Speech Dereverberation in the Time-Frequency Domain

Anna Oyzerman

Speech Dereverberation in the Time-Frequency Domain

Research Thesis

As Partial Fulfillment of the Requirements for
the Degree Master of Science in Electrical Engineering

Anna Oyzerman

Submitted to the Senate of the Technion—Israel Institute of Technology

Heshvan 5772 Haifa October 2012

Acknowledgement

I would like to express my gratitude to my research advisor Prof. Israel Cohen. His supervision, guidance and support helped me greatly throughout my research.

The Generous Financial Help Of The Technion Is Gratefully Acknowledged.
This research was supported by the Israel Science Foundation (grant no. 1130/11).

Contents

List of Figures

List of papers

- Oyzerman, A.; Cohen, I., "System Identification and Dereverberation of Speech Signals in the Single-Side-Band Transform Domain". Appeared in Proc. 20th European signal processing conference (EUSIPCO-2012), Bucharest, Romania, August 2012.

- Oyzerman, A.; Cohen, I., "Reverberant component estimation in the STFT domain in comparison to the statistical method", in preparation.

Abstract

In this thesis, we address the problem of dereverberation of speech signals, acquired in an enclosed space by a single microphone. We spectrally estimate and suppress the dereverberation in the time-frequency domain, using two representations: The short-time Fourier transform (STFT), and the Single-Side-Band transform (SSB).

In the first part of this thesis, the problems of system identification and dereverberation are addressed in the SSB domain. We derive an analytical relation between the input and output signals in the SSB domain, and formulate a system identification routine for a band-to-band approximation of that relation. The dereverberation problem is addressed using a statistical model for the acoustic impulse response (AIR) function. We present exact and approximate representations of the AIR and the reverberant signal in the SSB domain. The performance of the dereverberation algorithm is evaluated as a function of the representation complexity. The SSB and the STFT representations are compared; it is found that the STFT representation is more suitable for the dereverberation application.

In the second part of the thesis, we propose a new algorithm for the estimation of the reverberant component, which is a critical part of dereverberation in the spectral domain. In order to address the frequency variation of the reflection coefficients in the room, we formulate the problem directly in the STFT domain.

First, a system identification stage is performed. A white noise signal is played in the reverberant environment, and a filter is extracted relating the recorded output signal to its reverberant component. In order to reduce the complexity of the computation, we propose approximate representations of the filter, that use only some or none of the crossband filters. The latter case is referred to as a "band-to-band filter".

The performance of the system identification stage is analysed in terms of the mean-square error (MSE) between the actual reverberant component and its estimate by the

approximate representations. It is shown that the smallest MSE is achieved by using the band-to-band filter, due to the fact that the expression that relates the output signal to its reverberant component is most accurate in that case. Also, it was found that for a small number of crossbands, it is advantageous to increase the lengths of the filters in order to improve the estimate of the reverberant component.

In the last stage, we use the reverberant component estimate in a spectral enhancement algorithm for dereverberation. We measure the performance for different reverberant environments and for various distances between the source and the microphone. We show that our method achieves better results than an existing statistical method.

Notation

$x * y$	linear convolution
$\mathbf{x}^T$	vector transpose
$\|x\|$	absolute value of x
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
$Re\{\cdot\}$	real component of a complex number
$Im\{\cdot\}$	imaginary component of a complex number
$E\{\cdot\}$	mathematical expectation
T_{60}	Reverberation time
f_s	Sampling frequency
σ^2	variance od identically distributed Gaussian random noise
δ	damping constant (related to the reverberation time)
$x(n)$	time-domain signal
$h(n)$	impulse response function in time domain
Q	length of the impulse response function h
λ_x	spectral variance of signal x
r_{xx}	auto-correlation of signal x
ϵ	mean squared error
σ^2	variance of the identically distributed Gaussian random noise
δ	damping constant (related to the reverberation time)
$\tilde{\psi}$	analysis window
ψ	synthesis window
Y_d	direct path
Y_e	early reverberant path
Y_l	late reverberant path
Y_r	reverberant path

$X_{m,k}$	SSB domain signal
K	number of frequency bands in the SSB transform
$H_{m,m',k,k'}$	crossband filter in the SSB domain
k'	crossband
m'	crosstime sample
$H^{\mathbf{bb}}_{m,m',k}$	band-to-band filter in the SSB domain
$\mathbf{H}^{\mathbf{bb}}_{m,k}$	vector of band-to-band filter for time sample m and frequency band k.
$\mathbf{H}^{\mathbf{bb}}_{\mathbf{k}}$	a column-stack concatenation of the band-to-band filters at frequency band k
N_x	number of crosstime samples in cross band filter $H_{m,m',k,k'}$
N_{xh}	number of time samples in filter $H_{m,m',k,k'}$
$X(l,k)$	STFT domain signal
N	number of frequency bands in the STFT transform
$H(l,k,k')$	crossband filter in the STFT domain
$H^{\mathbf{bb}}(l,k,k)$	band-to-band filter in the STFT domain
$H^{\mathbf{2K}}(l,k,k')$	$2K$-crossbands' filter in the STFT domain
$\tilde{H}_r^{2K}(l,k)$	$2K$-crossbands' filter relating the output signal to its reverberant component
$\hat{H}(k,l)$	estimated filter
N_h	length of the crossband filter $H(l,k,k')$
$Y_r^{\mathbf{bb}}(l,k)$	approximation of Y_r using only the band-to-band filter
$Y^{\mathbf{sp}}(l,k)$	recorded speech signal

Abbreviations

AIR	Acoustic Impulse Response
DRR	Direct to Reverberation Ratio
DTFT	Discrete-Time Fourier transform
GSM	Gaussian Source Model
LP	Linear Prediction
LSA	Log-Spectral Amplitude
LS	Least Squares
LSD	Log-Spectral Distortion
LRSV	Late Spectral Reverberant Component
ML	Maximum Likelihood
MSE	Mean Squared Error
PSD	Power Spectral Density
LTI	Linear Time-Invariant
OM-LSA	Optimally-Modified Log Spectral Amplitude
RIR	Room Impulse Response
SIR	Signal to interference ratio
SNR	Signal to Noise Ratio
STFT	Short Time Fourier Transform
SSB	Single Side Band
SRR	Signal to Reverberation Ratio
RT	Reverberation Time

Chapter 1

Introduction

1.1 Background

1.1.1 Reverberation in Enclosed Spaces

When a speech signal is acquired in an enclosed space, the sound wave, created by the signal, hits the surrounding walls and other objects. Therefore the observed signal consists of a superposition of many delayed and attenuated copies of the speech signal. These multiple reflections can number several thousands and give rise to the effect known as reverberation.

The resulting signal consisted of direct path, and of multiple attenuated reflections, arriving in different delays with regards to the direct signal, which deteriorate the intelligibility of the speech, and reduce the performance of various signal processing applications. The deleterious effects increase as the distance between the talker and the microphones is increased. Thus, an efficient techniques are required, in order to reduce the reverberations effect, or by other words, to estimate the original direct path of the signal, without the reflections.

