2

Table 7.3: $\mathrm{SATT}_{\mathrm{seg}}$ *(dB)*

input		$\mathrm{NATT}_{\mathrm{seg}}$ (dB) DFT est.			$\mathrm{NATT}_{\mathrm{seg}}$ (dB) Amplitude est.				
noise source	SNR (dB)	NIG	Wiener	Laplace	RIG	Rayleigh	Supergauss^1	Supergauss^2	$\hat{A}_{C,5}^{(1)}$
white	5	14.7	16.1	16.3	13.5	12.9	15.1	14.3	12.5
	10	11.9	13.3	13.5	10.8	10.4	12.2	11.6	10.4
	15	9.7	10.8	11.3	8.6	8.2	9.8	9.4	8.6
F16	5	12.2	13.8	14.0	11.2	10.7	12.8	12.0	10.5
	10	9.8	11.2	11.6	8.7	8.5	10.2	9.7	8.7
	15	8.0	9.1	9.7	6.9	6.8	8.3	7.9	7.3
Car	5	12.4	13.6	13.7	11.5	10.8	12.9	12.3	10.7
	10	11.0	12.1	12.4	10.0	9.5	11.4	10.9	9.6
	15	9.9	10.8	11.3	8.6	8.2	10.2	9.8	8.5
Factory	5	9.6	11.5	11.2	8.7	8.7	10.3	9.5	8.5
	10	8.0	9.6	9.7	7.1	7.1	8.5	8.0	7.3
	15	6.8	8.0	8.4	5.8	5.8	7.1	6.7	6.3

Table 7.4: $\mathrm{NATT}_{\mathrm{seg}}$ *(dB)*

not measurable due to the use of a lower limit on the enhancement gain function of 0.0316 and the fact that the definition of segmental SNR uses a lower limit of -10 dB. Therefore, experiments were also performed without the use of a lower limit on the gain function and without a lower limit on the segmental SNR. These experiments, shown in Table 7.5, show a small improvement in terms of segmental SNR in the order of 0.1 dB when using the MNIG based MAP estimator with $d = 2$. Despite the fact that the performance improvement when incorporating dependency between real and imaginary parts of DFT coefficients is relatively small, the MNIG based estimator with $d = 2$ has the advantage over the MNIG density with $d = 1$ that closed form solutions of the Bessel functions in Eq. (7.26) do exist, leading to lower computational complexity.

input SNR (dB)	MNIG $d = 2$	MNIG $d = 1$
5	12.2	12.0
10	9.7	9.6
15	7.5	7.4

Table 7.5: *Comparison between MNIG with* $d = 2$ *and MNIG with* $d = 1$ *in terms of improvement in* $\mathrm{SNR}_{\mathrm{seg}}$ *(dB).*

7.6.3 Subjective Evaluation

Informal listening confirmed the aforementioned objective results. With respect to the complex DFT estimators, the proposed 1d-MNIG based estimator leads to much less suppressed speech than the Wiener filter. Also, compared to the Laplace based estimator the speech quality was judged slightly better. In terms of residual noise, the Laplace based estimator leads to slightly less residual noise, but slightly more musical tones than the 1d-MNIG based estimator. When comparing the Wiener filter and the 1d-MNIG based estimator in terms of residual noise, it gives the perceptual impression that the Wiener filter leads to more residual noise, but with a less musical character. However, when listening to the noise filtered signal $\mathbf{G}(i)\mathbf{n}_t(i)$, i.e. the noise only signal filtered with the speech enhancement gain functions, then it turns out that the 1d-MNIG based estimator leads to more residual noise than the Wiener filter, but also more concentrated at places where speech is present. That might lead to more masking of the residual noise by speech energy. No perceptual difference was found between the 1d-MNIG and the 2d-MNIG based estimator.

With respect to the amplitude estimators, the proposed RIG based amplitude estimator leads to much better speech quality than the amplitude estimator derived under the Rayleigh density. The speech sounds less suppressed, but also less reverberant. Moreover, the RIG based amplitude estimator leads to a reduced amount of residual noise. The speech quality of the RIG based estimator is slightly better than that of the Supergauss[1] estimator. The difference in terms of speech quality between the Supergauss[2] and the RIG based estimator is rather small. The RIG based estimator leads to slightly more residual noise, but with a less musical character than the Supergauss[2] estimator. In comparison to the MMSE estimator $\hat{A}_{C,5}^{(1)}$ the RIG based estimator has more or less the same speech quality, but much less residual noise.

7.7 Conclusions

In this chapter we presented a new class of complex DFT and amplitude estimators for DFT-domain based speech enhancement. The estimators are derived under a multivariate normal inverse Gaussian (MNIG) distribution for the DFT coefficients. The MNIG distribution is very flexible and can model a wide range of densities, from heavy-tailed to less heavy-tailed. Under the MNIG distribution we derived complex DFT and amplitude estimators. Measurements of speech histograms based on speech DFT coefficients and DFT amplitudes showed a slightly better fit for the MNIG and RIG distribution, respectively, than for the pre-selected super-Gaussian distributions.

Experimental results demonstrated improvement in comparison to complex DFT and amplitude estimators that are based on Gaussian and pre-selected super-Gaussian distributions. Further, the derived complex MNIG based estimator allows for vector processing, where dependency and correlation between vector elements can be taken into account. In experiments the 2-dimensional MNIG based estimator was used to jointly estimate the real and imaginary parts of DFT coefficients. Experiments showed very small improvements when using the 2-dimensional MNIG instead of the 1-dimensional MNIG, the latter assuming independent real and imaginary parts of DFT coefficients.

References

[1] D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *J. of the Royal Stat. Soc.., Series B*, 36(1):99 –102, 1974.

[2] M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

[3] T. A. Øigård, A. Hanssen, R. E. Hansen, and F. Godtliebsen. EM-estimation and modeling of heavy-tailed processes with the multivariate normal inverse Gaussian distribution. *Signal Processing*, 85:1655–1673, 2005.

[4] T. Eltoft, T. Kim, and T. Lee. On the multivariate laplace distribution. *IEEE Signal Processing Letters*, 13(5):300–303, May 2006.

[5] I. Cohen. Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models. *Signal Processing*, 86(4):698–709, 2006.

[6] O. E. Barndorff-Nielsen. Normal inverse gaussian distributions and stochastic volatility modellling. *Scand. Jornal of Statistics*, 24:1–13, 1997.

[7] T. Eltoft. The Rician inverse Gaussian distribution: a new model for non- Rayleigh signal amplitude statistics. *IEEE Trans. Image Processing*, 14(11):1722–1735, November 2005.

[8] M. Tweedie. Functions of a statistical variate with given means with special reference to laplacian distributions. *Proceedings of the Cambridge Philosophical Society*, 43:41–49, 1947.

[9] I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series and Products*. New York: Academic, 6th ed. edition, 2000.

[10] R. Martin and C. Breithaupt. Speech enhancement in the DFT domain using Laplacian speech priors. In *Int. Workshop on Acoustic, Echo and Noise Control*, pages 87–90, September 2003.

[11] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[12] S. Kullback. *Information Therory and Statistics*. New York, Dover, 1997.

[13] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[14] R. J. McAulay and M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-28(2):137–145, April 1980.

[15] P. J. Wolfe and S. J. Godsill. Efficient alternatives to the Ephraim and Malah suppressionrule for audio signal enhancement. *EURASIP Journal on Applied Signal Processing*, 10:1043–1051, 2003.

[16] J. R. Deller, J. H. L. Hansen, and J. G. Proakis. *Discrete-Time Processing of Speech Signals*. IEEE Press, Piscataway, NJ, 2000.

[17] J. Greenberg, P. Peterson, and P.M. Zurek. Intelligibility-weighted measures of speech-to-interference ratio and speech system performance. *J. Acoust. Soc. Amer.*, 94:3009–3010, 1993.

[18] DARPA. Timit, Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc 1-1.1, October 1990.

[19] A. Varga and H. J. M. Steeneken. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–253, 1993.

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

Chapter 8

Noise Tracking using DFT-Domain Subspace Decompositions

This chapter is based on the article accepted for publication as "Noise Tracking using DFT Domain Subspace Decompositions", by R. C. Hendriks, J. Jensen and R. Heusdens in the *IEEE Trans. Speech, Audio and Language Processing*, March 2008.

8.1 Introduction

In the Chapters 6 and 7, MMSE and MAP estimators under super-gaussian densities have been discussed. Common for these estimators is that they are dependent on knowledge of the noise power spectral density (PSD). This does not only hold for the estimators derived in Chapters 6 and 7, but holds in general for most, if not all, speech enhancement estimators, e.g. [1][2][3][4]. Since in general the noise PSD is unknown, it has to be estimated from the noisy speech signal. An overestimation of the noise PSD will lead to oversuppression and, as a consequence, to a potential loss of speech quality, while an underestimation will lead to an unnecessary high level of residual noise. An accurate tracking of the noise PSD is therefore essential to obtain proper quality of the enhanced speech signal. Furthermore, fast tracking is important for non-stationary noise. However, both fast and accurate noise tracking is very challenging, especially under these non-stationary noise conditions.

A conventional method for estimating the spectral noise variance is to exploit speech pauses. Here, a voice activity detector [5][6] (VAD) is used and only in case of speech absence the noise PSD is estimated and updated. Although this is effective when the noise is stationary, it often fails when the noise statistics change during speech presence. Moreover, accurate voice activity detection under very low signal-to-noise-ratio (SNR) conditions is not trivial.

Minimum statistics (MS) based noise trackers [7][8] offer a more advanced alternative to VAD based methods. These methods exploit the property that the minimum power level in a particular frequency bin seen across a sufficiently long time interval is due to the noise process. From this minimum the average noise power can be estimated by applying a bias compensation. The size of the time interval should be such that there is at least one noise-only observation within the window. The minimum size of the time window is therefore dependent on the duration of speech presence in a frequency bin. If the time window is chosen too short and speech energy is constantly present in the search window, MS will track the PSD of the noisy speech instead of the noise PSD. This will lead to an overestimate of the noise level. If, on the other hand, the time window is chosen too long, changes in the noise power level are not tracked or can only be tracked with a large delay.

In this chapter we present a novel approach for noise tracking, which updates the noise PSD for each DFT coefficient even when both speech and noise are present. This method is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. We exploit the fact that these correlation matrices can be decomposed using an eigenvalue decomposition into two sub-matrices of which the columns span two mutually orthogonal vector spaces, namely a signal (+ noise) subspace and a noise-only subspace. We use the property that speech signals can often be expressed as a linear combination of a small number of complex exponentials [9]. Therefore, speech signals seen in a particular frequency bin can be described by a low-rank model. In that case, the eigenvalues that describe the energy in the noise-only subspace allow for an update of the noise statistics, even when speech is constantly present. Noise types that are described by a low-rank model itself, i.e. deterministic types of noise, will be represented in the signal subspace as well and need different measures to be estimated. How to track these deterministic

types of noise will be discussed as well.

The remainder of this chapter is organized as follows. In Section 8.2 we illustrate the potential of the proposed method of noise tracking. In Section 8.3 we explain the signal model and the concept of DFT-domain subspace decompositions that we use to derive the noise tracking method. In Section 8.4, the procedure for noise variance estimation is presented. Furthermore, in Section 8.5 we focus on some implementational aspects of the proposed noise tracking algorithm. In Section 8.6 we present experimental results and discuss tracking of deterministic types of noise. Finally, in Section 8.7 concluding remarks are given.

8.2 Illustration of DFT-Domain Subspace Based Noise Tracking

To illustrate the potential of the proposed method of noise tracking, we compare our new method to the MS method, which is known as the state-of-the-art for noise tracking in single-microphone speech enhancement applications. To do so, we create a synthetic signal in which the speech signal is modelled by a sinusoid of approximately 190 Hz. With this simplistic, but relevant model of a speech signal we can simulate the situation where speech energy is constantly present and demonstrate that our proposed method has great potential for tracking of the noise PSD in the presence of speech. In the first 2 seconds, (125 time frames) the signal consists of white noise only. Then, after 2 seconds a sinusoidal component is turned on and remains constantly present with a global SNR of 5 dB. This sinusoid simulates the continuous presence of speech energy. Finally, 0.5 seconds later, at frame number $i = 157$, the noise PSD decreases by 6 dB while the sinusoid remains present. We use both the MS approach and the proposed method to estimate the noise PSD. In Fig. 8.1 we compare their estimated noise PSDs together with the true noise PSD obtained by recursively smoothed periodogram estimates. The dotted line denotes the true noise PSD, the dashed-dotted line the noise PSD estimated using minimum statistics and the dashed line the noise PSD estimated with the proposed approach, all in the same frequency bin. We see that in the first approximately 156 frames both methods lead to a fairly good estimate of the true noise PSD. After 156 frames, however the proposed method follows the decrease in the noise PSD even though the sinusoid is present, while the MS method, on the other hand, is not able to follow this change. Moreover, approximately 100 frames after the sinusoid is turned on, the MS approach takes the energy of the noisy sinusoid as the new minimum and wrongly updates the estimated noise PSD. To avoid this problem, the search window could be enlarged. However, enlarging the search window will result in a larger delay and is harmful for tracking changes in the noise power.

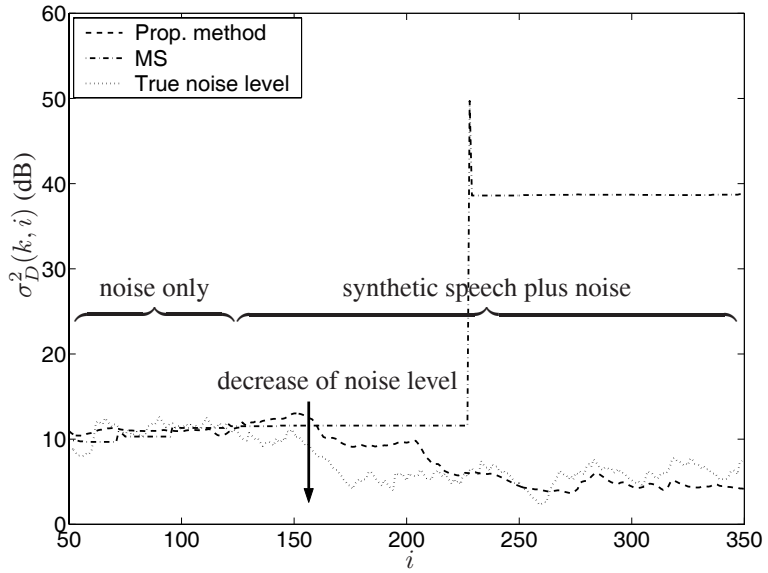

Figure 8.1: *Synthetic example of noise tracking.*

8.3 Signal Model and DFT-Domain Subspace Decompositions

In this chapter we consider the discrete Fourier transform of speech signals as being the outcome of a random process. That is, $Y(k,i)$, $X(k,i)$ an $D(k,i)$ are complex random variables denoting the noisy speech, clean speech and noise DFT coefficients of frame i and frequency bin k, with $k \in \{1,...,K\}$, and K the total number of frequency bins. We assume the noise to be additive, i.e. $Y(k,i) = X(k,i) + D(k,i)$, zero mean and uncorrelated with the clean speech signal, i.e., $E[X(k,i)D(k,i)] = 0$, $\forall\,(k,i)$.

We collect DFT coefficients per frequency bin k that originate from the time frames $i-p_1$ up to frame $i+p_2$ and form a vector $\mathbf{Y}(k,i) \in \mathbb{C}^M$ with $M = p_1+p_2+1$. That is,

$$\mathbf{Y}(k,i) = [Y(k,i-p_1),...,Y(k,i+p_2)]^T. \tag{8.1}$$

Let $\mathbf{C}_Y(k,i) \in \mathbb{C}^{M\times M}$ be the noisy speech correlation matrix related to frequency bin k and time frame i defined as

$$\mathbf{C}_Y(k,i) = E\left[\mathbf{Y}(k,i)\mathbf{Y}^H(k,i)\right], \tag{8.2}$$

where H indicates Hermitian transposition. The construction of $\mathbf{C}_Y(k,i)$ is illustrated in Fig. 8.2. Similarly we can define the speech correlation matrix $\mathbf{C}_X(k,i) \in \mathbb{C}^{M\times M}$, that is

$$\mathbf{C}_X(k,i) = E\left[\mathbf{X}(k,i)\mathbf{X}^H(k,i)\right],$$

and the noise correlation matrix $\mathbf{C}_D(k,i) \in \mathbb{C}^{M\times M}$, that is

$$\mathbf{C}_D(k,i) = E\left[\mathbf{D}(k,i)\mathbf{D}^H(k,i)\right].$$

Using the assumption that speech and noise are uncorrelated we can write the noisy speech correlation matrix $\mathbf{C}_Y(k,i)$ as

$$\mathbf{C}_Y(k,i) = \mathbf{C}_X(k,i) + \mathbf{C}_D(k,i).$$

Let us assume that $\mathbf{C}_D(k,i) = \sigma_D^2(k,i)I_M$, that is, the noise DFT coefficients in $\mathbf{D}(k,i)$ are uncorrelated. This assumption is valid when frames do not overlap and the correlation time of the noise is small enough [10]. In case of overlapping frames this assumption will be violated. This violation can be overcome by applying a pre-whitening transform, as we describe in Section 8.5.