The acoustic channel between a source and a microphone can be described by an Acoustic Impulse Response (AIR). This is the signal that is measured at the microphone in response to a source that produces a "delta impulse" of sound. The AIR can be divided into three segments, the direct path, early reflections, and late reflections, whose convolution with the desired signal results in the direct sound, early reverberation, and

late reverberation, respectively. The early reflections appear as separate delayed impulses in the AIR, while late reflections appear as a continuum exponential decay. This decay is a well-known property of the AIR, which has motivated the notion of another highly important concept- the reverberation time. It is defined as the time that is necessary to reach a 60 dB decay of the sound energy after switching off a sound source, and denoted by T_{60}. Intuitively, it quantifies the severity of reverberation within a room, and is affected by the volume of the enclosed space and the acoustic properties of the reflecting surfaces [1].

1.1.2 Previous works

Reverberation reduction processes can be divided into many categories. We categorized the reverberation reduction processes depending on whether or not the AIR needs to be estimated. We then obtain two main categories, viz. reverberation cancellation and reverberation suppression.

Reverberation cancellation

This category consists of methods known as blind deconvolution. The acoustic impulse responses are identified blindly, using only the observed microphone signals, and then used to design an inverse filter that compensates for the effect of the acoustic channels. The first stage of this method is called blind system identification, and usually applied on a multi-channel system, where the cross-correlation between each two observations is computed to form a cross-correlation matrix, whose solution corresponds to the AIRs of the different channels. To achieve a solution of such system of equations, two requirements must be satisfied [2]:

1. The unknown channels must not include common zeros.

2. The correlation matrix of the source signal must be full rank.

An approach to solve this problem was proposed by Xu et al. [2], via Least Squares (LS) method. Another possible method is the eigen decomposition, proposed by Gurelli and Nikias [3], where the algorithm estimates the orders and root locations of the channel transfer functions. The input signal is allowed to be non-stationary and the channel

transfer functions may be non-minimum phase, as well as non-causal. Using multichannel LMS and Newton adaptive filters both in the time and frequency domains, were proposed by Huang and Benesty [4, 5].

Other approaches in this category include the usage of the cepstrum for blind system identification between two channels [6]. It is shown that the channels can be reconstructed from their phases using an iterative approach, where the phases are identified from the cepstra of the observed data [6, 7]. A method introduced by Triki and Slock [8] comprises multichannel Linear Prediction (LP) to whiten the input signal and subsequent multichannel linear prediction which is used to identify the channels. Finally, Hopgood and Rayner [9] proposed to use an autoregressive model of channel impulse response, which is assumed to be stationary, in contrast to the FIR model employed in all the above methods. The parameters of the all-pole channel filter can be identified by observing several frames of the input signal. Further extensions based on this idea have been developed by Evers and Hopgood [10], and Evers, Hopgood, and Bell [11].

The second stage of the method is called inverse filtering. After determining the AIRs from blind system identification algorithm, dereverberation can be achieved by an inverse system. However, direct inversion of an acoustic channel is problematic due to number of reasons. First, AIRs long durations (typically thousands of coefficients) require high computational efforts. Second, acoustic channels typically exhibit non-minimum phase characteristics [12]. Finally, acoustic channels may contain spectral nulls, which cause noise amplification.

Several approaches have been studied for single channel inversion. For example, single channel LS inverse filters [13], or inverse filtering where the impulse response is decomposed into a minimum-phase component and an all-pass component, and equalized separately [14]. Vary [15] proposed to filter the signal in the time domain, while the filter coefficients are calculated in the STFT domain. This exhibits smaller values of signal delay and complexity. In the multichannel case, an exact inverse can be found by application of multichannel least squares design [16, 17]. Adaptive versions have also been considered by Nelson, Orduna-Bustamante and Hamada [18]. If there are no common zeros between the two channel transfer functions, exact inverse filtering can be performed, with inverse filters of length similar to the channel length [16, 17].

While good results can be achieved, the methods in this category suffer from several limitations: They usually require a multi-channel system, (2) channels cannot be identified uniquely when they contain common zeros, (3) observation noise causes severe problems, and (4) some methods require knowledge of the order of the unknown system [19].

Reverberation suppression

Methods in the second category do not require an estimate of the AIR and explicitly exploit the characteristics of speech, the effect of reverberation on speech, or the characteristics of the AIR. The first group of methods in this category is based on spatial processing, or so called, the beamforming. These methods often use a priori knowledge regarding the direction of arrival of the desired source. The microphone signals can be processed to enhance or attenuate signals emanating from particular directions. The most direct and straightforward technique in this category, is the Delay-and-sum Beamformer (DSB), in which the microphone signals are delayed, to compensate for different times of arrival, and then weighted and summed [20]. Examples for other possible designs are beamformers with three of four subarrays [21, 22], adaptive beamformers or subbands beamformers, where the signals are co-phased in each frequency band separately [22–24]. The drawback of the beamformer approach is that it works best for strongly localized sources and is less effective when the sound field is diffuse, as in the reverberation case. To address this problem three-dimensional array can be employed, where additional beams are steered in the direction of the strong initial reflections [25, 26]. Another approach is the matched filter beamformer, where the microphone signals are deconvolved with the room impulse responses [25, 27–30]. However, since diffuse reverberant sound comes from all directions in a room [1], it will always enter the look-direction of the beam and hence will be only partially suppressed.

Another method in this category is the Linear Prediction (LP) Residual method, introduced by Gaubitch, Naylor and Ward [31] and by Yegnanarayana and Satyanarayana [32]. This approach exploits the fact that the peaks in the LP residual signal correspond to excitation events in voiced speech, together with additional random peaks due to reverberation. It is assumed that the effect of reverberation on the Autoregressive (AR) coefficients is insignificant. Griebel and Brandstein [33] employ coarse room impulse response

estimates and apply a matched filter to obtain weighting functions for the reverberant residuals. Yegnanarayana *et al.* [34] use multichannel Hilbert envelopes to represent the strength of the peaks in the prediction residuals, and then sum it, to form a weight vector, which is applied to the prediction residual of one of the microphones. Yegnanarayana and Murthy [32] derive a weighting function based on the signal-to-reverberant ratio in different regions of the prediction residual. The approach presented by Gillespie, Malvar and Florencio [35] uses the kurtosis of the residual as a reverberation metric. This method was extended by adding a spectral subtraction stage to further suppress the remaining reverberation [36].

Nakatani *et al.* [37] formulated speech dereverberation as a maximum likelihood (ML) problem with multi-channel LP, using the Gaussian Source Model (GSM) to describe the autocorrelation matrix of the output signals. Later on, the ML approach was extended to the STFT domain, employing the crossband effect compensation [38].

Although these methods do attenuate the reverberation, they also significantly reduce naturalness in the dereverberated speech. A proposed solution for this problem can be using a spatio-temporal averaging approach, where the speech signals are first spatially averaged and the prediction residual is further enhanced using temporal averaging of neighbouring larynx cycles [39–41].

Recently, spectral enhancement methods have been used for speech dereverberation. Spectral enhancement of noisy speech is often formulated as estimation of speech spectral components from a speech signal degraded by statistically independent additive noise. One of the earlier methods in this category is spectral subtraction [42], where the effect of overlap-masking is reduced by subtracting reverberation from the recorded signal [43]. The method estimates the short-term Power Spectral Density (PSD) of late reverberation based on a statistical reverberation model. This model exploits the fact that the envelope of the AIR decays exponentially and depends on a single parameter that is related to the reverberation time of the room. The estimation of the late reverberation spectral variance, in this method, was derived under the assumption that the source-microphone distance is larger than the critical distance. When the source-microphone distance is smaller than the critical distance the LRSV estimator overestimates the real late component. For that reason, the described model was modified by Habets [44], to produce a so called

generalized model, that takes into account the energy contribution of the direct path. Based on this model, the author propose a new LRSV estimator, and then performs the spectral post filtering stage, via OM-LSA method [45].

Löllmann, and Vary [46] employ the spectral method for dereverberation in hearing aids. The proposed system derives a low delay filter-bank, and then performs time-domain filtering with coefficients adapted in the uniform or non-uniform non-uniform (Bark-scaled) frequency-domain. Jeub *et al.* [47], in addition to using the statistical model, use a dual-channel Wiener filter, in order to imitate binaural cues of the human auditory system.

An example of combining several methods for reverberation suppression is presented by Habets, Gaubitch and Naylor [48]. First, late reverberation is suppressed using a statistical reverberation model. Secondly, early reverberation and residual late reverberation are suppressed using a linear prediction (LP) residual processing technique. Another example of such combination is presented by Krishnamoorthy, Prasanna, and Mahadeva [49] and by Wu and Wang [36], where after the spectral subtraction step, the processed speech is subjected to identification and enhancement of high signal-to-reverberation ratio (SRR) regions in the temporal domain.

Although enhancement approaches to dereverberation do not assume explicit knowledge of the room impulse response, blind identification of other features is often required. Nevertheless, many of these methods are computationally efficient and suitable for real-time implementation.

1.2 Motivation and Goals

Currently, the statistical methods for dereverberation rely on an AIR model formulated in the time domain. Therefore, the distorting component of the recorded signal ,which is usually the late reverberant component, is also derived in the time domain, and only then transformed to the STFT domain for the spectral enhancement step. The disadvantage of this routine is that the transition from the time domain to the STFT domain does not take into account the crossband filters, as required for accurate representation [50]. Moreover, the time domain formulation doesn't allow one to incorporate the frequency

dependency of the reverberation phenomena [51,52]. Our goal in this thesis is to represent the AIR and the reverberant signal directly in the time-frequency domain, and based on that, to derive an estimator for the reverberant component of the signal. This formulation will address both accurate and simplified representations, and will analyse the effect of the approximations on the algorithm performance.

Another goal of this work is to incorporate time-frequency domain system identification into the dereverberation algorithm, which has been shown to be a useful tool for dereverberation application [16, 36], as it enables one to characterize the distorting element and later on to suppress it. Here we perform the identification routine directly in the time-frequency domain, in order to estimate the time-frequency domain filter that causes the reverberant effect, and then find the representation of the reverberant component. In contrary to the methods that use a specific AIR model, our approach does not require any prior knowledge regarding systems' parameters, and is not confined to using any specific statistical model.

Finally, we are testing the possibility of incorporating the Single-Side-Band (SSB) transform in the dereverberation application. The SSB transform is a real-valued time-frequency representation, and therefore it is often preferred in applications involving communications, coding systems, and speech processing applications. Hence, we formulate the dereverberation and identification problems in the SSB domain, and compare their performances to those in the STFT representation.

1.3 Overview of the Thesis

In this thesis, we address the dereverberation problem, using two time-frequency representations: the Single-Side-Band (SSB) representation and the Short-Time Fourier transform (STFT) representation.

For the SSB domain representation we first represent the AIR of the enclosure as an LTI system, and develop an analytic relation between the input and the output signals in the SSB domain. Then, a system identification routine is formulated for a band-to-band approximation of that relation. Thus, given an enclosed space with an unknown AIR, one can compute its approximate version by the identification procedure, and use it later in

order to perform dereverberation by one of the methods mentioned above.

Next, we incorporate the SSB representation into the dereverberation routine. We compute the late spectral variance estimation using the exact and approximate representations of the AIR and the reverberant signal, and perform post filtering using the OM-LSA spectral enhancement method [45]. Our experimental results show that the band-to-band approximation gives sufficient performance, and that each additional cross-band can contribute to a small improvement. Finally we compare the dereverberation performance using the SSB transform to that of the STFT. We find out the STFT performance is higher, apparently due to the fact that the SSB transform combines the phase information into the amplitude representation.

The second time-frequency representation that is addressed in this thesis, is the STFT representation. We present a method of estimating the reverberant component of the recorded signal, using the AIR representation in the STFT domain. This allows us to perform the dereverberation without any prior knowledge of the system's parameters and without confining ourselves to a specific statistical model. Moreover, formulating the problem directly in the STFT domain allows one to incorporate the frequency dependency of the reflective surfaces in the AIR model.

The method is carried out as follows: first we perform a system identification in order to find the filter which relates the recorded signal to its reverberant part. A number of approximate formulations are proposed for representing that relation. At the next stage, we use the identified filter to derive the estimator for the reverberant component of the recorded speech signal, and also to use in the existing spectral enhancement algorithm for dereverberation.

Our experimental results show that the simplest version of the identified filter where we use only the band-to-band filter, gives the lowest estimation error, due to the fact that the mathematical formulation is most accurate in that case. We also test the effect of the filter length on the estimation, and show that for a small number of crossbands we would prefer long filters, while for a large number of crossbands, short filters achieve lower estimation errors. Finally, we evaluate the new method in comparison to existing generalized statistical method, and show that, under specific assumptions, our method yields a higher performance

1.4 Structure

This thesis is organized as follows. In Chapter 2, we briefly present the work performed by Avargel and Cohen [50] on the subject of system identification in the STFT domain. There, we focus on the representation of LTI systems in the STFT domain, and on formulating the system identification problem for a desired number of crossbands. In Chapter 3, we present an existing method for dereverberation via spectral enhancement algorithms, which uses a statistical model for the AIR of the enclosed space. Our contribution begins in Chapter 4, where we utilize the Single-Side-Band (SSB) representation for speech dereverberation. We first derive the representation of LTI systems in the SSB domain, and then formulate the band-to-band identification problem for that representation, solved via a Least Squares (LS) optimization criterion. Finally we perform the dereverberation in the SSB domain, for full and approximate representations, and compare the results with those of the STFT domain representation. In Chapter 5 we formulate the dereverberation problem directly in the STFT domain, and propose a new method for estimation of the reverberant component using system identification with crossband filters. We analyse the relationship between the number of crossbands and the filter's length, and the accuracy of the estimation. Finally, in Chapter 6 we conclude our work and propose directions for future research.

Chapter 2

Dereverberation Using Spectral Enhancement

2.1 Introduction

Speech signals that are received by a microphone at a distance from the speech source usually contain reverberation, ambient noise and other interferences. Reverberation is the process of multi-path propagation of an acoustic sound from its source to a microphone. The received microphone signal generally consists of a direct sound, reflections that arrive shortly after the direct sound (commonly called *early reverberation*) and reflections that arrive after the early reverberation (commonly called *late reverberation*). The combination of the direct sound and early reverberation is sometimes referred to as the early speech component. Early reverberation mainly contributes to spectral colouration, while late reverberation changes the waveform's temporal envelope, and as a result, degrade speech fidelity and intelligibility.

In a reverberant room, speech intelligibility initially decreases with increasing source-microphone distance, but beyond the so-called critical distance speech intelligibility is approximately constant. The critical distance is the distance at which the direct-path energy is equal to the energy of all reflections. For distances smaller than the critical distance the measure depends on the source microphone distance, the reverberation time, and the volume of the room. Beyond the critical distance the measure depends only on the reverberation time.