As mentioned before, we make use of the property that speech signals can often be modelled by a sum of complex exponentials. In particular this is true for voiced speech sounds [9]. Under this signal model and under assumption that the frame size is long enough, ideally each frequency bin will observe at most one complex exponential across time. The clean speech correlation matrix $\mathbf{C}_X(k,i)$ can therefore be assumed to be of low-rank. When the noise-only subspace is of full-rank and the speech signal can be described using such a low-rank signal subspace, the eigenvalues that describe the energy in the noise-only subspace allow for an update of the noise PSD, even when speech is constantly present. A validation of the low-rank assumption of $\mathbf{C}_X(k,i)$ is given in Section 8.4.3.

Let $\mathbf{C}_X(k,i) = \mathbf{U}\mathbf{\Lambda}_X\mathbf{U}^H$ denote the eigenvalue decomposition of the clean speech correlation matrix related to frequency bin k and time frame i. Here, $\mathbf{U} \in \mathbb{C}^{M\times M}$ is a unitary matrix and contains the eigenvectors as columns and

$$\mathbf{\Lambda}_X = \mathrm{diag}\left(\lambda_{X_1}, ..., \lambda_{X_Q}, 0, ..., 0\right)$$

is a diagonal matrix with the non-negative eigenvalues

$$\lambda_{X_1} \geq \lambda_{X_2} \geq ... \geq \lambda_{X_Q} \geq 0$$

on the main diagonal and where $Q \leq M$ is the dimension of the signal subspace. Using the assumption that $\mathbf{C}_D(k,i)$ is a scaled diagonal matrix and $X(k,i)$ and $D(k,i)$ are uncorrelated we can write the eigenvalue decomposition of $\mathbf{C}_Y(k,i)$ as

$$\mathbf{C}_Y(k,i) = \mathbf{U}\left(\mathbf{\Lambda}_X(k,i) + \sigma_D^2(k,i)I_M\right)\mathbf{U}^H, \tag{8.3}$$

i.e. $\mathbf{C}_Y(k,i)$, $\mathbf{C}_X(k,i)$ and $\mathbf{C}_D(k,i)$ have the same eigenvectors and the eigenvalues of $\mathbf{C}_Y(k,i)$ are simply obtained by adding the eigenvalues of $\mathbf{C}_X(k,i)$ and $\mathbf{C}_D(k,i)$.

The eigenvector matrix $\mathbf{U}$ can be partitioned as $\mathbf{U} = [\mathbf{U}_1; \mathbf{U}_2]$, where the columns of $\mathbf{U}_1 \in \mathbb{C}^{M\times Q}$ form a basis for the signal subspace and the columns of $\mathbf{U}_2 \in \mathbb{C}^{M\times M-Q}$ form a basis for the noise-only subspace. Assuming that there indeed exists a low-dimensional signal subspace, i.e. $Q < M$, the eigenvalues in the noise-only subspace can be used to determine the noise PSD $\sigma_D^2(k,i)$, as the noise-only subspace eigenvalue matrix equals $I_{(M-Q)}\sigma_D^2(k,i)$.

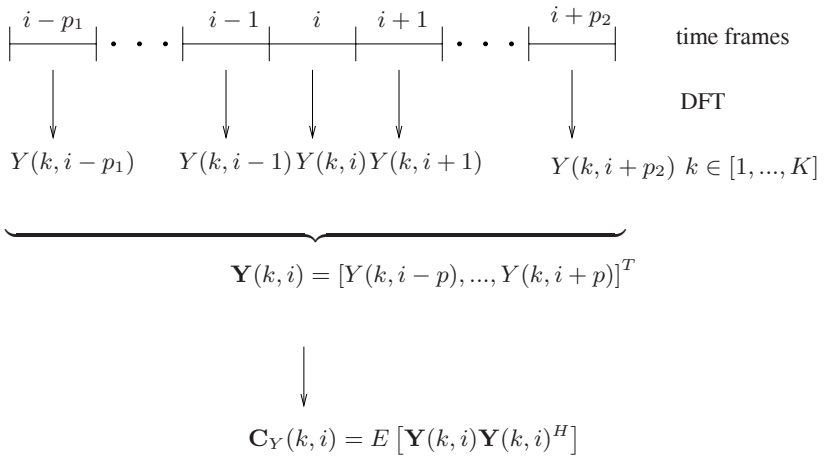

Figure 8.2: *Schematic overview of how correlation matrices in the DFT domain are computed.*

8.4 Estimation of $\sigma_D^2(k,i)$

In the previous section we considered the eigenvalue decomposition of $\mathbf{C}_Y(k,i)$ in order to estimate the noise PSD from the eigenvalues in the noise-only subspace. However, in practice the correlation matrix $\mathbf{C}_Y(k,i)$ in Eq. (8.2) is unknown and estimated based on realizations. Therefore, we consider in this section estimation of $\sigma_D^2(k,i)$ based on an estimate of the correlation matrix $\mathbf{C}_Y(k,i)$.

The correlation matrix $\mathbf{C}_Y(k,i)$ can be estimated from a limited number of samples by

$$\hat{\mathbf{C}}_Y(k,i) = \frac{1}{L}\mathcal{Y}(k,i)\mathcal{Y}^H(k,i), \tag{8.4}$$

where $\mathcal{Y} \in \mathbb{C}^{M\times L}$ is a Hankel-structured data-matrix defined as

$$\mathcal{Y}(k,i) = \begin{pmatrix} y(k,i-n_1) & \cdots & y(k,i-n_1+L-1) \\ \vdots & & \vdots \\ y(k,i-n_1+M-1) & \cdots & y(k,i+n_2) \end{pmatrix}, \tag{8.5}$$

where the small letters y indicate realizations of the random variable Y.

Let $\hat{\lambda}_{Y_l}$, indicate an eigenvalue of the estimated correlation matrix $\hat{\mathbf{C}}_Y(k,i)$. Given the eigenvalue decomposition of $\hat{\mathbf{C}}_Y(k,i)$ and the dimension of the signal subspace Q, it is shown in Appendix C.3, that under the assumption that the vector $\mathbf{Y}(k,i)$ has a multivariate Gaussian density, a maximum likelihood estimate of the noise PSD is given by

$$\hat{\sigma}_D^2(k,i) = \frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}. \tag{8.6}$$

That is, the noise PSD is estimated by taking the average of the eigenvalues in the noise-only subspace.

In order to compute Eq. (8.6) it is necessary to estimate the signal subspace dimension Q. Estimation of Q for noisy signals is a well-known problem for large data-records, and can be performed using e.g. Akaike information criterium (AIC) [11][12], minimum description length (MDL) criterium [12][13] or the Bayesian information criterium (BIC) [14]. However, when $\mathbf{C}_Y(k,i)$ is estimated based on a few data samples only, which is the case in our situation[1], existing model order estimators lead to inaccurate estimates of Q. Moreover, due to the inaccurate model order estimation and not always clear distinction between the noise-only and signal subspace, the noise power spectral estimate may be biased depending on whether the dimension of the signal subspace is over or underestimated. To increase the accuracy of the estimated model order, we present an alternative approach for model order estimation in Section 8.4.1, where we assume that some *a priori* knowledge of the noise level in each frequency bin is available. In order to correct for a possible bias we introduce a bias compensation factor for the estimation of $\hat{\sigma}_D^2(k,i)$ in Section 8.4.2.

8.4.1 Model Order Estimation

We consider an alternative approach for estimation of the signal subspace dimension, where we exploit the fact that some *a priori* information of the noise PSD is present. In this work we use the noise PSD estimate of the previous frame. This implicitly assumes relatively slowly varying noise. However, this does not limit the practical performance as will be shown in simulation experiments in Section 8.6. There it is shown that a change in the noise level of 15 dB per second can successfully be tracked. Furthermore, we assume that the eigenvalues in the noise-only subspace have an exponential distribution. Although we can not give any strict theoretical reasons that the distribution is truly exponential, the choice for an exponential distribution for the noise eigenvalues shows a reasonable fit in validation experiments [15].

A noisy eigenvalue $\hat{\lambda}_{Y_l}$ is decided to belong to the signal subspace when the probability of observing an eigenvalue equal or larger than $\hat{\lambda}_{Y_l}$ is smaller than a pre-chosen minimum probability P_{min}. We can write this as

$$\int_{\hat{\lambda}_{Y_l}}^{+\infty} f_{\Lambda_D}(\lambda_D) d\lambda_D < P_{min}, \tag{8.7}$$

where $f_{\Lambda_D}(\lambda_D)$ denotes the assumed pdf of the noise eigenvalues with its mean equal to the *a priori* known noise PSD, which we will take to be the noise PSD estimate of the previous frame. The decision procedure is visualized in Fig. 8.3. The dotted curve in Fig. 8.3 denotes the exponential pdf f_{Λ_D} of the noise eigenvalues belonging to the noise-only subspace. This approach can be seen within a hypothesis based framework

[1] In our experiments $\mathbf{C}_Y(k,i) \in \mathbb{C}^{M\times M}$ has dimension $M = 7$ and is estimated from 13 data-samples, i.e. 13 DFT coefficients at a fixed frequency bin k and from consecutive (overlapping) frames. Using much more data-samples might lead to a higher model order Q because the speech signal might becomes less stationary and as a consequence more complex exponentials are needed to model the speech signal.

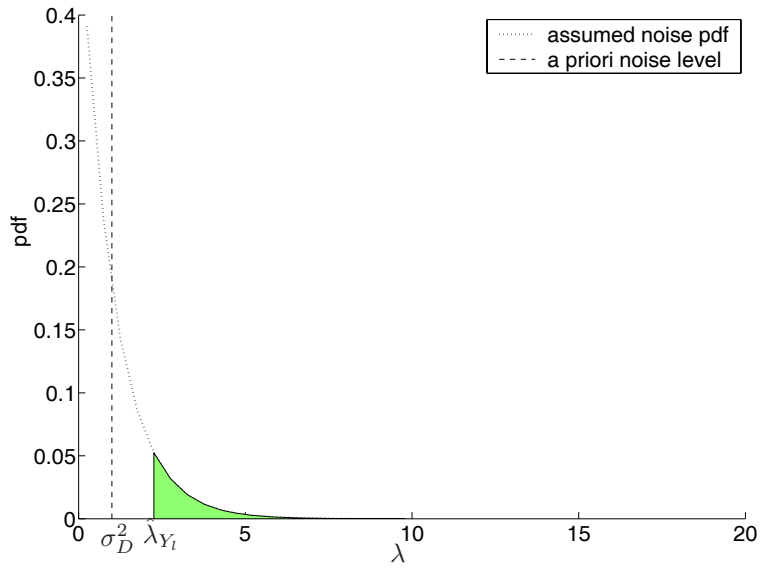

Figure 8.3: *Example showing how the noise-only subspace dimension is determined.*

where H_0 and H_1 are defined as

$$\begin{aligned} H_0: &\quad \hat{\lambda}_{Y_l} \text{ belongs to the noise-only subspace} \\ H_1: &\quad \hat{\lambda}_{Y_l} \text{ belongs to the signal subspace.} \end{aligned} \tag{8.8}$$

Given a threshold λ_{th}, H_1 is decided when $\hat{\lambda}_{Y_l} > \lambda_{th}$. When $\hat{\lambda}_{Y_l} \leq \lambda_{th}$, $\hat{\lambda}_{Y_l}$ is decided to belong to the noise-only subspace. The hypothesis is evaluated for all eigenvalues in increasing order until H_0 is rejected, which determines then the dimension of the noise and the signal subspace. The threshold λ_{th} can be expressed in terms of the false alarm probability $P_{fa} = P_{min}$ and is given by $\lambda_{th} = -\sigma_D^2 \ln P_{fa}$ [16].

For evaluation, the proposed model order is compared to an MDL based model order estimator. Comparing to the existing MDL based model order estimator [12] is not completely fair and will be in advantage of the proposed method, because it uses *a priori* knowledge on the noise variance while the traditional MDL estimator in [12] does not. Therefore, to have a fair comparison we derive in Appendix C.2 a modified MDL model order estimator where *a priori* knowledge on the noise variance is also taken into account.

For the comparison a synthetic signal was constructed, consisting of a sinusoid at frequency bin number 11 in additive white noise. The sinusoid will not only have a contribution to bin $k = 11$, but to neighboring bins as well, because the period of the sinusoid is not an integer multiple of the minimum period visible with the used DFT size. The overall SNR between the sinusoid and white noise was 0 dB. For each frequency bin we estimate a correlation matrix $\mathbf{C}_Y(k,i) \in \mathbb{C}^{7\times 7}$ and use either the proposed approach or the modified MDL method to estimate the dimension of the

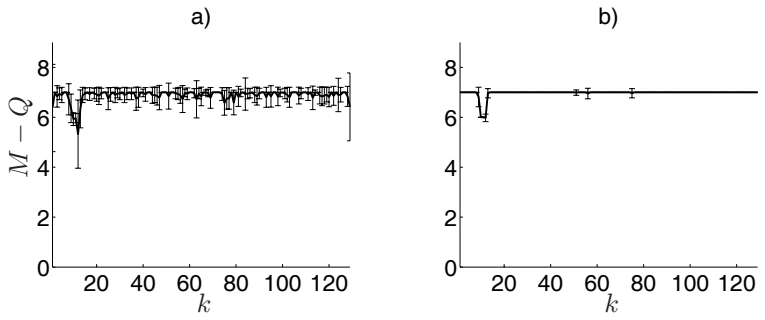

Figure 8.4: *a) MDL model order estimator with a priori knowledge on noise variance b) proposed model order estimator.*

noise-only subspace. At those frequency bins where the sinusoid is present, a noise-only subspace dimension of 6 is expected, while at all other bins a noise-only subspace dimension of 7 is expected.

In Fig. 8.4 the outcome of the comparison between modified MDL derived in Appendix C.2 and the proposed method are shown with $\mathbf{C}_Y(k,i) \in \mathbb{C}^{M\times M}$ estimated based on a data-matrix $\mathcal{Y} \in \mathbb{C}^{7\times 7}$. For each successive frequency bin the model order is estimated. This is repeated for many frames. The average noise-only subspace dimension and the variance of noise-only subspace dimension are shown in Fig. 8.4. We see that the modified MDL approach leads to a larger variance in the estimated model order than the proposed approach. We use in the following the proposed approach to estimate the model order of the noise-only subspace because of its smaller variance.

8.4.2 Bias Compensation of $\hat{\sigma}_D^2(k,i)$

When the dimension Q of the signal subspace is overestimated or underestimated, evaluating Eq. (8.6) can result in the introduction of a bias in the noise PSD estimate. To correct for such a bias in the estimated noise PSD as a result of consistent over or underestimates of Q, we introduce a signal subspace dimension dependent bias compensation factor $B(Q)$ and compute $\hat{\sigma}_D^2(k,i)$ as

$$\hat{\sigma}_D^2(k,i) = \frac{1}{B(Q)} \frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i). \tag{8.9}$$

The argumentation that we use to define the bias compensation factor is similar to the one introduced in [17].

The use of this bias compensation factor $B(Q)$ is based on the fact that

$$E\left[\frac{1}{(M-Q)} \sum_{l=Q+1}^{M} \hat{\lambda}_{Y_l}(k,i)\right]$$

is proportional to σ_D^2. We therefore write

$$\hat{\sigma}_D^2(k,i) = \frac{\sigma_D^2(k,i)}{E\left[\frac{1}{(M-Q)}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}(k,i)\right]}\frac{1}{(M-Q)}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}(k,i),$$

with

$$B(Q) = \frac{E\left[\frac{1}{(M-Q)}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}(k,i)\right]}{\sigma_D^2(k,i)}. \tag{8.10}$$

In order to compute the bias compensation factor $B(Q)$, for $Q = 0, 1, ..., M$, we approximate Eq. (8.10) by making use of a training procedure based on speech data degraded by white noise with a known variance $\sigma_D^2(k,i) = 1 \; \forall \; (k,i)$. Let $\tilde{B}(k,i)$ be defined as

$$\tilde{B}(k,i) \quad = \quad \frac{\frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_{Y_l}(k,i)}{\sigma_D^2}. \tag{8.11}$$

and let $\mathcal{Q}(Q)$ be the set of time-frequency points in the training data for which the signal subspace dimension is estimated to be Q. $B(Q)$, $Q = 0, 1, ..., M$, is then computed by averaging $\tilde{B}(k,i)$ over the set $\mathcal{Q}(Q)$ leading to

$$B(Q) = \frac{1}{|\mathcal{Q}(Q)|}\sum_{(k,i)\in\mathcal{Q}(Q)}\left[\tilde{B}(k,i)\right],$$

where $|\mathcal{Q}(Q)|$ is the cardinality of the set $\mathcal{Q}(Q)$. Notice that computing the bias compensation factor in the training phase using the same signal subspace dimension estimator as when used in practice has the advantage that it can help to overcome systematic errors due to the signal subspace dimension estimator. Further, notice that $B(Q)$ can show some dependency on the SNR of the training data. This can be taken into account by computing $B(Q)$ also as a function of SNR.