Since the acoustic behaviour in real rooms is too complex to model explicitly, we make use of Statistical Room Acoustics (SRA). SRA provides a statistical description of the acoustic impulse response (AIR) of the system between the source and the microphone in terms of a few key quantities, e.g., source-microphone distance, and reverberation time. The crucial assumption of SRA is that the sound field has almost uniform distribution throughout the room volume. This assumption is valid if the following conditions hold [1, 53, 54]:

1. Dimensions of the room are relatively large compared to the longest wavelength of the sound of interest.

2. The average spacing of the resonance frequencies of the room must be smaller than one third of their bandwidth.

3. The source and the microphone are located in the interior of the room, at least a half-wavelength away from the walls.

In this chapter we describe a spectral enhancement method to suppress late reverberation and ambient noise, i.e., to estimate the early speech component. The method assumes a statistical model for the acoustic impulse response (AIR) of a room, that relies on the fact that its late component is dependent mostly on the reverberation time of the room [55]. We present here two statistical models: a Polack's model [56], and a generalized model [44], which is a modification of Polack's model that addresses the case where the source-microphone distance is smaller then the critical distance. Using a chosen statistical model, the algorithm provides a computation of the late reverberant component of the noisy signal, and then suppresses it via a post filtering stage.

2.2 Statistical Reverberation Models

2.2.1 Polacks Statistical Model

Polack [56] developed a time-domain model, where an AIR is described as a realization of a non-stationary stochastic process. This model is defined as

$$h(n) = \begin{cases} b(n)e^{-\delta n} & \text{for } n \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where n denotes the discrete time index, $b(n)$ is a zero-mean Gaussian noise, and δ is linked to the reverberation time T_{60} through

$$\delta = \frac{3 \ln 10}{T_{60} f_s} \tag{2.2}$$

where f_s denotes the sampling frequency. In contrast to the model in (2.1), the reverberation time is frequency dependent due to frequency dependent reflection coefficients of walls and other objects and the frequency dependent absorption coefficient of air [1, 51]. It should be noted that Polacks model is only valid in cases for which the distance between the source and the measurement point is greater than the critical distance D_c. In this case, the echo density is high enough such that the space can be considered to be in a fully diffused or mixed state.

The energy envelope of the AIR can be expressed as

$$E\left\{h^2(n)\right\} = \sigma^2 e^{-2\delta n} \tag{2.3}$$

where σ^2 denotes the variance of $b(n)$, and E denotes spatial expectation.

2.2.2 Generalized Statistical Model

In [44], a generalized statistical model was proposed, which can be used when the source-microphone distance is smaller than the critical distance. To model the contribution of the direct-path, the AIR $h(n)$ is divided into two segments:

$$h(n) = \begin{cases} h_d(n) & \text{for } 0 \leq n < n_d \\ h_r(n) & \text{for } n \geq n_d \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

The value n_d is chosen such that $h_d(n)$ contains the direct-path and $h_r(n)$ contains all the reflections. In practice, the direct-path is deterministic and could be modelled using a Dirac pulse. Unfortunately this would prevent us from creating a statistical model. To be able to model the energy related to the direct-path the following model of $h(n)$ is proposed:

$$h(n) = \begin{cases} b_d(n)e^{-\delta n} & \text{for } 0 \leq n < n_d \\ b_r(n)e^{-\delta n} & \text{for } n \geq n_d \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

where $b_d(n)$ and $b_r(n)$ are a white zero-mean Gaussian stationary noise. Under the SRA conditions, it is assumed that the direct and reverberant component of the AIR are uncorrelated [53].

The energy envelope of $h(n)$ can be expressed as

$$E\left\{h^2(n)\right\} = \begin{cases} \sigma_d^2 e^{-2\delta n} & \text{for } 0 \leq n < n_d \\ \sigma_r^2 e^{-2\delta n} & \text{for } n \geq n_d \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

where σ_d^2 and σ_r^2 denote the variances of $b_d(n)$ and $b_r(n)$, respectively. When $\sigma_d^2 < \sigma_r^2$, the contribution of the direct-path can be neglected. Therefore, it is assumed that $\sigma_d^2 \geq sigma_r^2$. For the case where $\sigma_d^2 = \sigma_r^2$ the generalized statistical model is equivalent to Polacks statistical model.

2.3 Single-microphone Spectral Enhancement

In this section we discuss the spectral enhancement of a noisy and reverberant microphone signal. We start by formulating the spectral enhancement problem, where we show how the spectrum of the early speech component can be estimated. Then we show how to estimate the spectrum of the late component, which is needed for the post-filtering.

2.3.1 Problem Formulation

The reverberant signal results from the convolution of the anechoic speech signal and a causal AIR. In this section we assume that the AIR is time-invariant and that its length

is infinite. The reverberant speech signal at discrete-time n can be written as

$$z(n) = \sum_{l=-\infty}^{\infty} x(l)h(n-l) \tag{2.7}$$

The direct-path is modelled by $h(0)$. Since our main goal is to suppress late reverberation we split the AIR into two components such that

$$h(n) = \begin{cases} h_e(n) & \text{for } 0 \leq n < n_e \\ h_l(n) & \text{for } n \geq n_e \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

where n_e is chosen such that $h_e(n)$ consists of the direct-path and a few early reflections and $h_l(n)$ consists of all later reflections. The fraction n_e/f_s can be used to define the time instance from where the late reverberation should be suppressed. In practice, n_e/f_s usually ranges from 30 to 60 ms. Using (2.8) we can write the microphone signal $y(n)$ as

$$y(n) = \underbrace{\sum_{l=n-n_e+1}^{n} y(l)h_e(n-l)}_{z_e(l)} + \underbrace{\sum_{l=-\infty}^{n-n_e} x(l)h_l(n-l) + v(n)}_{z_l(n)} \tag{2.9}$$

where $z_e(n)$ is the early speech component, $z_l(n)$ denotes the late reverberant speech component, and $v(n)$ denotes the additive ambient noise component.

Here we formulate the problem of estimating $z_e(n)$, or in other words suppressing $z_l(n)$, using spectral enhancement. The noisy and reverberant speech signal is first transformed to the STFT domain:

$$Y(l,k) = \sum_{n=0}^{K-1} y(n+lR)w(n)e^{-j\frac{2\pi}{K}nk} \tag{2.10}$$

where $w(n)$ is the analysis window of size K, and R is the number of samples separating two successive frames. $Y(l,k)$ can be used to estimate the spectral variance $\lambda_v(l,k) = E\left\{|V(l,k)|^2\right\}$ of the ambient noise and to estimate the spectral variance $\lambda_{z_l}(l,k) = E\left\{|Z_l(l,k)|^2\right\}$ of the late reverberant signal component. While the spectral variance of the noise is slowly time varying, the spectral variance of late reverberant signal component is highly time-varying due to the non-stationarity of the speech signal. An estimator for $\lambda_{z_l}(l,k)$ is derived in sec (2.4). For now, we assume that an estimate of the late reverberant spectral variance is available.

The spectral enhancement problem can be formulated as deriving an estimator $\hat{Z}_e(l,k)$ for the speech spectral coefficients such that a certain distortion measure is minimized [57].

2.3.2 Generalized Post-filter

Here, OM-LSA method for spectral enhancement is used to obtain an estimate of the spectral component $Z_e(l,k)$ [45]. Let $H_1(l,k)$ and $H_0(l,k)$ denote the hypotheses for speech presence and absence, respectively. Such that

$$H_1(l,k) : X(l,k) = Z_e(l,k) + Z_l(l,k) + V(l,k) \tag{2.11}$$

$$H_0(l,k) : X(l,k) = Z_l(l,k) + V(l,k) \tag{2.12}$$

The *a posteriori* SIR is then defined as

$$\gamma(l,k) = \frac{|Y(l,k)|}{\lambda_{z_l}(l,k) + \lambda_v(l,k)} \tag{2.13}$$

and the *a priori* SIR is defined as

$$\xi(l,k) = \frac{\lambda_{z_e}(l,k)}{\lambda_{z_l}(l,k) + \lambda_v(l,k)} \tag{2.14}$$

The spectral variance of noise can be estimated using [58]. The *a priori* SIR cannot be calculated directly since the spectral variance $\lambda_{z_e}(l,k)$ is unknown. The decision directed estimator [57] is used here to estimate the *a priori* SIR, which is given by

$$\hat{\xi}(l,k) = \max\left\{ \eta \frac{\left|\hat{Z}_e(l-1,k)\right|^2}{\lambda_{z_l}(l-1,k) + \lambda_v(l-1,k)} + (1-\eta)\max\left\{\gamma(l,k) - 1, 0\right\}, \xi_{min} \right\} \tag{2.15}$$

where ξ_{min} is the lower-bound on the *a priori* SIR that helps to reduce the musical noise, and $0 \leq \eta \leq 1$ is a weighting factor, usually chosen close to one.

When $H_1(l,k)$ is assumed to be true, and the interference signals are mutually independent, the log spectral amplitude (LSA) gain function is used, given by

$$G_{H_1}(l,k) = \frac{\xi(l,k)}{1+\xi(l,k)} \exp\left(\frac{1}{2}\int_{\varsigma(l,k)}^{\infty} \frac{e^{-}t}{t} dt\right) \tag{2.16}$$

where

$$\varsigma(l,k) \triangleq \frac{\xi(l,k)}{1+\xi(l,k)}\gamma(l,k). \tag{2.17}$$

When H_0 is assumed to be true, a lower bound $G_{H_0}(l,k)$ is applied given by

$$G_{H_0}(l,k) = G_{\min}\frac{\lambda_v(l,k)}{\lambda_{z_l}(l,k) + \lambda_v(l,k)} \tag{2.18}$$

The OM-LSA spectral gain function is obtained as a weighted geometric mean of the gains associated with the speech presence probability denoted by $p(l,k)$

$$G_{OM-LSA}(l,k) = \{G_{H_1}(l,k)\}^{p(l,k)} \{G_{H_0}(l,k)\}^{1-p(l,k)} \tag{2.19}$$

The spectral early component now can be estimated by applying the OM-LSA spectral gain function to each component of the received signal, i.e.,

$$\hat{Z}_e(l,k) = G_{OM-LSA}(l,k)Y(l,k) \tag{2.20}$$

2.4 Late Reverberant Spectral Variance Estimator

In this section we derive a spectral variance estimator for the late reverberant spectral component, $\lambda_{z_l}(l,k)$, using the generalized statistical reverberation model described in Sect.(2.2.2). We start by analysing the autocorrelation of the reverberant signal $z(n)$. The autocorrelation of the reverberant signal $z(n)$ at discrete time n and lag τ is defined as

$$r_{zz}(n, n+\tau) = E\{z(n)z(n+\tau)\} \tag{2.21}$$

Using (2.9) and (2.21), we obtain

$$\begin{aligned} r_{zz}(n, n+\tau) = &\sum_{l=n-n_+1}^{n} \sum_{l'=n-n_+1+\tau}^{n+\tau} E\{x(l)x(l')h_d(n-l)h_d(n+\tau-l')\} \\ &+ \sum_{l=-\infty}^{n-n_d} \sum_{l'=-\infty}^{n-n_d+\tau} E\{x(l)x(l')h_r(n-l)h_r(n+\tau-l')\} \end{aligned} \tag{2.22}$$

Assuming that the stochastic processes h and x are mutually independent, and using (2.6) and the fact that $b_d(n)$ and $b_r(n)$ are a zero-mean white Gaussian noise, it follows that

$$r_{zz}(n, n+\tau) = r_{z_d z_d}(n, n+\tau) + r_{z_r z_r}(n, n+\tau) \tag{2.23}$$

with

$$r_{z_d z_d}(n, n+\tau) = e^{-2\delta n} \sum_{l=n-n_d+1}^{n} E\{x(l)x(l+\tau)\} \sigma_d^2 e^{2\delta l} \tag{2.24}$$

and

$$\begin{aligned} r_{z_r z_r}(n, n+\tau) &= e^{-2\delta n} \sum_{l=-\infty}^{n-n_d} E\{x(l)x(l+\tau)\} \sigma_r^2 e^{2\delta l} \\ &= e^{-2\delta n} \sum_{l=n-2n_d+1}^{n-n_d} E\{x(l)x(l+\tau)\} \sigma_r^2 e^{2\delta l} + e^{-2\delta n} \sum_{l=-\infty}^{n-2n_d} E\{x(l)x(l+\tau)\} \sigma_r^2 e^{2\delta l} \end{aligned} \tag{2.25}$$

The first term in (2.23) depends on the direct signal between time $nn_d + 1$ and n, and the second depends on the reverberant signal. Let us consider the autocorrelation at time $n_0 = nn_d$. Using (2.24) and (2.25) autocorrelation of $r_{zz}(n - n_d, n - n_d + \tau)$ can be expressed

$$\begin{aligned} r_{zz}(n-n_d, n-n_d+\tau) = e^{-2\delta(n-n_d)} \sum_{l=n-2n_d+1}^{n-n_d} E\left\{x(l)x(l+\tau)\right\}\sigma_d^2 e^{2\delta l} + \\ + e^{-2\delta(n-n_d)} \sum_{l=-\infty}^{n-2n_d} E\left\{x(l)x(l+\tau)\right\}\sigma_r^2 e^{2\delta l} \end{aligned} \tag{2.26}$$

Let us define

$$\kappa = \frac{\sigma_r^2}{\sigma_d^2}, \text{ with } \kappa \leq 1 \tag{2.27}$$

From (2.26), (2.25) and (2.27) the term $r_{z_r z_r}(n, n+\tau)$ can be expressed as

$$r_{z_r z_r}(n, n+\tau) = \kappa e^{2\delta n_d} r_{z_d z_d}(n-n_d, n-n_d+\tau) + e^{2\delta n_d} r_{z_r z_r}(n-n_d, n-n_d+\tau) \tag{2.28}$$

Using (2.23) we can rewrite (2.28)

$$r_{z_r z_r}(n, n+\tau) = (1-\kappa)e^{2\delta n_d} r_{z_d z_d}(n-n_d, n-n_d+\tau) + \kappa e^{2\delta n_d} r_{z_d z_d}(n-n_d, n-n_d+\tau) \tag{2.29}$$

The late reverberant component can now be obtained using

$$r_{z_l z_l}(n, n+\tau) = (1-\kappa)e^{2\delta(n_e-n_d)} r_{z_r z_r}(n-n_e+n_d, n-n_e+n_d\tau) \tag{2.30}$$