8.4.3 Dimension of $\mathbf{C}_Y(k,i)$

A requirement for the noise-only subspace to exist is that the signal subspace is not of full rank. For many speech sounds it holds that they can be modelled using a (limited) number of basis functions. Consider, for example, the voiced speech sounds that can be modelled using a sum of complex exponentials. In that case, a particular frequency bin containing a harmonic will only observe a small number of complex exponentials and results in a low dimensional signal subspace. The dimension of the correlation matrix can then be chosen such that the noise-only subspace has sufficiently high dimension to make an accurate estimate of the noise variance. To show that the dimension of the signal subspace is usually relatively low, we estimated for each DFT coefficient in the time-frequency plane the correlation matrix $\mathbf{C}_X \in \mathbb{C}^{M\times M}$ with $M = 7$. For each estimated correlation matrix we defined the model order as the number of eigenvalues needed to contain at least 95 % of the energy. In Fig. 8.5 we illustrate this experiment. The clean speech signal is shown in Fig. 8.5a. For each time-frequency point

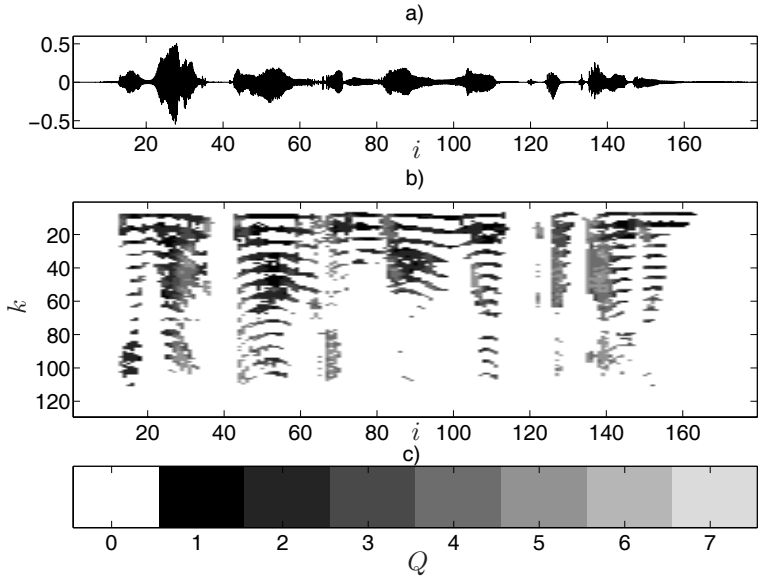

Figure 8.5: *a) Clean speech signal. b) Dimension of the signal subspace Q for each time-frequency point (k,i). Q is estimated by measuring in how many of the M eigenvalues 95 % of the energy is distributed. c) Color legend.*

the estimated model order Q is indicated in Fig. 8.5b using colors from the legend in Fig. 8.5c. The white color in the legend indicates speech absence, i.e. $Q = 0$. Time-frequency points are classified as speech absence when their energy is 40 dB below the DFT coefficient with maximum energy. We see that in general the dimension of the signal subspace Q is relatively low, especially at the harmonic tracks. Further we see that $M = 7$ is a sufficient dimension for the correlation matrix, since the model order of 5 is hardly exceeded.

8.5 Implementational Aspects

In this section we focus on some implementational aspects and present a summary of the proposed algorithm.

8.5.1 Pre-Whitening

In Section 8.3 the assumption was made that $\mathbf{C}_D(k,i) = \sigma_D^2(k,i)\mathbf{I}_M$. Although this assumption holds as long as the DFT coefficients in $\mathbf{D}(k,i)$ are computed from time frames that are not overlapping and/or when the correlation time of the noise is small enough [10], this assumption becomes less valid when an overlap is introduced. In this section we show how the inter-frame correlation is affected by the window overlap

and indicate how a pre-whitening matrix can be obtained such that the aforementioned assumption is fulfilled.

Let $D_t(m)$ denote a time-domain sample considered as a random variable, let $\overline{D_t(m)}$ indicate complex conjugation of $D_t(m)$ and let P denote the frame shift. Let $C_D(k,i;p)$ denote the correlation between a noise DFT coefficient $D(k,i+p)$ and $D(k,i)$ with frame lag p. The correlation $C_D(k,i;p)$ can then be written as

$$\begin{aligned}
C_D(k,i;p) &= E[D(k,i+p)\overline{D(k,i)}] \\
&= E\left[\left(\sum_{m=0}^{K-1} D_t(m+(i+p)P)e^{-\frac{j2\pi km}{K}}\right)\overline{\sum_{n=0}^{K-1} D_t(n+iP)e^{-\frac{j2\pi kn}{K}}}\right] \\
&= e^{\frac{j2\pi kpP}{K}} E\left[\sum_{m=pP}^{K-1+pP} D_t(m+iP)e^{-\frac{j2\pi km}{K}}\sum_{n=0}^{K-1}\overline{D_t(n+iP)}e^{\frac{j2\pi kn}{K}}\right] \\
&= \underbrace{e^{j2\pi kpP/K}\sum_{m=pP}^{K-1} E\left[|D_t(m+iP)|^2\right]}_{\tilde{C}_D(k,i;p)} \\
&+ \underbrace{\sum_{m=pP}^{K-1+pP}\sum_{\substack{n=0\\ m\neq n+pP}}^{K-1} e^{j2\pi k(pP+n-m)/K} E\left[D_t(m+iP)\overline{D_t(n+iP)}\right]}_{C_C(k,i;p)}.
\end{aligned} \tag{8.12}$$

We conclude that the correlation $C_D(k,i;p)$ consists of two components; a term $\tilde{C}_D(k,i;p)$ and a term $C_C(k,i;p)$. $C_C(k,i;p)$ contains all the cross-terms and is dependent on the cross-correlation between the time samples. In general it holds that $C_C(k,i;p)$ decreases for increasing P. Also, the shorter the correlation time in the noise, the smaller $C_C(k,i;p)$ becomes. For $C_D(k,i;p)$ with $p>0$ it follows from Eq. (8.12) that even if the time-domain process $D_t(\cdot)$ is completely uncorrelated $C_D(k,i;p)\neq 0$, unless $P > K-1$, which means no overlap between consecutive frames.

Using simulations with white noise training data we can estimate the first term $\tilde{C}_D(k,i;p)$ for a given overlap and also take some windowing effects into account. The second term $C_C(k,i;p)$ is signal dependent and is therefore in general unknown.

We can write $\tilde{C}_D(k,i;p)$ in Toeplitz matrix form, that is

$$\tilde{\mathbf{C}}_D(k,i) = \begin{pmatrix} \tilde{C}_D(k,i;0) & \tilde{C}_D(k,i;1) & \cdots & \tilde{C}_D(k,i;p) \\ \tilde{C}_D(k,i;-1) & \tilde{C}_D(k,i;0) & & \\ \vdots & & \ddots & \\ \tilde{C}_D(k,i;-p) & \cdots & & \tilde{C}_D(k,i;0) \end{pmatrix}.$$

Let the relative error between the two correlation matrices $\mathbf{C}_D(k,i)$ and $\tilde{\mathbf{C}}_D(k,i)$

be defined as,

$$\mathrm{Err}_{\mathrm{rel}}(\mathbf{C}_D(k,i),\tilde{\mathbf{C}}_D(k,i)) = \frac{\|\mathbf{C}_D(k,i)-\tilde{\mathbf{C}}_D(k,i)\|_{\mathrm{F}}^2}{\|\mathbf{C}_D(k,i)\|_{\mathrm{F}}^2},$$

with $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm [18]. In a simulation environment we can then compute the error that would have been made between $\mathbf{C}_D(k,i)$ and $\tilde{\mathbf{C}}_D(k,i)$ by neglecting the second correlation term $C_C(k,i;p)$.

To investigate the influence of neglecting the second correlation term $C_C(k,i;p)$ we conducted an experiment where $K = 256$ and $P = 32$, i.e. the overlap between time frames was 87.5 %. Then we computed for 3 different non-white noise sources, i.e. babble noise, factory noise 1 and factory noise 2, the true correlation matrix $\mathbf{C}_D(k,i)$ and computed $\tilde{\mathbf{C}}_D(k,i)$ based on white noise. The relative error $\mathrm{Err}_{\mathrm{rel}}(\mathbf{C}_{\mathrm{D}}(\mathrm{k,i}),\tilde{\mathbf{C}}_{\mathrm{D}}(\mathrm{k,i}))$ that is made by replacing $\mathbf{C}_D(k,i)$ by $\tilde{\mathbf{C}}_D(k,i)$ based on white noise and averaged over all frequency bins is shown in Table 8.1. We see that the relative error is always lower than $12 \cdot 10^{-3}$. This indicates that neglecting the cross-terms leads to a relatively small error for these type of noise sources and that $\mathbf{C}_D(k,i)$ is mainly determined by $\tilde{\mathbf{C}}_D(k,i)$. In the experimental results presented in Section 8.6 we will therefore neglect the correlation term $C_C(k,i;p)$ and use a correlation matrix $\tilde{\mathbf{C}}_D(k,i)$ trained on white noise to whiten possibly colored noise in $\mathbf{Y}(k,i)$.

Let $\mathbf{C}^{\frac{1}{2}}$ denote the principle square root of a matrix $\mathbf{C}$ [19]. The whitening of a vector $\mathbf{Y}(k,i)$ can then be written as

$$\mathbf{Y}_{pre}(k,i) = \tilde{\mathbf{C}}_D^{-\frac{1}{2}}(k,i)\mathbf{Y}(k,i). \tag{8.13}$$

$\mathbf{Y}_{pre}(k,i)$ is then used in Eq. (8.2). We denote the noise PSD when estimated in the whitened domain by $\hat{\sigma}^2_{D,\mathrm{pre}}(k,i)$. Notice, that if $\sigma^2_D(k,i)$ is estimated in the whitened domain, we have to correct with a scaling factor $\frac{\mathrm{tr}[\tilde{\mathbf{C}}_D(k)]}{M}$, with $\mathrm{tr}[\,\cdot\,]$ the trace operator [19], to obtain the noise PSD estimate in the non-whitened domain.

For some highly correlated noise types, i.e. with long correlation time, the aforementioned assumption of neglecting the correlation term $C_C(k,i;p)$ might be less valid. In that case Eq. (8.13) is not sufficient to whiten the noise process. A possible solution is to estimate the whitening transform matrix $\mathbf{C}_D(k,i)$ online during speech absence using a VAD. A somewhat more advanced method would be to exploit the signal subspace dimension estimator and update the estimated correlation matrix when the estimated noise-only subspace is full rank, i.e. $Q = 0$. Moreover, using a smaller overlap would also make the assumption of neglecting the correlation term $C_C(k,i;p)$ more valid. However, the experimental results that are presented in Section 8.6 are obtained using Eq. (8.13) with an overlap of 87.5 % between successive frames.

8.5.2 Algorithm Summary

In order to apply the proposed algorithm, the following steps should be taken:

1. Compute $\hat{\mathbf{C}}_Y(k,i)$ using Eq. (8.4) and (8.5). The DFT coefficients necessary to form data-matrix $\mathcal{Y}$ in Eq. (8.5) are computed using an FFT of frames with a

noise type	Babble	Factory 1	Factory 2
$\mathrm{Err}_{\mathrm{rel}}(\mathbf{C}(\mathrm{k,i}), \tilde{\mathbf{C}}(\mathrm{k,i}))$	$12 \cdot 10^{-3}$	$5.5 \cdot 10^{-3}$	$8.2 \cdot 10^{-3}$

Table 8.1: *Relative error for three non-white noise sources.*

pre-defined overlap. The choice for this overlap is a tradeoff between variance reduction of $\hat{\mathbf{C}}_Y(k,i)$ and stationarity of the data in the data-matrix.

2. Apply pre-whitening using Eq. (8.13) to remove the correlation in the noise introduced in step 1.

3. Compute the eigenvalue decomposition of the correlation matrix $\hat{\mathbf{C}}_Y(k,i)$ in the pre-whitened domain.

4. Estimate the noise PSD $\sigma^2_{D,\mathrm{pre}}(k,i)$ using Eq. (8.9)

5. Correct for scaling due to the pre-whitening in step 2

$$\hat{\sigma}^2_D(k,i) = \frac{tr\left[\tilde{\mathbf{C}}_D(k,i)\right]}{M}\hat{\sigma}^2_{D,\mathrm{pre}}(k,i).$$

8.6 Experimental Results

For performance evaluation we compare the proposed method with the minimum statistics based noise tracking algorithm [8] and with the situation where the noise PSD is computed using an ideal VAD. The speech and noise signals originate from the Noizeus database [20]. This database was extended with stationary computer generated white Gaussian noise, babble noise from the Noisex-92 database [21], noise originating from a passing train and non-stationary white Gaussian noise, respectively. Noisy signals are constructed synthetically at input SNRs of 0, 5, 10 and 15 dB. For the non-stationary white Gaussian noise, the initial noise level is 0, 5, 10 and 15 dB, respectively, and then gradually increases in one second by 15 dB where it stays at that level for 2 seconds after which it decreases again by 15 dB in one second. All signals are filtered at telephone bandwidth and sampled at 8 kHz. The noisy time-domain signals are divided in frames of 256 samples with 50 % overlap. For both analysis and synthesis a square root Hann window is used. The DFT coefficients that are used to form the data-matrix $\mathcal{Y}$ originate from time frames taken with an overlap of 87.5 %. The dimensions of $\mathcal{Y}$ were chosen as $M = L = 7$ and $n_1 = n_2 = 6$. The estimated noise PSDs $\hat{\sigma}^2_D(k,i)$ are smoothed using an exponential smoother with adaptive smoothing factors [8].

8.6.1 Performance Evaluation

To illustrate the noise tracking performance of the proposed approach within a typical example of noisy speech, we concatenated four speech signals and degraded this by noise originating from a passing train at 5 dB global SNR. In Fig. 8.6 the estimated

noise PSDs are shown for the proposed approach and the MS approach together with the true noise variance for a single frequency bin $k = 20$. This bin index corresponds to a frequency band centered around 625 Hz. We see that the proposed approach follows the increase in the noise level much better than the minimum statistics approach. This is due to the fact that the proposed approach can track changes in the noise level during speech presence. The MS approach on the other hand is limited in its update rate due to its search window and the fact that it can not track the noise when speech is continuously present in a bin. This results for MS in the delayed tracking of a rising noise level in Fig. 8.6.

In Fig. 8.7 another example is shown where the same speech signal is degraded by the non-stationary white noise described above. The initial part of the speech signal is degraded at an SNR of 10 dB. We again see that the proposed approach tracks the increase in noise level much faster than the MS approach.

Objective Performance Evaluation

For objective performance evaluation we use the segmental relative estimation error defined in [22] as

$$\text{Err}_{\text{seg}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{k=1}^{K}\left[\hat{\sigma}_D^2(k,i) - \sigma_D^2(k,i)\right]^2}{\sum_{k=1}^{K}\sigma_D^4(k,i)},$$

where N is the total number of frames in the signal and where $\sigma_D^2(k,i)$ is the ideal noise PSD measured using noise periodograms smoothed over time using an exponential window, i.e.

$$\sigma_D^2(k,i) = \alpha\sigma_D^2(k,i-1) + (1-\alpha)|D(k,i)|^2,$$

with a smoothing factor $\alpha = 0.9$ [8]. The measure Err_{seg} is non-symmetric and is more sensitive to overestimates than to underestimates. Therefore, we propose a symmetric segmental logarithmic estimation error, defined as

$$\text{LOG-Err}_{\text{seg}} = \frac{1}{NK}\sum_{k=1}^{K}\sum_{i=1}^{N}\left|10\log\left[\frac{\sigma_D^2(k,i)}{\hat{\sigma}_D^2(k,i)}\right]\right|.$$

In order to evaluate the influence of the proposed noise tracking algorithm on speech enhancement performance we use the estimated noise PSDs within a DFT-domain based speech enhancement algorithm, similar as the one depicted in Fig. 1.2. As estimator we use the MMSE amplitude estimator under the generalized Gamma model as presented in Chapter 6 with $\gamma = 2$ and $\nu = 0.1$. The maximum suppression was limited to 0.1 for perceptual reasons. For *a priori* SNR estimation we use the decision-directed (DD) approach [1] where a smoothing factor $\alpha = 0.98$ was used as proposed in [1]. For performance comparison we use segmental SNR, i.e.,

$$\text{SNR}_{\text{seg}} = \frac{1}{N}\sum_{i=1}^{N}10\log_{10}\frac{\sum_k |x(k,i)|^2}{\sum_k |x(k,i)-\hat{x}(k,i)|^2}, \tag{8.14}$$

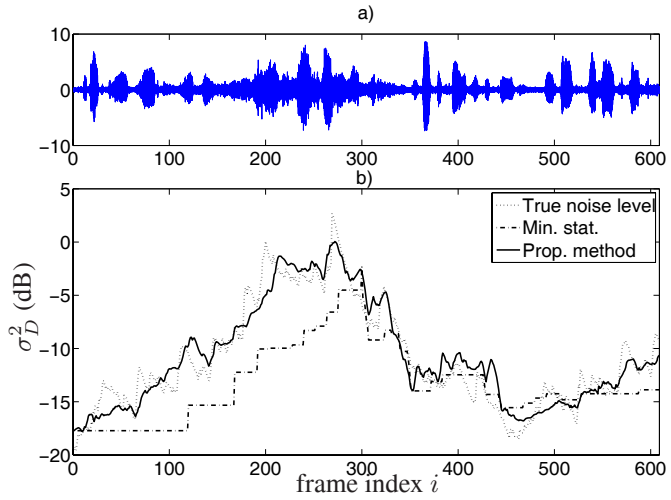

Figure 8.6: *Comparison between proposed method and minimum statistics. The estimated noise levels for bin $k = 20$ are shown. The noisy signal consists of speech degraded by non-stationary train noise at an overall input SNR of 5 dB.*

where $x(k,i)$ is a realization of a clean speech DFT and $\hat{x}(k,i)$ is its clean speech DFT estimate, respectively. The definition of segmental SNR in Eq. (8.14) differs slightly from the ones used in Chapters 6 and 7 where an upper and lower limit was applied on the measured SNR. We left out these limits, because our interest here is to express the performance difference between different noise trackers. The tracking performance of a sudden increase or decrease of the noise level is better reflected without these limits. Notice that the performance measured using $\mathrm{SNR}_{\mathrm{seg}}$ is unlike LOG-$\mathrm{Err}_{\mathrm{seg}}$ and $\mathrm{Err}_{\mathrm{seg}}$ not only influenced by the noise tracking algorithm, but also by the chosen gain function and *a priori* SNR estimator.

In Tables 8.2-8.4 we show performance evaluations for several noise types averaged over speech signals originating from the Noizeus database. We compare noise tracking using VAD, MS and the proposed approach. We see that in general for all three objective measures the performance is increased when using the proposed approach. Especially for noise sources that are characterized by a gradual change in the noise power (passing train and non-stationary white Gaussian noise) we see that the proposed approach outperforms MS and VAD. This is mainly due to the fact that a continuous update of the noise PSD allows for a faster update of changes in the noise power. Such an improved estimate of the noise PSD will increase the overall quality of a speech enhancement system as also reflected by the performance expressed in terms of $\mathrm{SNR}_{\mathrm{seg}}$ in Table 8.4.

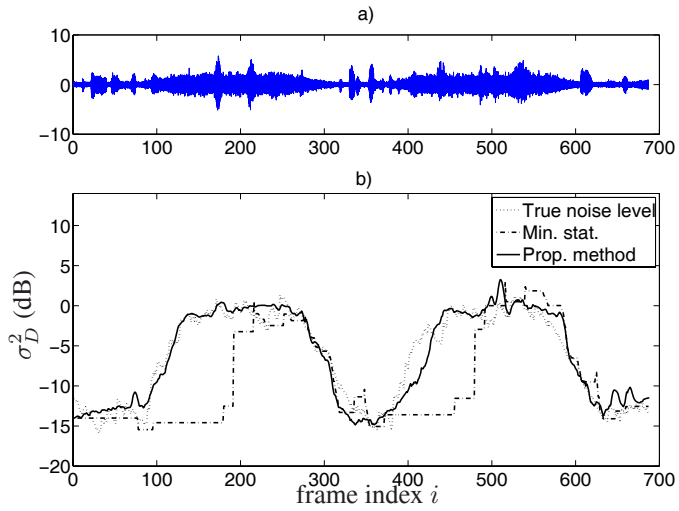

Figure 8.7: *Comparison between proposed method and minimum statistics. The estimated noise levels for bin $k = 20$ are shown. The noisy signal consists of speech degraded by non-stationary white noise.*

Subjective Performance Evaluation

For subjective evaluation an OAB listening test was performed with 8 participants, the authors not included. Here, O is the original clean speech signal and A and B are two noisy signals that are enhanced using the aforementioned DFT-domain based speech enhancement algorithm with two different methods for noise tracking. Method A uses the proposed noise tracking method, and method B uses the minimum statistics approach. The listeners were presented first the original signal followed by the two different enhanced signals A and B played in random order. The participants had to indicate their preference for excerpt A or B. Each series was repeated 4 times, with each time a randomized order of the signals A and B. In this listening test we used four different types of additive noise at two different SNRs, namely, white noise, street noise, noise originating from a passing train and non-stationary white noise at SNRs of 5 dB and 15 dB. For each noise type and noise power level we presented the listeners two female sentences and two male sentences. The average preference for method A under each test condition is shown in Table 8.5. Under all test conditions the proposed method for noise tracking was preferred over the minimum statistics approach.

8.6.2 Deterministic Noise

Deterministic noise components can in principle not be tracked with the proposed method, since they will appear in the signal subspace and not in the noise-only subspace. The noise is thus implicitly assumed to be stochastic. This is not only a property of the proposed method. Minimum statistics [8] implicitly assumes the noise to

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	1.6	0.31	0.17
	5	0.77	0.35	0.20
	10	0.75	0.35	0.22
	15	1.2	0.42	0.29
street	0	26.7	0.41	0.35
	5	18.3	0.43	0.31
	10	17.4	0.48	0.34
	15	18.6	0.94	0.53
white	0	0.19	0.13	0.074
	5	0.20	0.13	0.093
	10	0.18	0.15	0.15
	15	0.19	0.45	0.24
babble	0	0.47	0.48	0.34
	5	0.47	0.47	0.37
	10	0.47	0.47	0.40
	15	0.47	0.50	0.43
passing train	0	0.52	0.33	0.16
	5	0.52	0.38	0.18
	10	0.52	0.45	0.28
	15	0.52	1.2	0.37
non-stationary WGN	0	0.66	0.33	0.068
	5	0.66	0.35	0.075
	10	0.66	0.38	0.096
	15	0.66	0.42	0.13

Table 8.2: *Performance in terms of* Err_{seg}

be stochastic as well. More specifically, the bias-compensation that is applied within minimum statistics is based on the assumption that the noise is stochastic. However, it is applied to deterministic components as well. A consequence of this is that after bias-compensation the deterministic noise components are in general slightly overestimated. However, in practice minimum statistics is less sensitive than the proposed method when violating this assumption.

When deterministic noise components are present they are often mixed with stochastic noise components. Therefore it is not obvious how to estimate them. One way to estimate the deterministic noise components as well, is to make use of the fact that for stochastic noise the minimum of the last T minimum statistics based noise PSD estimates $\hat{\sigma}^2_{D,min}$ is always smaller than or equal to the current noise PSD estimate made by the proposed noise tracker (8.9), i.e.

$$\min\left[\hat{\sigma}^2_{D,min}(k, i-T+1), ..., \hat{\sigma}^2_{D,min}(k,i)\right] \leq \hat{\sigma}^2_D(k,i).$$

Whenever this minimum is larger than $\hat{\sigma}^2_D(k,i)$ this is due to the fact that deterministic noise components are present. In that case we can estimate the deterministic part of

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	3.3	2.6	1.9
	5	3.2	2.9	1.9
	10	3.0	2.6	2.0
	15	3.1	2.7	2.1
street	0	4.1	2.3	2.0
	5	4.0	2.8	2.0
	10	4.4	3.1	2.2
	15	3.6	2.7	2.5
white	0	1.6	1.3	1.0
	5	1.6	1.3	1.1
	10	1.6	1.4	1.2
	15	1.6	1.5	1.5
babble	0	4.4	3.9	2.1
	5	4.4	3.7	2.3
	10	4.4	3.5	2.6
	15	4.4	3.4	3.0
passing train	0	7.7	3.7	1.6
	5	7.7	3.6	1.9
	10	7.7	3.5	2.3
	15	7.7	3.6	3.0
non-stationary WGN	0	8.6	4.0	0.94
	5	8.6	4.1	1.0
	10	8.6	4.1	1.1
	15	8.6	4.0	1.4

Table 8.3: *Performance in terms of* LOG-Err$_{\text{seg}}$

$\sigma_D^2(k,i)$ by

$$\hat{\sigma}_{D,det}^2(k,i) = \left[\hat{\sigma}_{D,min}^2(k,i) - \hat{\sigma}_D^2(k,i)\right] B_{min}^{-1}(k,i), \tag{8.15}$$

where B_{min} is the bias-compensation as used in the minimum statistics method and which is used here to correct for the wrongly applied bias compensation on the deterministic component. The total estimate of $\sigma_D^2(k,i)$ is then given by adding $\hat{\sigma}_{D,det}^2(k,i)$ and the estimate obtained by the proposed method in (8.9).

In Fig. 8.8a a comparison is shown where a speech signal was degraded by white noise (filtered at telephone bandwidth) at an SNR of 5 dB. As deterministic noise a signal consisting of a sum of three harmonically related sinusoids with fundamental frequency of 656 Hz was added at an SNR of 10 dB with respect to the original clean speech signal. We see in Fig. 8.8a that with the DFT-domain subspace noise tracking approach it is not possible to estimate the sinusoidal noise components. In Fig. 8.8b we combine DFT-domain subspace noise tracking method with (8.15) and see that we also determine the deterministic noise components. In Table 8.6 we show a comparison in terms of LOG-Err$_{\text{seg}}$ for speech signals degraded by the above described noise.

noise source	input SNR (dB)	VAD	MS	prop. method
train	0	-3.3	-4.1	-2.3
	5	-0.29	-0.48	1.1
	10	3.9	3.6	4.7
	15	7.4	7.4	8.0
street	0	-4.1	-4.1	-3.5
	5	-1.1	-0.85	0.34
	10	2.9	3.2	4.1
	15	6.8	7.0	7.1
white	0	-0.65	-0.22	0.42
	5	2.9	3.1	3.7
	10	6.3	6.4	6.9
	15	9.8	9.8	10.0
babble	0	-8.4	-8.8	-6.8
	5	-4.1	-4.2	-2.8
	10	0.26	0.25	1.2
	15	4.6	4.7	5.1
passing train	0	-6.2	-4.3	-1.7
	5	-1.8	-0.037	1.8
	10	2.6	4.2	4.9
	15	7.0	8.3	8.4
non-stationary WGN	0	-19.2	-14.5	-9.4
	5	-14.6	-10.5	-5.8
	10	-10.0	-6.5	-2.5
	15	-5.5	-2.5	0.60

Table 8.4: *Performance in terms of* $\mathrm{SNR}_{\mathrm{seg}}$ (dB)

Here the SNR between the stochastic noise and the speech signal is 5 dB, and the SNR between the deterministic noise and the speech signal is 0, 5, 10 and 15 dB, respectively. Moreover, we also show a comparison for the natural noise source *Destroyer operations room background noise* that originates from the Noisex-92 database [21]. This is a noise source containing both stochastic and some deterministic components The comparison is made between minimum statistics, the proposed DFT-domain subspace noise tracking approach and the DFT-domain subspace noise tracking method combined with (8.15). The obtained distortion for these partly deterministic noise types is decreased by combining the proposed noise tracker with (8.15). Notice that the experimental results in Section 8.6.1 are based on the use of the DFT-domain subspace noise tracker without the use of a deterministic noise tracker.

noise source	input SNR	mean score for method A
white noise	5 dB	82.0 %
	15 dB	85.9 %
street noise	5 dB	77.3 %
	15 dB	67.2 %
passing train	5 dB	77.3 %
	15 dB	92.2 %
Non-stat. white noise	5 dB	96.1 %
	15 dB	89.1 %

Table 8.5: *Listening test results.*

noise source	input SNR (dB)	MS	prop. noise tracker	prop. noise tracker combined with (8.15)
white noise with sinusoids	0	1.2	2.1	1.1
	5	1.2	1.7	1.1
	10	1.2	1.5	1.1
	15	1.2	1.3	1.2
Destroyer operations room	0	1.8	1.7	1.5
	5	1.9	1.9	1.7
	10	2.0	2.1	1.9
	15	2.1	2.5	2.3

Table 8.6: *Performance in terms of* LOG-Err_{seg} *to compare the influence of a deterministic noise tracker.*

8.7 Concluding Remarks

In this chapter we presented a novel approach for noise tracking. The method is based on construction of correlation matrices in the DFT-domain per time-frequency point. Each correlation matrix can be decomposed into a signal subspace and a noise-only subspace. When the signal subspace is not full rank, the noise-only subspace can be used to estimate the noise PSD. The advantage of this approach is that the noise PSD can be updated for a DFT coefficient where both speech and noise are present. Comparisons showed that the presented method decreases the error between the true noise and the estimated noise spectrum, compared to the minimum statistics. Further, enhancement performance is improved, especially for speech signals degraded by noise types that change gradually in power. Deterministic noise sources appear in the signal subspace and can not be estimated by observing the noise-only subspace. However, these noise components can be tracked by observing T last minimum statistics based noise PSD estimates.

The improved noise tracking performance of the proposed DFT-domain subspace noise tracker over minimum statistics comes with an increase in the computational

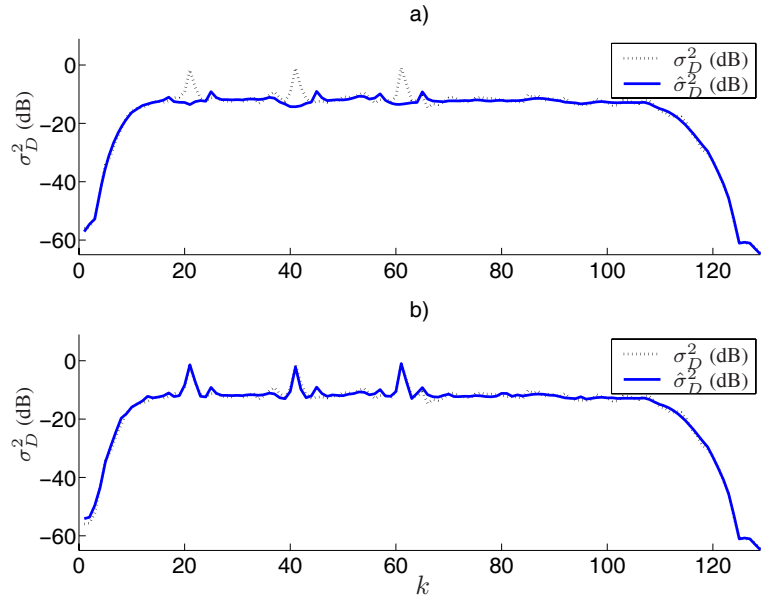

Figure 8.8: *a) Noise tracking performed with DFT-domain subspace decompositions only. b) Noise tracking performed with DFT-domain subspace decompositions combined with a tracker for deterministic components.*

complexity. Although the dimensions of the correlation matrices are rather small, most of the computation time is spent on eigenvalue decompositions of the noisy correlation matrices. However, the MATLAB implementation of the proposed algorithm runs approximately two times real time on a PC with pentium 4 processor.

References

[1] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[2] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

[3] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[4] P. J. Wolfe and S. J. Godsill. Efficient alternatives to the Ephraim and Malah suppressionrule for audio signal enhancement. *EURASIP Journal on Applied Signal Processing*, 10:1043–1051, 2003.

[5] J. Sohn, N. S. Kim, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Lett.*, 6(1):1–3, January 1999.

[6] J. Chang, N. S. Kim, and S. K. Mitra. Voice activity detection based on multiple statistical models. *IEEE Trans. Signal Processing*, 54(6):1965–1976, June 2006.

[7] R. Martin. Spectral subtraction based on minimum statistics. In *Proc. Eur. Signal Processing Conf.*, pages 1182–1185, 1994.

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

[9] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. Acoust., Speech, Signal Processing*, 34(4):744–754, Aug. 1986.

[10] D. R. Brillinger. *Time Series: Data Analysis and Theory*. SIAM, Philadelphia, 2001.

[11] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, 19(6):716–723, 1974.

[12] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33:387–392, 1985.

[13] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[14] P. Stoica and Y. Selén. Model-order selection: a review of information criteriun rules. *IEEE signal Processing magazine*, 21(4):36–47, 2004.

[15] R. C. Hendriks, J. Jensen, and R. Heusdens. Noise tracking using DFT domain subspace decompositions. Technical report ICT-2007-03, 2007.

[16] S. K. Kay. *Fundamentals of Statistical signal processing*, volume 2. Prentice Hall, Upper Saddle River, NJ, 1998.