The parameter κ is related to Direct to Reverberation Ratio (DRR), which is defined as

$$\frac{E_d}{E_r} = \frac{\sum_{l=0}^{n_d} h^2(l)}{\sum_{l=n_d+1}^{\infty} h^2(l)} = \frac{\sum_{l=0}^{n_d} \sigma_d^2 e^{-2\delta l}}{\sum_{l=n_d+1}^{\infty} \sigma_r^2 e^{-2\delta l}} = \frac{\sigma_d^2\left(1-e^{-2\delta n_d}\right)}{\sigma_r^2 e^{-2\delta n_d}} \tag{2.31}$$

Now the parameter κ can be expressed in terms of E_d and E_r:

$$\kappa = \frac{\sigma_r^2}{\sigma_d^2} = \frac{\left(1-e^{-2\delta n_d}\right)}{e^{-2\delta n_d}} \frac{E_r}{E_d} \tag{2.32}$$

In general the DRR is frequency dependent, as shown in [1]. Hence, to improve the accuracy of the model κ, as well as δ should be frequency dependent. In the following we assume that n_d is equal to the number of samples separating two successive STFT frames, denoted by R. Note that the value ne should be chosen such that n_e/R is an integer value.

Under these assumptions the counterparts of (2.29) and (2.30) in terms of the spectral variances are:

$$\lambda_{z_r}(l,k) = (1-\kappa(k))e^{-2\delta(k)R}\lambda_{z_r}(l-1,k) + \kappa e^{-2\delta(k)R}\lambda_z(l-1,k) \tag{2.33}$$

We can also substitute the expression $\lambda_z(l,k) = \lambda_{z_d}(l,k) + \lambda_{z_r}(l,k)$ to receive the form:

$$\lambda_{z_r}(l,k) = e^{-2\delta(k)R}\lambda_{z_r}(l-1,k) + \kappa e^{-2\delta(k)R}\lambda_{z_d}(l-1,k) \tag{2.34}$$

As we can see from those expressions, only the source can increase the reverberant energy in the room and the absorption of the energy is completely determined by the reverberation time of the room.

The late reverberant spectral variance (LRSV) is then given by

$$\lambda_{z_l}(l,k) = e^{-2\delta(k)(n_e-R)}\lambda_{z_r}(l-\frac{n_e}{R}+1,k) \tag{2.35}$$

For $\kappa(k) = 1$ the LRSV is given by

$$\lambda_{z_l}(l,k) = e^{-2\delta(k)n_e}\lambda_z(l-n_e,k) \tag{2.36}$$

which is equivalent to the LRSV estimator proposed by [43].

2.5 Summary

In this chapter, a single-microphone speech dereverberation method was described. This method uses a statistical model for the AIR of the room. We have presented Polack's model and a generalized model, and elaborated on the dereverberation algorithm using the generalized model. This model depends on the reverberation time of room and on the direct to reverberation ratio (DRR), or its related parameter, kappa. Under the assumption that the parameters of the model are known, we showed how a late spectral reverberant variance (LRSV) can be extracted from the received reverberant and noisy signal. Together with the estimation of noise variance, the LRSV is used to compute the *a-priori* SIR and the post-filter, via OM-LSA estimation method. The generalized model is especially advantageous in case the source microphone distance is smaller than the critical distance.

Chapter 3

System Identification in the STFT Domain with Crossband Filtering

3.1 Introduction

For the completeness of understanding of this thesis, we present in this chapter the work of Avargel and Cohen [50]. Here we concentrate on the subject of system identification with long impulse responses in the STFT domain, which is of major importance in many signal processing application, including dereverberation [16, 59]. When dealing with dereverberation problem, we try to cancel or to suppress the effect caused by the acoustic system. Therefore, we might need first to identify this system, and then incorporate its expression in the different methods for dereverberation.

In this chapter we first review the development of the relation between the crossband filters in the STFT domain and the impulse response in time domain. This relation shows that the number of crossband filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows, employed for the STFT representation. Then we consider an offline system identification in the STFT domain with changing number of crossbands, using the least squares (LS) optimization criterion.

This chapter is organized as follows. Section 3.2 describes an LTI system representation in the STFT domain. Section 3.3 formulates the system identification problem in the STFT domain, using exact representation and $2K$-crossband approximation.

3.2 Representation of LTI Systems in the STFT Domain

The STFT representation of a signal $x(n)$ is given by

$$x_{l,k} = \sum_m x(m)\tilde{\psi}^*_{l,k}(m)\,, \tag{3.1}$$

where

$$\tilde{\psi}_{l,k}(n) \triangleq \tilde{\psi}(n - pL)e^{j\frac{2\pi}{N}k(n-pL)}\,, \tag{3.2}$$

$\tilde{\psi}(n)$ denotes an analysis window (or analysis filter) of length N, l is the frame index, k represents the frequency-band index, L is the discrete-time shift, and * denotes complex conjugation. The inverse STFT, is given by

$$x(n) = \sum_l \sum_{k=0}^{N-1} x_{l,k}\psi_{l,k}(n)\,, \tag{3.3}$$

where

$$\psi_{l,k}(n) \triangleq \psi(n - pL)e^{j\frac{2\pi}{N}k(n-pL)} \tag{3.4}$$

and $\psi(n)$ denotes a synthesis window (or synthesis filter) of length N. Let $h(n)$ denote a length Q impulse response of an LTI system, whose input $x(n)$ and output $d(n)$ are related by

$$d(n) = \sum_{i=0}^{Q-1} h(i)x(n - i)\,. \tag{3.5}$$

Using (3.1) and (3.5), the STFT representation of $d(n)$ can be written as

$$d_{l,k} = \sum_{m,p} h(p)x(m - p)\tilde{\psi}^*_{l,k}(m) \tag{3.6}$$

Substituting (3.3) into (3.6), we obtain

$$\begin{aligned} d_{l,k} &= \sum_{m,p} h(p) \sum_{k'=0}^{N-1} \sum_{l'} x_{l',k'}\psi_{l',k'}(m - p)\tilde{\psi}^*_{l,k}(m) \\ &= \sum_{k'=0}^{N-1} \sum_{l'} x_{l',k'}h_{l,k,l',k'} \end{aligned} \tag{3.7}$$

where

$$h_{l,k,l',k'} = \sum_{m,p} h(p)\psi_{l',k'}(m - p)\tilde{\psi}^*_{l,k}(m) \tag{3.8}$$

may be interpreted as the STFT of $h(n)$ using a composite analysis window $\sum_m \psi_{l',k'}(m-p)\tilde{\psi}^*_{l,k}(m)$. Substituting (3.2) and (3.4) into (3.8), we obtain

$$\begin{aligned} h_{l,k,l',k'} &= \sum_{m,p} h(p)\psi(m-p-l'L)e^{j\frac{2\pi}{N}k'(m-p-l'L)}\tilde{\psi}(m-pL)e^{-j\frac{2\pi}{N}k(m-pL)} \\ &= \sum_p h(p)\sum_m \tilde{\psi}(m)e^{-j\frac{2\pi}{N}km}\psi\left((l-l')L-p+m\right)e^{j\frac{2\pi}{N}k'((l-l')L-p+m)} \\ &= \left\{h(n)*\phi_{k,k'}(n)\right\}|_{n=(l-l')L} \triangleq h_{l-l',k,k'}\,, \end{aligned} \tag{3.9}$$

where $*$ denotes convolution with respect to the time index n, and

$$\phi_{k,k'}(n) \triangleq e^{j\frac{2\pi}{N}k'n}\sum_m \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')}\,. \tag{3.10}$$