[17] R. Martin. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, 86(6):1215–1229, June 2006.

[18] C. W. Therrien. *Discrete Random Signals and Statistical Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ, 1992.

[19] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, 1999.

[20] Y. Hu and P. Loizou. Subjective comparison of speech enhancement algorithms. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, volume 1, pages 153–156, May 2006.

[21] A. Varga and H. J. M. Steeneken. Noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247–253, 1993.

[22] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Processing*, 11(5):446–475, September 2003.

Chapter 9

Conclusions and Discussion

9.1 Summary and Discussion of Results

In this thesis we considered DFT-domain based speech enhancement and focussed on three different aspects. First we considered methods to improve estimation of the *a priori* SNR. Good estimation of the *a priori* SNR is important, because most speech enhancement estimators are a function of this quantity. Wrong estimates, i.e. an underestimate or overestimate of the *a priori* SNR, can lead to oversuppression or undersuppression of the noisy speech DFT coefficients, respectively.

Most methods for *a priori* SNR estimation are (partly) based on an estimate of the noisy speech PSD. In order to improve estimation of the noisy speech PSD, and therefore also estimation of the *a priori* SNR, we exploited the fact that speech is a time-varying process and presented an adaptive time-segmentation algorithm for noisy speech. This algorithm finds for each frame a corresponding segment in which the data can be considered stationary. We applied this segmentation algorithm to obtain better estimates of the noisy speech PSD. We showed that estimation of the *a priori* SNR with the decision-directed approach can be improved using this improved estimate of the noisy speech PSD instead of a periodogram based estimate. Further, we presented a backward decision-directed approach which can be combined with the standard (forward) decision-directed approach to overcome a consistent overestimate or underestimate of the *a priori* SNR at the start of stationary regions. This backward decision-directed approach is dependent on the next future frame instead of on the previous frame. As such it overcomes wrong estimates of the *a priori* SNR at the start of stationary regions.

Secondly, we investigated estimators for clean speech DFT coefficients that take properties of speech DFT coefficients into account. We presented an MMSE estimator under a combined stochastic-deterministic speech model. The use of a deterministic speech model is based on the idea that certain speech sounds have a more deterministic character, e.g. voiced sounds. Especially in frequency bins containing a speech harmonic this combined stochastic-deterministic speech model leads to improved enhancement performance in comparison to the use of a stochastic model alone. Besides

estimators under a combined stochastic-deterministic model, we presented estimators under super-Gaussian densities as well. MMSE estimators for complex DFT coefficients and DFT magnitudes are derived under the assumption that they have a double-sided and single-sided generalized Gamma density, respectively. These estimators are a generalization of existing MMSE estimators proposed in [1] and [2]. Further, MAP estimators for complex DFT coefficients and DFT magnitudes were derived under the assumption that clean speech DFT coefficients have a multivariate normal inverse Gaussian density (MNIG). The MNIG density can model scalar processes as well as vector processes. The estimators derived under the MNIG density can exploit this property and take the dependency between real and imaginary parts of DFT coefficients into account. This is a potential advantage over the generalized Gamma density based estimators, since these estimators assume that real and imaginary parts are independent. However, taking the dependency into account leads only to a very small improvement in terms of enhancement performance. A second advantage of estimators derived under the MNIG density over the generalized Gamma density is the fact that under the MNIG density the models in the complex DFT domain and the polar domain are consistent, which is not the case for the generalized Gamma density.

Thirdly, we presented a new method for tracking of noise statistics. The method that we propose is based on the eigenvalue decomposition of correlation matrices that are constructed from time series of noisy DFT coefficients. The noise level is estimated per frequency bin by observing the noise-only subspace of these correlation matrices. Hence, even when speech is continuously present, the proposed method can estimate the noise level. Changes in the noise PSD can be tracked much faster than with existing noise tracking algorithms like minimum statistics [3] that has a delay of approximately 1 second when the noise level increases.

9.1.1 Discussion on Contributions

In this section we present a further discussion on the contributions of this thesis.

1. **Adaptive time segmentation for speech enhancement**

   The algorithm for adaptive time segmentation of noisy speech is rather general and uses the noisy speech signal as an input. It can therefore be combined with almost any enhancement system. The algorithm uses time frames from the past as well as frames from the future in order to determine the segmentation of the noisy speech signal. Because latency is application dependent, the number of future frames that are used in the algorithm can be adjusted. For human-to-human applications hardly any (e.g. hearing aids) or some (e.g. mobile telephony) latency is allowed, while for human-to-machine communication the amount of latency is less of an issue. The more latency is allowed in the enhancement system, the higher the gain of using the adaptive time-segmentation algorithm. However, it was shown that the performance obtained with a latency of approximately 30 ms is fairly close to the use of an infinite latency.

   In this thesis we used the adaptive time segmentation in Chapter 3 to improve estimation of the noisy speech PSD. These improved estimates of the noisy speech

PSD are then used to obtain improved estimates of the *a priori* SNR or of the speech PSD. In Chapter 4 adaptive time segmentation was used to combine the forward and backward decision-directed approach, in order to obtain improved estimates of the *a priori* SNR. In [4] the adaptive time-segmentation algorithm was used within a subspace based speech enhancement context, where the adaptive segmentation was used to obtain better estimates of the noisy speech correlation matrix.

2. **Improved *a priori* SNR estimation**

Two methods were presented to improve estimation of the *a priori* SNR. The first method was presented in Chapter 3 and is based on a modification of the decision-directed approach. Instead of a noisy speech periodogram, an estimate of the noisy PSD based on the presented adaptive time segmentation was used. As mentioned in the discussion on the previous contribution, the adaptive time segmentation implies a certain delay when using future frames to determine the segmentation. Although some improvement is still obtained when no future information is used, most improvement is obtained when a delay of approximately 30 ms is allowed.

The second method for improved *a priori* SNR estimation is based on the in Chapter 4 presented backward decision-directed approach. Combined with the forward decision-directed approach an improved estimate of the *a priori* SNR is made. A potential disadvantage of this method is that the backward decision-directed approach has a dependency on future frames and therefore implies a delay. By combining the forward and the backward approach it is possible to limit the delay to one frame.

Both methods for improved *a priori* SNR estimation are thus based on the use of some future information. Although the resulting delay can be limited, both methods are less applicable when absolutely no delay is allowed. By combining these two approaches for improved *a priori* SNR estimation an improvement in terms of segmental SNR of approximately 0.5 dB can be achieved over the situation when only one of the two methods is used. The obtained performance improvement is mainly reflected in terms of a better noise suppression. However, it results in somewhat more suppressed speech as well.

The two methods for improved *a priori* SNR estimation can be combined with any DFT-domain based clean speech estimator that can be expressed in terms of the *a priori* SNR, e.g. estimators presented in [5][1][6][2][7] and the estimators presented in Chapter 6. These methods for improved *a priori* SNR estimation can be used in combination with estimators under the combined stochastic-deterministic model that is proposed in Chapter 5 as well. However, it is not expected that this combination will lead to the same amount of enhancement performance as it does for the estimator used in Chapters 3 and 4. This can be explained by the fact that under the deterministic model the estimator is not dependent on the *a priori* SNR. Estimators under the MNIG density that are presented in Chapter 7 can not be written as a function of the *a priori* SNR

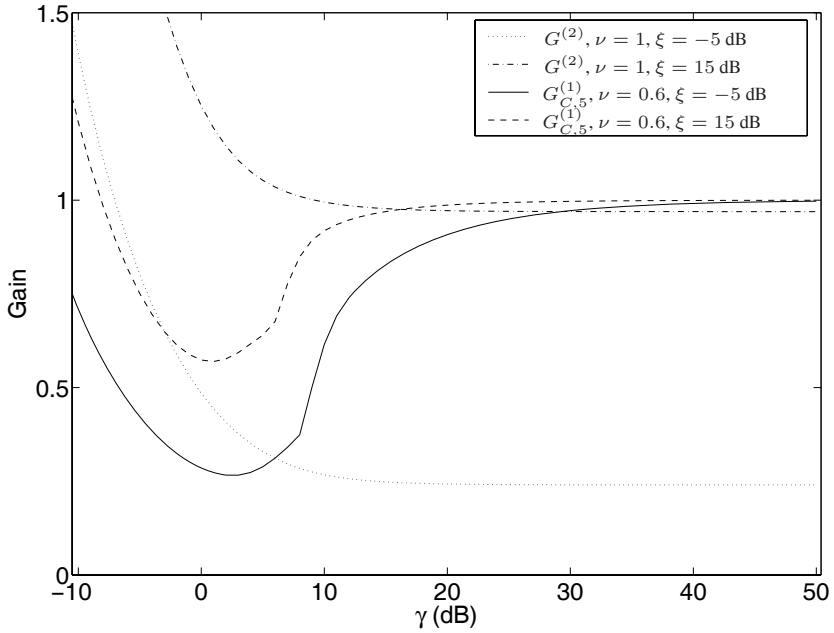

Figure 9.1: *Comparison between gain curves.*

and can therefore not directly be combined with these improved *a priori* SNR estimators.

The amount of performance improvement that is obtained by combining a clean speech estimator with the forward-backward decision-directed approach for *a priori* SNR estimation is dependent on the type of estimator that is used. More specifically, some estimators are able to compensate for under or overestimates of the *a priori* SNR. This also holds, under certain parameter settings, for the complex DFT and DFT magnitude estimators that are derived under the generalized Gamma density in Chapter 6. In particular this is true when the assumed underlying density for the speech DFT coefficients corresponds to a super-Gaussian density. Estimators under these densities have the tendency to increase the value of the gain function when the *a priori* SNR is underestimated and much lower than the *a posteriori* SNR. A similar situation occurs when the *a priori* SNR is overestimated and larger than the *a posteriori* SNR. In that case, the value of the gain function is somewhat decreased. To visualize this effect, we show in Fig. 9.1 gain curves of the MMSE magnitude estimator under a generalized Gamma density with $\gamma = 1$ and $\nu = 0.6$ (denoted by $\hat{G}^{(1)}_{C,5}$, $\nu = 0.6$) and under the generalized Gamma density with $\gamma = 2$ and $\nu = 1$ (the Rayleigh density, denoted by $\hat{G}^{(2)}$, $\nu = 1$). The latter gain function corresponds to the MMSE magnitude estimator presented in [1]. The gain curves in Fig. 9.1 are plotted as a function of the *a posteriori* SNR for the *a priori* SNRs of -5 dB

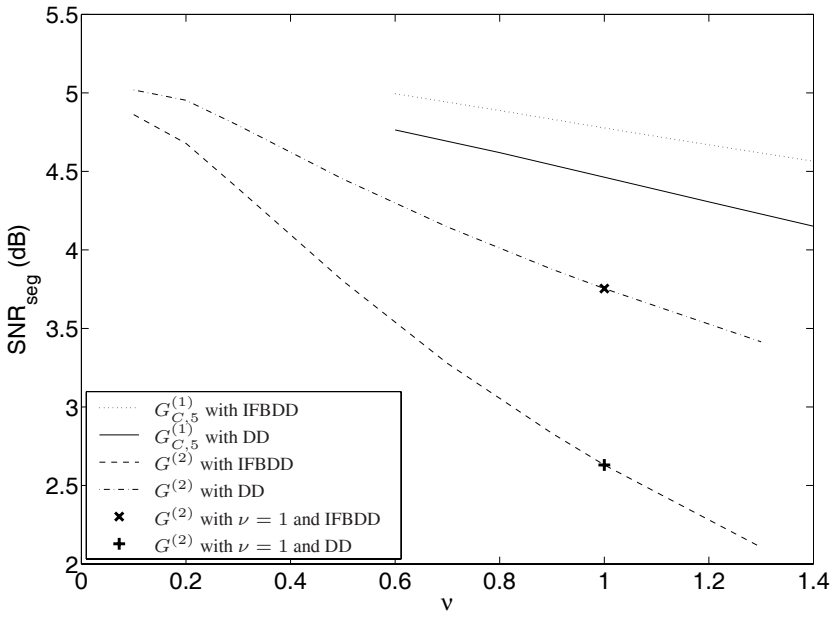

Figure 9.2: *Comparison between MMSE amplitude estimators under the generalized gamma density with decision-directed approach and with iterative forward-backward decision-directed approach.*

and 15 dB. We see that the gain curves for the estimator with parameter settings $\gamma = 1$ and $\nu = 0.6$ show a much higher degree of sensitivity to the *a posteriori* SNR than the estimator with parameter settings $\gamma = 2$ and $\nu = 1$. This sensitivity to the *a posteriori* SNR makes it possible to compensate for an underestimated or overestimated *a priori* SNR.

Estimators with these properties have a smaller benefit of improved *a priori* SNR estimation. To demonstrate the impact of this mechanism we compare the decision-directed approach and the iterative forward-backward decision-directed approach in terms of enhancement performance. Both *a priori* SNR estimators are combined with an MMSE magnitude estimator under the generalized Gamma density for a wide range of parameter settings. The performance is shown in Fig. 9.2 in terms of average segmental SNR versus the ν-parameter. We see that the performance difference between the decision-directed approach and the iterative forward-backward decision-directed approach is decreased when the parameter settings of the assumed underlying density for the DFT magnitudes corresponds to super-Gaussian densities ($\gamma = 1$ and $\gamma = 2$ with small ν). The estimator that was used in Chapter 4 corresponds to parameter settings $\gamma = 2$ and $\nu = 1$ (Rayleigh density) and is indicated in Fig. 9.2 by the $+$ and the $\times$ for the decision-directed approach and the iterative forward-backward decision-directed approach, respectively.

3. **Clean speech DFT estimator under a combined stochastic-deterministic model**

   In Chapter 5 experiments were presented where the estimator derived under a combined stochastic-deterministic model was compared with an estimator using a stochastic model alone. For both estimators we used in these comparisons as stochastic speech model the Gaussian and Laplace density. The combined stochastic-deterministic model is not restricted to be used with these densities, but can also be combined with the generalized Gamma or MNIG density that were used in Chapters 6 and 7, respectively. However, the performance improvement of the combined stochastic-deterministic model over the use of a stochastic model alone gets smaller when the stochastic speech model that is used tends to be more super-Gaussian. This can be explained by the fact that both the deterministic model and stochastic super-Gaussian models partly improve on the same aspects. More specifically, both the deterministic model and stochastic super-Gaussian models are better resistant towards underestimates or overestimates of the *a priori* SNR. Secondly, both the deterministic and super-Gaussian models are improved models for DFT coefficients containing speech harmonics.

4. **Clean speech DFT estimators under super-Gaussian densities**

   Clean speech complex DFT and magnitude DFT estimators were presented under the generalized Gamma density and under the multivariate normal inverse Gaussian (MNIG) density. Objective performance of the estimators under the MNIG density is somewhat better than for the estimators under the the generalized Gamma density. Moreover, estimators under the MNIG density have some theoretical advantages over estimators under the generalize Gamma density. For the MNIG density, the models in the complex DFT domain and the polar domain are consistent, which is not the case for the generalized Gamma density. Secondly, the MNIG density can model vector processes and can take the dependency between the real part and imaginary part of DFT coefficients into account. However, taking the dependency between real and imaginary parts of DFT coefficients into account hardly leads to any performance improvement in practice. Despite these theoretical advantages, the computational complexity for computing the parameters that specify the estimator under the MNIG density is higher than for estimators under the generalized Gamma density. This makes the MNIG based estimators less applicable in situations where computational complexity is an issue.

5. **Tracking of noise statistics**

   The DFT-domain subspace based method for noise tracking that was presented in Chapter 8 clearly improves over existing, state-of-the-art, noise tracking methods like minimum statistics, especially for non-stationary noise where the tracking delay is considerably reduced. This followed from direct measurements of the noise tracking performance as well as from measurements of the enhancement performance when combined with an enhancement system. The computational complexity of the algorithm is mainly dominated by the eigenvalue

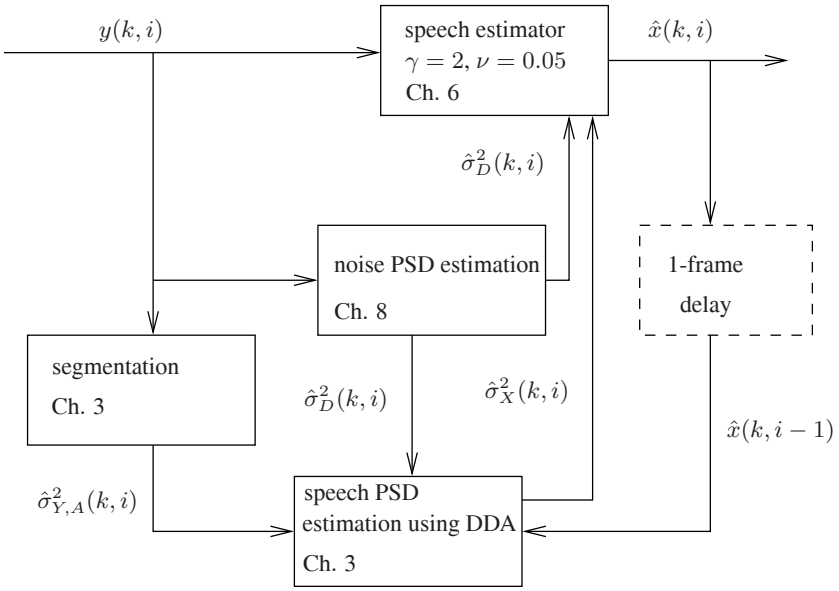

Figure 9.3: *DFT-domain enhancement scheme based on contributions of this thesis.*

decompositions that have to be performed and is therefore somewhat higher than for minimum statistics based noise tracking. The proposed noise tracking method is not dependent on the type of clean speech estimator and can as such be combined within any type of DFT-domain based enhancement algorithm, for example the estimators presented in Chapters 5, 6 and 7 or the improved *a priori* SNR estimators from Chapters 3 and 4.

9.1.2 Comparison to State-of-the-art Speech Enhancement System

As discussed in Section 9.1.1, some of the contributions of this thesis can be combined to form a complete DFT-domain based speech enhancement system. In this section we make a comparison between such a system and a state-of-the-art system by means of a listening test. The system that is based on contributions of this thesis is depicted in Fig. 9.3 and will be referred to as method A. Here we use the adaptive time segmentation discussed in Chapter 3 to obtain an estimate of the noisy speech PSD, denoted by $\hat{\sigma}_{Y,A}^2(k,i)$. For estimation of the noise PSD we use the DFT-domain subspace based method that is proposed in Chapter 8. Using these estimates of the noise PSD, noisy speech PSD and the estimated clean speech DFT from the previous frame, the speech PSD is estimated based on the method for improved a priori SNR estimation that was presented in Chapter 3, denoted by DDA. To estimate the magnitude of the clean speech DFT coefficients, the magnitude estimator under the generalized Gamma density is used, which is proposed in Chapter 6. Here we use parameter settings $\gamma = 2$

and $\nu = 0.05$, that are based on initial experiments with this complete system.

The state-or-the-art system that we compare to is shown in Fig. 9.4. We refer to this system as method B. For noise PSD estimation we use minimum statistics as proposed in [3]. Estimation of the speech PSD is based on the decision-directed approach [1] and the estimator that we use for estimation of the clean speech magnitudes is the joint MAP amplitude and phase estimator presented in [7].

The listening test that we perform is a so-called OAB test with 8 participants. Here, O is the original clean speech signal and A and B are two noisy signals that are enhanced using method A and method B, respectively. The listeners were presented first the original signal followed by the two different enhanced signals A and B played in random order. The participants had to indicate their preference for excerpt A or B. Each series was repeated 4 times, with each time a randomized order of the signals A and B. This leads for each set to a score of 0, 25, 50, 75 or 100 %, indicating the relative preference for excerpt A. This procedure is repeated for four different speech signals (two male and two female sentences). The speech signals that were used in the listening test were degraded under 6 different conditions, namely with three different noise types (white noise, street noise and noise from a passing train) and at two different SNRs (5 dB and 15 dB). The histograms of the listening-test scores for each of the test conditions are shown in Fig. 9.5. From these histograms we see that under all conditions the majority of the participants had a preference for method A, the system based on contributions of this thesis. Notice that for speech signals degraded by passing train noise at 5 dB SNR a clear minority of the participants preferred method B. For these signals the noise source is rather non-stationary and increases a lot in power. The proposed DFT-domain subspace based method for noise tracking performs a much better tracking of the noise level than minimum statistics, as also shown in Fig. 8.6. This leads to more suppression of the noise, but implies more suppression of the speech as well. Analysis of the results and discussion with the participants of the listening test revealed that some of the participants preferred method B for these signals, because the underestimated noise PSD leads to less suppression of speech. Too much suppression of speech could be overcome by limiting the clean speech estimator to a maximum suppression. The average preference for method A under each test condition is shown in Table 9.1. A Wilcoxon signed-ranks test [8] was performed to find out whether the difference between the two tested systems based on the scores of the listening test is significant. The P-values of the Wilcoxon signed-ranks test are shown in Table 9.1 as well. They indicate that under all test conditions, the difference between our proposed system and the state-of-the-art system is significant at a significance level of $1.2 \cdot 10^{-3}$.

9.2 Directions for Future Research

Although the world of DFT-domain based speech enhancement has almost become mature, there are still many challenges that can lead to a further quality improvement of single-channel speech enhancement algorithms.

An important aspect is the further development of algorithms for fast and accurate tracking of noise statistics, without constraining these systems too work only under

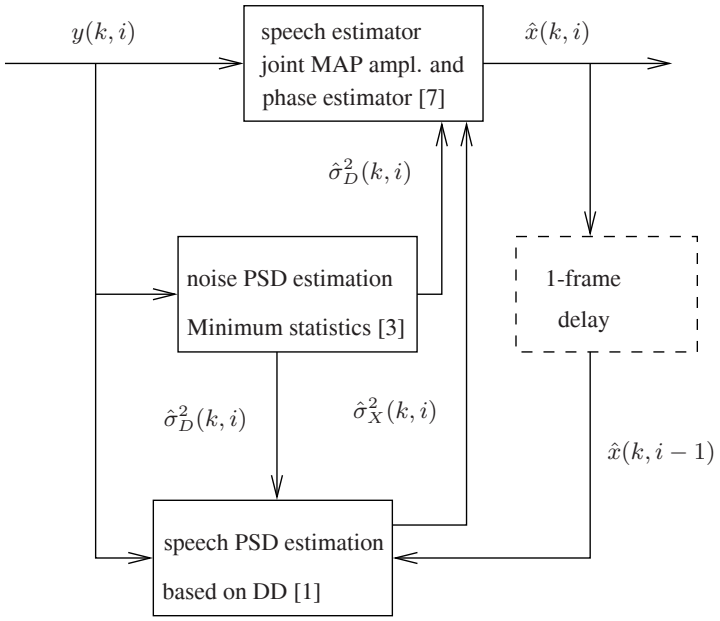

Figure 9.4: *State-of-the-art enhancement scheme.*

specific situations, e.g. a certain speaker or noise-type. Imposing such constraints restricts the applicability of these algorithms. Fast and accurate tracking of noise statistics is of vital importance, because all clean speech estimators are dependent on these noise statistics. Using more sophisticated models and making better use of the available data can help in the development of improved noise tracking algorithms. The method for tracking of noise statistics that we presented in Chapter 8 is based on processing of the noisy speech data in the DFT subspace domain. It was shown that large improvements are obtained over minimum statistics based schemes. A further development of these types of approaches, that even allow tracking of noise statistics when speech is constantly present, will lead to a further improvement of the enhancement performance. A suggestion for further development of the method presented in Chapter 8 is to take into account that speech is a time-varying process. By doing so, estimation of the noisy speech correlation matrix in Eq. (8.4) can be improved, possibly resulting in a lower dimension of the signal subspace and lower variance on the estimated noisy speech correlation matrix and as a consequence a better estimate of the noise PSD.

Another important direction for future research is to study how intelligibility of enhanced speech signals can be improved with respect to existing enhancement algorithms. In general, existing methods improve quality in terms of noise suppression, but decrease quality in terms of speech intelligibility. A challenging direction of research would be to investigate how the decrease in intelligibility can be restricted while still obtaining good noise reduction.

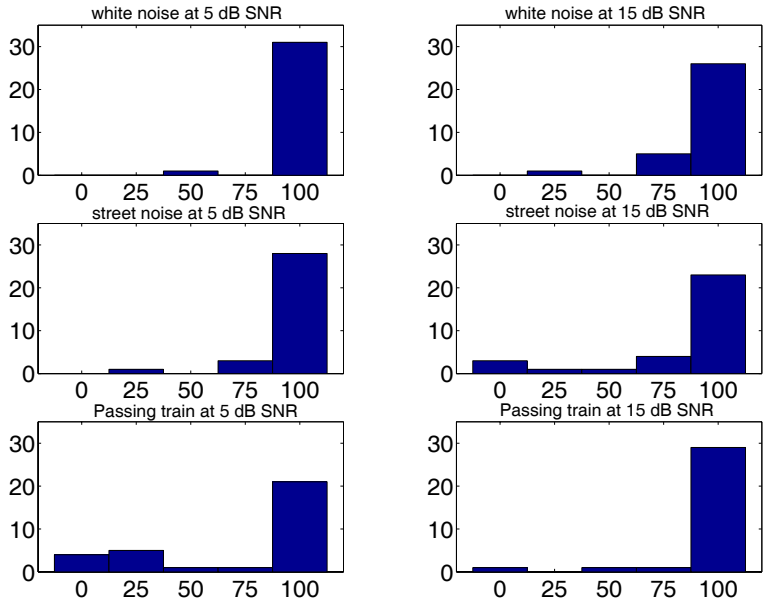

Figure 9.5: *Histograms of listening-test scores indicating the preference for method A.*

An interesting development with respect to clean speech estimators is presented in [9][10], where a methodology is proposed to estimate speech enhancement gain functions directly from speech signals in a training data-base. The advantage of this procedure is that no explicit assumption is made about the density of the speech DFT coefficients. Moreover, the approach that is followed in [9][10] takes into account that the true *a priori* SNR is unknown, but that only an estimate is available. This data-driven approach has the advantage that certain systematic modelling errors and estimation inaccuracies and consistencies can automatically be taken into account. Moreover, this approach can optimize for all kind of distortion measures for which analytical results are hard to obtain.

Another interesting direction for future research is to take correlation across time into account. Most speech enhancement algorithms apply the estimators independently for each time and frequency point, see e.g. Chapters 6 and 7, but also [1][2][7]. When no special measures are taken, these types of methods will lead to musical noise when estimates of the noise and clean speech PSD are substituted in the clean speech estimator. Smoothing methods like the decision-directed approach are used to avoid this. Although the decision-directed approach leads to a reduction of musical noise, it is also conflicting with the assumption that the clean speech estimators can be applied independently across time.

It would be interesting to develop estimators that automatically take (some of) the correlation between DFT coefficients across time into account. As an example, the framework presented in Chapter 8 could be very well used for this purpose. In

noise source	input SNR	mean score for method A	P-value	significant
white noise	5 dB	98.4 %	$2.6 \cdot 10^{-8}$	yes
	15 dB	93.8 %	$1.8 \cdot 10^{-7}$	yes
street noise	5 dB	95.3 %	$9.5 \cdot 10^{-8}$	yes
	15 dB	83.6 %	$6.1 \cdot 10^{-5}$	yes
passing train	5 dB	73.4 %	$1.2 \cdot 10^{-3}$	yes
	15 dB	94.5 %	$3.0 \cdot 10^{-7}$	yes

Table 9.1: *Results of the listening test.*

this framework, correlation matrices are constructed from time series of noisy DFT coefficients. In Chapter 8 the noise PSD is obtained from these correlation matrices by exploiting the noise-only subspace. Instead of only estimating the noise PSD, this framework could be applied as well to estimate a series of clean DFT coefficients while taking into account their correlation across time.

Besides further development of single and multi-microphone enhancement schemes, it would also be interesting to investigate how multiple speech enhancement systems can cooperate in an adaptive manner. More specifically, the current generation of voice processors work individually, although in some situations several voice processors share the same information. Consider the situation where several hearing aid users are on the same location. Instead of processing the noisy acoustical environment for all persons individually, their processors should be combined, form a larger microphone array and distribute the work load. Research on this jointly type of processing will be challenging, and might lead to a different and new view on speech enhancement and might change the insight in how to solve the speech enhancement problem.

References

[1] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, December 1984.

[2] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Processing*, 13(5):845–856, Sept. 2005.

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

[4] R. C. Hendriks, R. Heusdens, and J. Jensen. Improved subspace based speech enhancement using an adaptive time segmentation. In *Proc. IEEE First BENELUX/DSP Valley Signal Processing Symposium*, pages 163–166, April 2005.

[5] N. Wiener. *Extrapolation, Interpolation and Smoothing of Stationary Time Series: With Engineering Applications*. MIT Press, principles of electrical engineering series edition, 1949.

[6] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(2):443–445, April 1985.

[7] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, May 2005.

[8] D. J. Sheskin. *Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, 3rd edition edition, 2004.

[9] J. Jensen and R. Heusdens. A numerical approach for estimating optimal gain functions in single-channel DFT based speech enhancement. In *Proc. European Signal Proc. Conf. Eusipco*, Florence, Italy, 2006.

[10] J. S. Erkelens, J. Jensen, and R. Heusdens. A general optimization procedure for spectral speech enhancement methods. In *Proc. European Signal Proc. Conf. Eusipco*, 2006.

Appendix A

Derivations for Chapter 6

A.1 Second Moments

In this appendix we derive expressions for the second moments of the random variables with densities $f_A(a)$ and $f_{X_\Re}(x_\Re)$ as given by Eqs. (6.1) and (6.2) for the cases $\gamma = 1$ and $\gamma = 2$.

A.1.1 The Single-sided Prior $f_A(a)$

The second moment of A for the case $\gamma = 1$

Using the assumption $E\{X\} = 0$, we can write $\sigma_X^2 = E\{A^2\}$. Using [1, Eq. 3.381.4] and (6.1) with $\gamma = 1$, it can be shown that

$$\sigma_X^2 = \int_{-\infty}^{\infty} a^2 f_A(a) da = \frac{\nu(\nu+1)}{\beta^2}. \tag{A.1}$$

The second moment of A for the case $\gamma = 2$

Using the assumption $E\{X\} = 0$, we can write $\sigma_X^2 = E\{A^2\}$. Using [1, Eq. 3.381.4], Eq. (6.1) with $\gamma = 2$ and the substitution $a = \sqrt{t}$, it can be shown that

$$\sigma_X^2 = \int_{-\infty}^{\infty} a^2 f_A(a) da = \frac{\nu}{\beta}. \tag{A.2}$$

A.1.2 The Two-sided Prior $f_{X_\Re}(x_\Re)$

The second moment of $X_\Re$ for $\gamma = 1$

Using the assumption $E\{X_\Re\} = 0$, we can write $\sigma_{X_\Re}^2 = E\{X_\Re^2\}$, i.e. the variance equals the second moment. Using [1, Eq. 3.462.9] and (6.2) with $\gamma = 1$, it can be shown that

$$\sigma_{X_\Re}^2 = \int_{-\infty}^{\infty} x_\Re^2 f_{X_\Re}(x_\Re) dx_\Re = \frac{\nu(\nu+1)}{\beta^2}. \tag{A.3}$$

The second moment of $X_\Re$ for $\gamma = 2$

Using the assumption $E\{X_\Re\} = 0$, we can write $\sigma^2_{X_\Re} = E\{X^2_\Re\}$. Using [1, Eq. 3.462.9] and (6.2) with $\gamma = 2$, it can be shown that

$$\sigma^2_{X_\Re} = \int_{-\infty}^{\infty} x^2_\Re f_{X_\Re}(x_\Re) dx_\Re = \frac{\nu}{\beta}. \tag{A.4}$$

A.2 Modified MAP Estimator

The estimator originally proposed in [2] was computed as

$$\max_a \log f_A(a) f_{R|A}(r|a) \tag{A.5}$$

with $f_{R|A}(r|a)$ as in Eq. (6.6). However, in [2] $f_A(a)$ was not used in the form Eq. (6.1) with $\gamma = 1$, but in a slightly different form:

$$f_A(a) = \frac{a^{\nu-1}}{\Gamma(\nu)} \left(\frac{\mu}{\sigma_X}\right)^\nu \exp\left\{-a\frac{\mu}{\sigma_X}\right\}, \tag{A.6}$$

where μ and ν were treated as independent parameters although μ is in fact completely specified by ν, see Eq. (A.9) below. Since an analytical solution to Eq. (A.5) is hard to find, the approximation Eq. (6.13) for the Bessel function was made *before* taking the derivative with respect to a in Eq. (A.5). This led to the gain function

$$G^{(1)}_{MAP} = u + \sqrt{u^2 + \frac{\nu' - 0.5}{2\zeta}} \quad , \quad u = 1/2 - \frac{\mu}{4\sqrt{\zeta\xi}}, \tag{A.7}$$

where $\nu' = \nu - 1$, and which is only valid for $\nu' > 0.5$. A joint amplitude and phase MAP estimator was proposed as well. The gain function $G^{(1)}_{JMAP}$ of the joint MAP estimator is given by

$$G^{(1)}_{JMAP} = u + \sqrt{u^2 + \frac{\nu'}{2\zeta}} \quad , \quad u = 1/2 - \frac{\mu}{4\sqrt{\zeta\xi}}, \tag{A.8}$$

which allows for a broader range of ν'-values, namely $\nu' > 0$. The parameters ν' and μ were estimated in [2] by fitting Eq. (A.6) to clean-speech amplitude distributions conditioned on a small range of high values of estimated *a priori* SNR.