From (3.9), $h_{l,k,l',k'}$ depends on $(l-l')$ rather than on l and l' separately. Substituting (3.9) into (3.7), we obtain

$$d_{l,k} = \sum_{k'=0}^{N-1}\sum_{l'} x_{l',k'}h_{l-l',k,k'} = \sum_{k'=0}^{N-1}\sum_{l'} x_{l-l',k'}h_{l',k,k'}\,, \tag{3.11}$$

where $h_{l-l',k,k'}$ may be interpreted as a response to an impulse $\delta_{l-l',k-k'}$ in the STFT domain. The impulse response $h_{l,k,k'}$ in the STFT domain is related to the impulse response $h(n)$ in the time domain by

$$h_{l,k,k'} = \left\{h(n)*\phi_{k,k'}(n)\right\}|_{n=pL} \triangleq \bar{h}_{n,k,k'}\big|_{n=pL}\,, \tag{3.12}$$

where $*$ denotes convolution with respect to the time index n and

$$\begin{aligned} \phi_{k,k'}(n) &\triangleq e^{j\frac{2\pi}{N}k'n}\sum_m \tilde{\psi}(m)\psi(n+m)e^{-j\frac{2\pi}{N}m(k-k')} \\ &= e^{j\frac{2\pi}{N}k'n}\psi_{n,k-k'}\,, \end{aligned} \tag{3.13}$$

where $\psi_{n,k}$ is the STFT representation of the synthesis window $\psi(n)$ calculated with a decimation factor $L=1$. Equation (3.11) indicates that for a given frequency-band index k, the temporal signal $d_{l,k}$ can be obtained by convolving the signal $x_{l,k'}$ in each frequency-band k' ($k'=0,1,\ldots,N-1$) with the corresponding filter $h_{l,k,k'}$ and then summing over all the outputs. We refer to $h_{l,k,k'}$ for $k=k'$ as a band-to-band filter and for $k\neq k'$ as a crossband filter. Cross-band filters are used for cancelling the aliasing effects caused by

the sub-sampling. It can also be seen from (3.12) that the length of each crossband filter is given by

$$N_h = \left\lceil \frac{Q+N-1}{L} \right\rceil + \left\lceil \frac{N}{L} \right\rceil - 1\,. \tag{3.14}$$

Had the analysis and the synthesis windows been ideal low-pass filters with bandwidth $f_s/2N$ (where f_s is the sampling frequency), a perfect STFT representation of the system $h(n)$ could be achieved by using just the band-to-band filter $h_{n,k,k}$. However, their bandwidths are generally greater than $f_s/2N$ and therefore, $h_{n,k,k'}$ is not zero for $k \neq k'$. Nevertheless, the energy of a crossband filter from frequency-band k' to frequency-band k decreases as $|k-k'|$ increases, since the overlap between the windows in the frequency domain becomes smaller. As a result, relatively few crossband filters (up to eight) need to be considered in order to capture most of the energy of the STFT representation of $h(n)$.

In general, the number of cross-band filters required for the representation of an impulse response is mainly determined by the analysis and synthesis windows, while the length of the crossband filters (with respect to the time index n) is related to the length of the impulse response.

3.3 System Identification in the STFT Domain

In this section, we consider system identification in the STFT domain and address the problem of estimating the crossband filters of the system using an LS optimization criterion for each frequency-band. Throughout this section, scalar variables are written with lowercase letters and vectors are indicated with lowercase boldface letters. Capital boldface letters are used for matrices and norms are always ℓ_2 norms.

The input signal $x(n)$ passes through an unknown system characterized by its impulse response $h(n)$, obtaining the desired signal $d(n)$. Together with the background noise signal $v(n)$, the system output signal is given by

$$y(n) = d(n) + v(n) = h(n) * x(n) + v(n)\,. \tag{3.15}$$

From (3.15) and (3.11), the STFT of $y(n)$ may be written as

$$y_{l,k} = d_{l,k} + v_{l,k} = \sum_{k'=0}^{N-1} \sum_{l'=0}^{N_h-1} x_{l-l',k'} h_{l',k,k'} + v_{l,k}\,, \tag{3.16}$$

where N_h is the length of the crossband filters. Defining N_x as the length of $x_{l,k}$ in frequency band k, we can write the length of $y_{l,k}$ for a fixed k as $N_y = N_x + N_h - 1$.

Let $\mathbf{h}_{k,k'}$ denote the crossband filter from frequency-band k' to frequency-band k

$$\mathbf{h}_{k,k'} = \begin{bmatrix} h_{0,k,k'} & h_{1,k,k'} & \cdots & h_{N_h-1,k,k'} \end{bmatrix}^T \tag{3.17}$$

and let $\mathbf{h}_k$ denote a column-stack concatenation of the filters $\{\mathbf{h}_{k,k'}\}_{k'=0}^{N-1}$

$$\mathbf{h}_k = \begin{bmatrix} \mathbf{h}_{k,0}^T & \mathbf{h}_{k,1}^T & \cdots & \cdots & \mathbf{h}_{k,N-1}^T \end{bmatrix}^T . \tag{3.18}$$

Let

$$\mathbf{X}_k = \begin{bmatrix} x_{0,k} & 0 & \cdots & \cdots & 0 \\ x_{1,k} & x_{0,k} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{N_y-1,k} & \cdots & \cdots & \cdots & x_{N_y+N_h-2,k} \end{bmatrix} \tag{3.19}$$

represent an $N_y \times N_h$ Toeplitz matrix constructed from the input signal STFT coefficients of the k-th frequency-band, and let $\mathbf{\Delta}_k$ be a concatenation of $\{\mathbf{X}_k\}_{k=0}^{N-1}$ along the column dimension

$$\mathbf{\Delta}_k = \begin{bmatrix} \mathbf{X}_0 & \mathbf{X}_1 & \cdots & \cdots & \mathbf{X}_{N-1} \end{bmatrix} . \tag{3.20}$$

Then, (3.16) can be written in a vector form as

$$\mathbf{y}_k = \mathbf{d}_k + \boldsymbol{v}_k = \mathbf{\Delta}_k \mathbf{h}_k + \boldsymbol{v}_k \, , \tag{3.21}$$

where

$$\mathbf{y}_k = \begin{bmatrix} y_{0,k} & y_{1,k} & y_{2,k} & \cdots & y_{N_y-1,k} \end{bmatrix}^T \tag{3.22}$$

represents the output signal STFT coefficients of the k-th frequency-band, and the vectors $\mathbf{d}_k$ and $\boldsymbol{v}_k$ are defined similarly.