The first of the modifications we make to this estimator concerns the number of free parameters in Eqs. (A.6), (A.7), and (A.8). We see that μ and σ_S do not appear independently in Eq. (A.6), but only as the quotient μ/σ_X, and therefore only represent one degree of freedom. The parameter ν represents the second degree of freedom. Since $E\{A^2\}$ equals σ^2_X by definition, it follows from Eq. (A.1) that

$$\mu = \sqrt{\nu(\nu+1)}. \tag{A.9}$$

The second modification concerns the order in which the approximation of the bessel function is used and the derivative of Eq. (A.5) is taken. More specifically, we

compute the amplitude MAP estimator by *first* taking the derivative and *then* using the large-argument approximation $\mathcal{I}_1/\mathcal{I}_0 \approx 1$, where $\mathcal{I}_1$ is the first-order modified Bessel function of the first kind. Interestingly, the resulting MAP gain function is identical to the joint MAP gain function in Eq. (A.8), with μ given by Eq. (A.9).

References

[1] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, Inc., 6 edition, 2000.

[2] T. Lotter and P. Vary. Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model. *EURASIP Journal on Applied Signal Processing*, 7:1110–1126, 2005.

Appendix B

Derivations for Chapter 7

B.1

In this appendix we outline the steps to compute the MAP estimator under the MNIG distribution for the complex DFT coefficients. The derivative of $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ can be computed as

$$\begin{aligned}
\frac{d}{d\mathbf{x}} \ln[f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})] &= \frac{d}{d\mathbf{x}} \ln[f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x})] + \frac{d}{d\mathbf{x}} \ln[f_{\mathbf{X}}(\mathbf{x})] \\
&= {\boldsymbol{\lambda}_{\mathbf{D}}}^{-1}(\mathbf{y}-\mathbf{x}) + \frac{\frac{d}{d\mathbf{x}} \int_{\lambda_X} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int_{\lambda_X} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \\
&= {\boldsymbol{\lambda}_{\mathbf{D}}}^{-1}(\mathbf{y}-\mathbf{x}) - \frac{\int_{\lambda_X} \lambda_X^{-1} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int_{\lambda_X} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \boldsymbol{\Gamma}^{-1}\mathbf{x} \\
&= {\boldsymbol{\lambda}_{\mathbf{D}}}^{-1}(\mathbf{y}-\mathbf{x}) - E\left[\Lambda_X^{-1}|\mathbf{x}\right] \boldsymbol{\Gamma}^{-1}\mathbf{x}. \qquad \text{(B.1)}
\end{aligned}$$

Solving

$$\frac{d}{d\mathbf{x}} \ln[f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})] = 0$$

then leads to (7.25). Further, using [1, Th. 3.471,9] it can be shown that

$$\begin{aligned}
E\left[\Lambda_X^{-1}|\mathbf{x}\right] &= \frac{\int_{\lambda_X} \lambda_X^{-1} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int_{\lambda_X} f_{\mathbf{X}|\Lambda_X}(\mathbf{x}|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \\
&= \frac{\int_{\lambda_X=0}^{\infty} {\lambda_X^{-1}}^{\frac{1}{2}-d/2} \exp\left[-\frac{1}{2}\left((\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})\lambda_X^{-1} + \alpha^2\lambda_X\right)\right] d\lambda_X}{\int_{\lambda_X=0}^{\infty} {\lambda_X^{-1}}^{\frac{1}{2}-d/2} \exp\left[-\frac{1}{2}\left((\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})\lambda_X^{-1} + \alpha^2\lambda_X\right)\right] d\lambda_X} \\
&= \left(\frac{\alpha^2}{\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x}}\right)^{\frac{1}{2}} \frac{\mathcal{K}_{\frac{3+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})}\right)}{\mathcal{K}_{\frac{1+d}{2}}\left(\sqrt{\alpha^2(\delta^2 + \mathbf{x}^T\boldsymbol{\Gamma}^{-1}\mathbf{x})}\right)}, \qquad \text{(B.2)}
\end{aligned}$$

where $\mathcal{K}_{d'}$ denotes the modified Bessel function of the second kind with order d'.

B.2

In this appendix we outline the steps to compute the MAP estimator under the RIG distribution for the DFT amplitudes coefficients. The derivative of $f_{A|R}(a|r)$ can be computed as

$$\begin{aligned}
\frac{d}{da}\ln[f_{A|R}(a|r)] &= \frac{d}{da}\ln[f_{R|A}(r|a)] + \frac{d}{da}\ln[f_A(a)] \\
&= 2\frac{-a+r}{\sigma_D^2} - \frac{1}{2a} + \frac{\int \frac{d}{da} f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \\
&= 2\frac{-a+r}{\sigma_D^2} + \frac{1}{2a} - a\frac{\int \lambda_X^{-1} f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X}{\int f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \\
&= 2\frac{-a+r}{\sigma_D^2} + \frac{1}{2a} - aE\left[\Lambda_X^{-1}|a\right] \qquad \text{(B.3)}
\end{aligned}$$

Solving

$$\frac{d}{da}\ln[f_{A|R}(a|r)] = 0 \qquad \text{(B.4)}$$

then leads to (7.35). Further, using [1, Th. 3.471,9] it can be shown that

$$\begin{aligned}
E\left[\Lambda_X^{-1}|a\right] &= \frac{\int \lambda_X^{-1} f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda}(\lambda_X) d\lambda_X}{\int f_{A|\Lambda_X}(a|\lambda_X) f_{\Lambda_X}(\lambda_X) d\lambda_X} \\
&= \frac{\int \lambda_X^{-3\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\delta^2+a^2\right)\lambda_X^{-1} - \frac{1}{2}\alpha^2\lambda_X\right] d\lambda_X}{\int \lambda_X^{-2\frac{1}{2}} \exp\left[-\frac{1}{2}\left(\delta^2+a^2\right)\lambda_X^{-1} - \frac{1}{2}\alpha^2\lambda_X\right] d\lambda_X} \\
&= \left(\frac{\alpha^2}{\delta^2+a^2}\right)^{\frac{1}{2}} \frac{\mathcal{K}_{2\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}{\mathcal{K}_{1\frac{1}{2}}\left(\alpha\sqrt{\delta^2+a^2}\right)}, \qquad \text{(B.5)}
\end{aligned}$$

where $\mathcal{K}_{d'}$ denotes the modified Bessel function of the second kind with order d'.

References

[1] I. Gradshteyn and I. Ryzhik. *Table of Integrals, Series and Products*. New York: Academic, 6th ed. edition, 2000.

Appendix C

Derivations for Chapter 8

C.1 Derivation of MDL Based Model Order Estimator Without a Priori Knowledge on the Noise Level

For completeness the most important steps in deriving the standard MDL model order estimator as derived in [1] (assuming no knowledge of the noise variance) are given here.

The MDL criterion is defined as [1]

$$MDL = -log(f(\mathbf{y}_1, ..., \mathbf{y}_N|\Theta)) + \frac{1}{2} z \log N,$$

where $\mathbf{y}_1, ..., \mathbf{y}_N$ are N iid zero mean M-dimensional multivariate Gaussian observation vectors, Θ a parameter vector of the model under consideration and z the degree of freedom. Let Θ^Q be the parameter vector of the assumed model, i.e. $\Theta^Q = \left[a_1, ...a_Q, \sigma_D^2, C_1^H, ..., C_Q^H\right]$, where a_l, with $l \in \{1, ..., Q\}$ are the eigenvalues in the signal subspace, σ_D^2 is the noise variance and C_l, with $l \in \{1, ..., Q\}$ are the eigenvectors in the signal subspace. The joint probability density $f(\mathbf{y}_1, ..., \mathbf{y}_N|\Theta^Q)$ can then be written as

$$f(\mathbf{y}_1, ..., \mathbf{y}_N|\Theta^Q) = \prod_{i=1}^{N} \frac{1}{\pi^M \det \mathbf{C}^{(Q)}} \exp\left[-\mathbf{y}_i^H \mathbf{C}^{(Q)^{-1}} \mathbf{y}_i\right]. \tag{C.1}$$

The log likelihood of Eq. (C.1) is then given by

$$L(\Theta^{(Q)}) = -N \log[\det \mathbf{C}^{(Q)}] - N \mathrm{tr}\left[\mathbf{C}^{(Q)^{-1}} \hat{\mathbf{C}}\right], \tag{C.2}$$

where $\hat{\mathbf{C}}$ is the estimate of the correlation matrix, $\hat{\mathbf{C}} = U \begin{pmatrix} \Lambda_Q & 0 \\ 0 & \Lambda_{M-Q} \end{pmatrix} U^H$. $\mathbf{C}^{(Q)}$ is now substituted with ML estimates:

$$\mathbf{C}^{(Q)} = U \begin{pmatrix} \Lambda_Q & 0 \\ 0 & \hat{\sigma}_D^2 I_{M-Q} \end{pmatrix} U^H,$$

where $\Lambda_Q \in \mathbb{R}^{Q\times Q}$ is a diagonal matrix with the estimated eigenvalues $\hat{\lambda}_l$ with $l \in \{1, ..., Q\}$ of the assumed Q-dimensional signal subspace on the main diagonal. Further, U is a ML estimate of the eigenvector matrix and $\hat{\sigma}_D^2 = \frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l$ is the ML estimate of the noise under the assumed Q-dimensional signal subspace. That U, Λ_Q and $\hat{\sigma}_D^2 = \frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l$ are ML estimates of the eigenvector matrix, the signal subspace eigenvalues, and noise variance will be shown in Appendix C.3.

Using the relation

$$\det\hat{\mathbf{C}} = \left(\prod_{l=1}^{Q}\hat{\lambda}_l\right)\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right) \Leftrightarrow \left(\prod_{l=1}^{Q}\hat{\lambda}_l^{-1}\right) = \frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{\det\hat{\mathbf{C}}},$$

it can be shown that $L(\Theta^{(Q)})$ can be written as:

$$L(\Theta^{(Q)}) \equiv N\log\left[\frac{\left(\prod_{l=Q+1}^{M}\hat{\lambda}_l\right)}{(\frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l)^{(M-Q)}}\right]. \tag{C.3}$$

Eq. (C.3) agrees with the result in [1].

C.2 MDL Model Order Estimator with a Priori Knowledge on the Noise Level

When a priori information on the noise level is present $\mathbf{C}^{(Q)}$ in (C.1) is substituted with:

$$\mathbf{C}^{(Q)} = U\begin{pmatrix}\Lambda_Q & 0\\ 0 & \sigma_D^2 I_{M-Q}\end{pmatrix}U^H,$$

$L(\Theta^{(Q)})$ then becomes

$$\begin{aligned} L(\Theta^{(Q)}) &= \underbrace{-N\log\left[\det\begin{pmatrix}\Lambda_Q & 0\\ 0 & \sigma_D^2 I_{M-Q}\end{pmatrix}\right]}_{A} \\ &- \underbrace{N\mathrm{tr}\left[\begin{pmatrix}\Lambda_Q^{-1} & 0\\ 0 & \sigma_D^{-2} I_{M-Q}\end{pmatrix}\begin{pmatrix}\Lambda_Q & 0\\ 0 & \Lambda_{M-Q}\end{pmatrix}\right]}_{B}. \end{aligned}$$

part A

$$\begin{aligned} A &= -N\log\left[\det\begin{pmatrix}\Lambda_Q & 0\\ 0 & {\sigma_D}^2 I_{M-Q}\end{pmatrix}\right] \\ &= N\log\left[\frac{\left(\prod_{l=1}^{Q}\hat{\lambda}_l^{-1}\right)}{(\sigma_D^2)^{(M-Q)}}\right] \end{aligned} \tag{C.4}$$

using the relation: $\det \hat{\mathbf{C}} = \left(\prod_{l=1}^{Q} \hat{\lambda}_l\right)\left(\prod_{l=Q+1}^{M} \lambda_l\right) \Leftrightarrow \left(\prod_{l=1}^{Q} \hat{\lambda}_l^{-1}\right) = \frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{\det \hat{\mathbf{C}}}$

$$\begin{aligned} A &= N \log \left[\frac{\frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{\det \hat{\mathbf{C}}}}{(\sigma_D{}^2)^{(M-Q)}} \right] \\ &= N \log \left[\frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}} \right] - N \log \left[\det \hat{\mathbf{C}}\right] \end{aligned}$$

part B

$$\begin{aligned} B &= N \mathrm{tr} \left[\begin{pmatrix} \Lambda_Q^{-1} & 0 \\ 0 & \sigma_D^{-2} I_{M-Q} \end{pmatrix} \begin{pmatrix} \Lambda_Q & 0 \\ 0 & \Lambda_{M-Q} \end{pmatrix} \right] \\ &= N \mathrm{tr} \left[\begin{pmatrix} I_Q & 0 \\ 0 & \sigma_D^{-2} \Lambda_{M-Q} \end{pmatrix} \right] \\ &= N \left(Q + \sigma_D^{-2} \sum_{l=Q+1}^{M} \hat{\lambda}_l \right) \\ &= N \left(Q + \sigma_D^{-2} (M-Q) \hat{\sigma}_D^2 \right) \end{aligned}$$

Combining part A and B gives:

$$\begin{aligned} L(\Theta^{(Q)}) &= N \log \left[\frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}} \right] - N \log \left[\det \hat{\mathbf{C}}\right] - N \left(Q + (M-Q) \frac{\hat{\sigma}^2}{\sigma_D^2} \right) \\ &\equiv N \log \left[\frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}} \right] - N \left(Q + (M-Q) \frac{\hat{\sigma}_D^2}{\sigma_D^2} \right) \\ &= N \log \left[\frac{\left(\prod_{l=Q+1}^{M} \hat{\lambda}_l\right)}{(\sigma_D^2)^{(M-Q)}} \right] - N \left(Q + \frac{\sum_{l=Q+1}^{M} \hat{\lambda}_l}{\sigma_D^2} \right) \end{aligned}$$

where we left out the constant $N \log \left[\det \hat{\mathbf{C}}\right]$.

C.3 ML Estimates for MDL and Modified MDL Estimator

In this appendix we derive maximum likelihood estimates for the noise variance σ_D^2, the eigenvectors C_l and the eigenvalues a_l, for $l \in \{1, .., Q\}$.

The ML estimate $\hat{\sigma}_D^2 = \frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l$ can be derived by maximization of Eq. (C.2) with respect to $\hat{\sigma}_D^2$, that is

$$\max_{\hat{\sigma}_D^2} L(\Theta^{(Q)})$$

$$\frac{dL(\Theta^{(Q)})}{d\hat{\sigma}_D^2} = -N\frac{(M-Q)}{\hat{\sigma}_D^2} + N\left(\frac{1}{\hat{\sigma}_D^2}\right)^2 \sum_{l=Q+1}^{M}\hat{\lambda}_l = 0,$$

which leads when solving for $\hat{\sigma}_D^2$ to $\hat{\sigma}_D^2 = \frac{1}{M-Q}\sum_{l=Q+1}^{M}\hat{\lambda}_l$

ML estimates of the eigenvectors and signal subspace eigenvalues of $\mathbf{C}^{(Q)}$ can be derived by considering the eigenvalue decomposition of $\mathbf{C}^{(Q)}$

$$\mathbf{C}^{(Q)} = \mathbf{C}\begin{pmatrix} A_Q & 0 \\ 0 & A_{M-Q} \end{pmatrix}\mathbf{C}^H,$$

When we use a priori information on the noise level we can write $\mathbf{A}_{M-Q} = \sigma^2\mathbf{I}_{M-Q}$. To find ML estimates of the eigenvectors $\mathbf{C}$ we consider the log-likelihood of Eq. (C.1), i.e.

$$\begin{aligned} L(\Theta^{(Q)}) &= -N\log[\det\mathbf{C}^{(Q)}] - N\mathrm{tr}\left[{\mathbf{C}^{(Q)}}^{-1}\hat{\mathbf{C}}\right] \\ &= -N\log\left[\prod_{l=1}^{M} a_l\right] - N\mathrm{tr}\left[\mathbf{C}\mathbf{A}^{-1}\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\right] \\ &= -N\log\left[\prod_{l=1}^{M} a_l\right] - N\mathrm{tr}\left[\mathbf{A}^{-1}\underbrace{\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}}_{\mathbf{G}}\right] \end{aligned} \tag{C.5}$$

Let $\mathbf{G} = \mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}$. The matrix $\mathbf{C}^H\mathbf{U}$ is now an orthogonal matrix and $\mathbf{C}^H\mathbf{U}\mathbf{\Lambda}\mathbf{U}^H\mathbf{C}$ the eigenvalue decomposition of $\mathbf{G}$. Now we can write

$$\begin{aligned} \mathrm{tr}\left[\mathbf{A}^{-1}\mathbf{G}\right] &= \mathrm{tr}\left[\begin{pmatrix} a_1^{-1} & 0 & \cdots & 0 \\ 0 & a_2^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_M^{-1} \end{pmatrix}\begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1M} \\ g_{21} & g_{22} & & \\ \vdots & & \ddots & \\ g_{M1} & & & g_{1MM} \end{pmatrix}\right] \\ &= \sum_{l=1}^{M} a_l^{-1}g_{ll} \end{aligned}$$

and

$$L(\Theta^{(Q)}) = -N\sum_{l=1}^{M}\log[a_l] - N\sum_{l=1}^{M}a_l^{-1}g_{ll}$$

ML estimates of a_l are then found as

$$\frac{\partial L}{\partial a_l} = -Na_l^{-1} + Na_l^{-2}g_{ll},$$

which leads to $a_l = g_{ll}$

Inserting this in L leads to

$$L = -N \log \left[\prod_{l=1}^{M} g_{ll} \right] - NM$$

To maximize L we need to minimize $\prod_{l=1}^{M} g_{ll}$. To find this minimum we use *Hadamards* inequality:

$$\det \mathbf{G} \leq \prod_{l=1}^{M} g_{ll},$$

with equality if and only if $\mathbf{G}$ is diagonal and $\mathbf{G}$ should be positive definite. We know that $\mathbf{G} = \mathbf{C}^H \mathbf{U} \mathbf{\Lambda} \mathbf{U}^H \mathbf{C}$. The orthogonal matrix $\mathbf{C}$ does not influence the determinant of G. Therefore we can choose $\mathbf{C} = \mathbf{U}$ such that *Hadamards* inequality leads to equality.

Let us now use the fact that $\mathbf{U}$ are ML estimates of $\mathbf{C}$. ML estimates of the eigenvalues of $\mathbf{C}^{(Q)}$ can then be computed by taking partial derivatives of Eq.(C.5), i.e.

$$\frac{\partial L(\Theta^{(Q)}}{\partial a_j} = -N \frac{1}{a_j} + N a_j^{-2} \hat{\lambda}_j = 0,$$

so that

$$a_j = \hat{\lambda}_j.$$

References

[1] M. Wax and T. Kailath. Detection of signals by information theoretic criteria. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-33:387–392, 1985.

List of Symbols

General

A	magnitude of clean speech DFT coefficient
d	noise DFT coefficient (realization)
D	noise DFT coefficient (random variable)
$\mathbf{D}$	vector of noise DFT coefficients D
DD	decision-directed
DFT	discrete Fourier transform
E	expectation operator
$f_Z(z)$	probability density function of a random variable Z
G	speech enhancement gain function
i	time frame index
$\mathcal{I}_d$	modified Bessel function of the first kind and order d
k	frequency bin index
K	frame size
P	frame shift
pdf	probability density function
PSD	power spectral density
R	magnitude of noisy speech DFT coefficient
SNR	signal-to-noise ratio
SNR_{seg}	segmental SNR
x	clean speech DFT coefficient (realization)
$\hat{x}$	estimate of clean speech DFT coefficient
x_t	time domain sample (realization)
$\mathbf{x}_t$	vector of time domain samples x_t
X	clean speech DFT coefficient (random variable)
X_t	time domain sample (random variable)
$X_{\Re}$	real part of a DFT coefficient X
$X_{\Im}$	imaginary part of a DFT coefficient X
$\mathbf{X}$	vector of clean speech DFT coefficients X
y	noisy speech DFT coefficient (realization)
Y	noisy speech DFT coefficient (random variable)
$Y_{\Re}$	real part of a DFT coefficient Y
$Y_{\Im}$	imaginary part of a DFT coefficient Y
$\mathbf{Y}$	vector of noisy DFT coefficients Y

α	smoothing factor
ζ	*a posteriori* SNR
Θ	phase of noisy speech DFT coefficient
λ_{th}	threshold
ξ	*a priori* SNR
$\hat{\xi}$	estimate of *a priori* SNR
σ_D^2	noise power spectral density
$\hat{\sigma}_D^2$	estimate of noise power spectral density
$\sigma_{D_\Re}^2$	variance of the real part of noise DFT coefficient D
σ_X^2	clean speech power spectral density
$\hat{\sigma}_X^2$	estimate of clean speech power spectral density
σ_Y^2	noisy speech power spectral density
$\hat{\sigma}_Y^2$	estimate of noisy speech power spectral density
$\hat{\sigma}_{Y,A}^2$	Estimate of σ_Y^2 using adaptive time segmentation
$\hat{\sigma}_{Y,B}^2$	Bartlett estimate of σ_Y^2
$\hat{\sigma}_{Y,P}^2$	periodogram estimate of σ_Y^2
Φ	phase of clean speech DFT coefficient
$\Re$	indicates a real part
$\Im$	indicates an imaginary part

Chapter 3

$C[0]$	correlation coefficient with lag zero
$\hat{c}[0]$	estimate of correlation coefficient with lag zero
$\hat{C}[0]$	estimator of correlation coefficient with lag zero
DDA	decision-directed approach with adaptive time segmentation
LRT	likelihood ratio test

Chapter 4

BDD	backward decision-directed
FDD	forward decision-directed
IFBDD	iterative forward-backward decision-directed
$\hat{\xi}_F$	ξ estimated with forward decision-directed approach
$\hat{\xi}_B$	ξ estimated with backward decision-directed approach

Chapter 5

am	speech absence
dm	speech is generated with a deterministic model
d_q	exponential decay factor of component q
M	random variable that indicates the type of speech model
sm	speech is generated with a stochastic model
SD	stochastic-deterministic
ν_q	frequency of component q

Chapter 6

$\hat{A}^{(1)}$	estimate of A based on generalized Gamma pdf with $\gamma = 1$
$\hat{A}^{(1)}_{\gg}$	$\hat{A}^{(1)}$ based on large argument approximation
$\hat{A}^{(1)}_{\ll,N}$	$\hat{A}^{(1)}$ based on small argument approximation
$\hat{A}^{(1)}_{C,N}$	$\hat{A}^{(1)}$ based on combined small and large argument approximation
$\hat{A}^{(2)}$	estimate of A based on generalized Gamma pdf with $\gamma = 2$
$\mathcal{D}_\nu$	parabolic cylinder function
${}_1\mathcal{F}_1(a;b;x)$	confluent hypergeometric function
$\zeta_\Re$	a *posteriori* SNR with respect to the real part of ζ

Chapter 7

MNIG	multivariate normal inverse Gaussian
$\mathcal{K}_d$	modified Bessel function of the second kind
RIG	Rayleigh inverse Gaussian
α	shape parameter of MNIG distribution
δ	scale parameter of MNIG distribution
$\mathbf{\Gamma}$	correlation matrix
Λ_X	scaling random variable
$\boldsymbol{\lambda}_D$	correlation matrix of vector $\mathbf{D}$

Chapter 8

$B(Q)$	signal subspace dimension dependent bias compensation
$\mathbf{C}$	correlation matrix
$\hat{\mathbf{C}}$	estimated correlation matrix
$C_D(k,i;p)$	correlation between $D(k,i+p)$ and $D(k,i)$ with frame lag p
M	dimension of the correlation matrix
MS	minimum statistics
Q	dimension of the signal subspace
tr	trace
$\mathbf{U}$	eigenvector matrix of correlation matrix $\mathbf{C}$
VAD	voice activity detector
λ	eigenvalue
$\hat{\lambda}$	eigenvalue based on an estimated correlation matrix
$\mathbf{\Lambda}$	Eigenvalue matrix of correlation matrix $\mathbf{C}$

Samenvatting

Het belang van het vakgebied van *speech enhancement*, of ook wel spraakverbetering genoemd, komt voort uit het groeiende gebruik van digitale spraakverwerkende applicaties zoals mobiele telefonie, digitale gehoorapparaten en verschillende mens-machine communicatiesystemen. De trend dat deze applicaties meer en meer mobiel worden gemaakt, vergroot de variëteit van potentiële storingsbronnen. *Speech enhancement* methoden kunnen worden gebruikt om de kwaliteit van deze spraakverwerkende applicaties te verhogen en robuuster te maken voor het gebruik onder ruizige condities.

De naam *speech enhancement* refereert naar een grote groep van methoden die allemaal tot doel hebben om bepaalde kwaliteitsaspecten van deze apparaten te verhogen. Voorbeelden daarvan zijn echoreductie in spraaksignalen, het kunstmatig verbreden van de bandbreedte van spraaksignalen, *packet loss concealment* en (additieve) ruisonderdrukking. In dit proefschrift richten we ons op additieve ruisonderdrukking met behulp van één microfoon. In het bijzonder richten we ons op methoden die werken in het discrete Fourier Transformatie (DFT) domein. Het hoofddoel van het gepresenteerde onderzoek is om bestaande methoden, welke gebaseerd zijn op het gebruik van slechts één microfoon, te verbeteren voor een uitgebreide range aan ruissoorten en ruisniveaus.

Het gepresenteerde onderzoek richt zich op drie verschillende onderwerpen. Allereerst onderzoeken we hoe een betere schatting van de *a priori* signal-to-noise ratio (SNR) uit ruizige spraak verkregen kan worden. Een goede schatting van de *a priori* SNR is van cruciaal belang voor *speech enhancement*, omdat veel *speech enhancement* schatters van deze parameter afhankelijk zijn. We richten ons op twee verschillende aspecten van het schatten van de *a priori* SNR. Als eerste presenteren we een adaptief tijd segmentatie algoritme, wat vervolgens gebruikt wordt om de variantie van de geschatte *a priori* SNR te verkleinen. Ten tweede beschrijven we een methode om de bias van de *a priori* SNR te verkleinen. Deze bias is vaak aanwezig tijdens overgangen tussen verschillende spraakklanken en overgangen van ruizige spraak naar ruis en vice versa. Het gebruik van deze verbeterde schatters voor de *a priori* SNR leidt tot objectieve en subjectieve verbeteringen van de kwaliteit.

Ten tweede onderzoeken we schatters voor schone spraak DFT coëfficiënten onder modellen waarbij rekening gehouden wordt met de eigenschappen van spraak. Dit onderwerp wordt van twee verschillende kanten benaderd. Ten eerste beschouwen we de afleiding van schone spraakschatters onder het gebruik van een gecombineerd

stochastisch/deterministisch model voor de complexe spraak DFT coëfficiënten. Het gebruik van dit model is gebaseerd op het feit dat bepaalde spraakklanken een meer deterministisch karakter hebben. Daarnaast beschouwen we schatters voor de complexe schone spraak DFT coëfficiënten en de magnitude van de DFT coëfficiënten onder de aanname dat deze super-Gaussisch verdeeld zijn. Deze aanname is gebaseerd op gemeten histogrammen van schone spraak DFT coëfficiënten. Onder super-Gaussische verdelingen beschrijven we twee verschillende schatters. *Minimum mean-square error* (MMSE) schatters worden afgeleid onder de aanname dat de spraak DFT coëfficiënten en spraak DFT magnitudes een gegeneraliseerde Gamma verdeling hebben. Maximum *a posteriori* (MAP) schatters worden afgeleid onder de aanname dat de spraak DFT coëfficiënten verdeeld zijn volgens een multivariate normale inverse Gaussische (MNIG) verdeling. Volgens objectieve experimenten is de performance van de schatters afgeleid onder de MNIG distributie iets beter dan de performance van de schatters afgeleid onder de gegeneraliseerde Gamma verdeling. Bovendien hebben de schatters afgeleid onder de MNIG verdeling een aantal theoretische voordelen boven de schatters afgeleid onder de gegeneraliseerde Gamma verdeling. Namelijk, de statistische modellen in het complexe DFT domein en in het polar domein zijn consistent onder de MNIG verdeling, wat niet het geval is voor schatters afgeleid onder de gegeneraliseerde Gamma verdeling. Verder is het met de MNIG verdeling ook mogelijk om vector processen te modelleren. Dit maakt het mogelijk om de afhankelijkheid tussen reëel en imaginair deel van de DFT coëfficiënten mee te nemen.

Als laatste presenteren we een methode voor het schatten van het vermogensdichtheidspectrum van de ruis. Het belang voor *speech enhancement* van een goede schatting van dit vermogensdichtheidspectrum volgt uit het feit dat alle schone spraak DFT-domein gebaseerde schatters hiervan afhankelijk zijn. De ontwikkelde methode is gebaseerd op de eigenwaarde decompositie van correlatie matrices die opgebouwd zijn uit een tijdserie van ruizige DFT coëfficiënten. De gepresenteerde methode maakt het mogelijk om, in tegenstelling tot de meeste bestaande methoden, het vermogensdichtheidspectrum van de ruis te schatten zelfs wanneer spraak continu aanwezig is. Verder is de vertraging in het schatten van een veranderend ruisspectrum aanzienlijk verkort in vergelijking tot bestaande *state-of-the-art* algoritmes.

Een aantal van de contributies in dit proefschrift kunnen worden gecombineerd tot een compleet *speech enhancement* systeem. Een vergelijking is gemaakt op basis van een luistertest tussen een systeem gebaseerd op bijdragen uit dit proefschrift en een *state-of-the-art speech enhancement* systeem. Hieruit komt naar voren dat het systeem gebaseerd op bijdragen uit dit proefschrift tot een significant betere kwaliteit leidt.

Acknowledgements

It is remarkable to see how someone's world can change within four years. Changes due to positive events and progress in work and private life, however, as well as due to unforeseeable sad moments. Now these four years that I worked towards my Ph.D. degree are over, time has come to thank those people that contributed to this work and my well being over the last years.

First of all I would like to thank my supervisors Jesper Jensen and Richard Heusdens for their great support and supervision. I very much enjoyed our discussions, the way we worked together, and of course our social events. Hopefully we can in some way continue our collaboration in the near future.

Also I would like to thank Jan Biemond for being my promotor. Your advice and comments on my thesis were very useful to me.

Jan Erkelens, thanks for working together, and for sharing your knowledge and your criticism with me as well as for sharing our office.

Much gratitude I owe to Rainer Martin, for giving me the possibility to work as a guest researcher at the Ruhr-University in Bochum. I felt very welcome in Bochum and enjoyed my stay very much. That brings me to thank Timo Gerkmann. I really enjoyed sharing our office and all the discussions we had during these three months. I certainly learnt to appreciate the beautiful city of Bochum with its own Königsallee and of course had fun in visiting the VFL.

I enjoyed being part of the ICT group during the last four years and would like to thank all my colleagues in this group for sharing this time with me. In particular I would like to say thanks to the members of the audio group; Pim Korten, Omar Niamut, Jan Østergaard, Ivo Batina and Ivo Shterev. Much gratitude I owe as well to the supporting staff; Anja van den Berg, Robbert Eggermont and Ben van den Boom.

Although work and research always gave me a lot of fun, satisfaction and sometimes even entertainment, it is also good to relax once in a while and worry about less serious things. Therefore, I would like to thank all my friends. In particular, my special thanks go towards Jos Eijkelestam for being the perfect organizer and initiator of the kart racing competition, board game evenings and so much more. Jos, your effort in this is very much appreciated.

My gratitude goes towards Ton and Lia Kleinbloesem for their great support to me and my family when my father passed away so suddenly.

Writing this thesis would have been much more difficult, if not impossible, without the support that I got from my family.

Jan and Coby, thanks for showing interest in my work and being ready to help whenever needed. Eric, thanks for your assistance in so many things over the last years. It is enjoyable to see how passionate both you and Joyce are when it is about winning a game. Joyce and Esther, thanks for being my sisters.

Mom, I still admire the way you can persist in reaching your goals. Certainly there is something to learn for me. Thanks for your support and guidance throughout all these years.

Above all, I would have liked to be able to thank my father here for his inspiration, support and for the times we were able to share together. Unfortunately, it has been made impossible by the temporality of life.

Rosalie, life during the last few years was sometimes not as easy as it seemed from the outside. Thanks for supporting me and my family during these moments, but of course also thanks for the many joyful moments and your unconditional love.

Curriculum Vitae

Richard Christian Hendriks was born in Schiedam, the Netherlands, on June 21st, 1980. He obtained his VWO-diploma from Scholengemeenschap Spieringshoek in Schiedam in 1998, after which he started his study Electrical Engineering at Delft University of Technology. In 2002 he worked during an internship of three months in the sound and image processing group at the Royal Institute of Technology, in Stockholm, Sweden. In 2003 he obtained his M.Sc. degree cum laude after doing his graduation work in the Information and Communication Theory group at Delft University of Technology on the topic *residual coding with LPC in the perceptual domain*.

In 2003 he started his Ph.D. study in the Information and Communication Theory group at Delft University of Technology and worked on the project *single-microphone enhancement of noisy speech signals*, which was funded by Philips Research and STW. During his Ph.D. study he was involved as a lecturer in several courses in the field of digital signal processing, and audio and speech processing. In 2005 he worked as a guest researcher at the Institute of Communication Acoustics at Ruhr-University Bochum, Germany.

Since 2007 he works as a postdoctoral researcher in the Information and Communication Theory group at Delft University of Technology. Currently he is involved in the project *intelligibility enhancement of noisy speech*, which is in cooperation with Oticon.