Let $\hat{h}_{l',k,k'}$ be an estimate of the crossband filter $h_{l',k,k'}$, and let $\hat{d}_{l,k}$ be the resulting estimate of $d_{l,k}$ using only $2K$ crossband filters around the frequency-band k, *i.e.*,

$$\hat{d}_{l,k} = \sum_{k'=k-K}^{k+K} \sum_{l'=0}^{N_h-1} \hat{h}_{l',k,k' \bmod N} x_{l-l',k' \bmod N} \, , \tag{3.23}$$

where we exploited the periodicity of the frequency-bands. Let $\hat{\tilde{\mathbf{h}}}_k$ be the $2K+1$ estimated filters at frequency band k

$$\hat{\tilde{\mathbf{h}}}_k = \begin{bmatrix} \hat{\mathbf{h}}_{k,(k-K) \bmod N}^T & \hat{\mathbf{h}}_{k,(k-K+1) \bmod N}^T & \cdots & \cdots & \hat{\mathbf{h}}_{k,(k+K) \bmod N}^T \end{bmatrix}^T , \tag{3.24}$$

where $\hat{\mathbf{h}}_{k,k'}$ is the estimated crossband filter from frequency-band k' to frequency-band k, and let $\tilde{\mathbf{\Delta}}_k$ be a concatenation of $\{\mathbf{X}_{k'}\}_{k'=(k-K) \bmod N}^{(k+K) \bmod N}$ along the column dimension

$$\tilde{\mathbf{\Delta}}_k = \left[\begin{array}{cccccc} \mathbf{X}_{(k-K) \bmod N} & \mathbf{X}_{(k-K+1) \bmod N} & \cdots & \cdots & \mathbf{X}_{(k+K) \bmod N} \end{array} \right] . \tag{3.25}$$

Then, the estimated desired signal can be written in a vector form as

$$\hat{\mathbf{d}}_k = \tilde{\mathbf{\Delta}}_k \hat{\tilde{\mathbf{h}}}_k \, , \tag{3.26}$$

Using the above notations, the LS optimization problem can be expressed as

$$\hat{\tilde{\mathbf{h}}}_k = \arg \min_{\tilde{\mathbf{h}}_k} \left\| \mathbf{y}_k - \tilde{\mathbf{\Delta}}_k \tilde{\mathbf{h}}_k \right\|^2 . \tag{3.27}$$

The solution to (3.27) is given by

$$\hat{\tilde{\mathbf{h}}}_k = \left(\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k \right)^{-1} \tilde{\mathbf{\Delta}}_k^H \mathbf{y}_k \, , \tag{3.28}$$

where we assumed that $\tilde{\mathbf{\Delta}}_k^H \tilde{\mathbf{\Delta}}_k$ is not singular. Substituting (3.28) into (3.26), we obtain an estimate of the desired signal in the STFT domain at the k-th frequency-band, using $2K$ crossband filters.

3.4 Summary

In this chapter we presented system identification problem formulation in the STFT domain. First we reviewed the development of the crossband filter, and showed how it relates to the impulse response function. At the second part, we dealt with the identification problem, using the LS criterion. The identification problem was formulated directly in the STFT, for changing number of crossband filters around the main band. This formulation allows one to find an optimal number of crossband filters to use in a specific problem, given the required computation complexity, and the SNR conditions.

Chapter 4

Dereverberation in the SSB Domain

4.1 Introduction

The Single-Side-Band (SSB) transform is an important time-frequency representation. Unlike the short-time Fourier transform (STFT), the SSB representation has real-valued channel signals instead of complex valued signals, and therefore it is often the choice in real-time low-cost applications involving communication, coding systems and speech processing. The SSB can be realized in an efficient manner by sharing computations among channels, employing efficient methods for decimation and interpolation, and by using fast algorithms for modulation and demodulation.

In this chapter, we employ the SSB transform in two related subjects: system identification and dereverberation. System identification is of major importance in many applications, including acoustic echo cancellation [60], beamforming [61], and dereverberation [16,36]. These applications share the common need of identifying the channel which is the source of the signal interference. An estimation for the clean signal can then be performed by a spectral subtraction of the interference.

As a first step in identification we derive an analytical expression for the impulse response of a linear time invariant (LTI) system in the SSB domain, and propose a possible approximation for that expression. We then present an offline system identification procedure for the approximation using a least squares (LS) criterion and investigate the performance of the identification for different signal-to-noise (SNR) conditions.

The second subject that is addressed is dereverberation via a spectral enhancement

method, that assumes a statistic model for the AIR [43,44]. Based on one of the statistical models proposed in [55,56], the algorithm estimates the late reverberant spectral variance (LRSV) component, which is the main contributor to the degradation of the signal. The clean speech signal is then estimated using one of the methods presented in [42, 45, 57].

In most of the existing methods, the AIR model is represented in the time domain, and the enhancement stage is performed in the STFT domain [43, 44, 62]. The drawback of using the STFT representation is the need to process complex signals, and incorporate crossband filters to achieve a sufficiently accurate representation [50], which complicates the algorithm's implementation. Therefore, we apply a formulation of the AIR model and the reverberated signal directly in the SSB domain, using approximate representations. Then we study how the dereverberation performance depends on the number of crossbands. Finally, we compare the performance using the SSB transform to the one obtained using the STFT representation.

This chapter is organized as follows. Section 4.2 describes an LTI system representation in the SSB domain. Section 4.3 addresses the problem of system identification. Section 4.4 presents the dereverberation in the SSB domain. Experimental results are demonstrated in Section 4.5.

4.2 Representation of LTI Systems in the SSB domain

In this section, we derive an analytical relation between the input and the output signals of an LTI system in the SSB domain. Throughout this paper, unless explicitly noted, the summation indexes range from $-\infty$ to ∞. The SSB representation of a signal $x(n)$ is given by

$$X_{m,k} = Re\left[\sum_n \tilde{\psi}(mM-n)x(n)e^{\frac{j\pi m}{2}}W_K^{-kn}\right] \tag{4.1}$$

where $\tilde{\psi}$ denotes the analysis window, m the frame index, k the frequency-band index, M the decimation factor and K represents the number of frequency bands used in the transform. W_K is defined as

$$W_K = e^{\frac{2\pi j}{K}}. \tag{4.2}$$

The inverse SSB transform is given by

$$x(n) = \frac{1}{K}\sum_{k=0}^{K-1}\sum_{m} Re\left[\psi(mM-n)X_{m,k}e^{-\frac{j\pi m}{2}}W_K^{kn}\right] \tag{4.3}$$

where ψ denotes the synthesis window. Let $h(n)$ denote an impulse response of an LTI system of length Q. The output signal in the SSB domain is given by

$$Y_{m,k} = Re\left[\sum_{n}\tilde{\psi}(n-mM)\sum_{l=0}^{Q-1}h(l)x(n-l)e^{\frac{j\pi m}{2}}W_K^{-kn}\right]. \tag{4.4}$$

After some manipulations $Y_{m,k}$ can be written as

$$Y_{m,k} = \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'} H_{m,m',k,k'}X_{m',k'} \tag{4.5}$$

where

$$H_{m,m',k,k'} = \sum_{n}\vartheta 1_{m,k,n}\sum_{l=0}^{Q-1}h(l)\vartheta 2_{m',k',n-l} \tag{4.6}$$

with

$$\begin{aligned}\vartheta 1_{m,k,n} &= \tilde{\psi}(mM-n)\cos\left(\frac{\pi m}{2}-\frac{2\pi kn}{K}\right)\\ \vartheta 2_{m',k',n} &= \psi(n-m'M)\cos\left(\frac{\pi m'}{2}-\frac{2\pi k'n}{K}\right)\end{aligned} \tag{4.7}$$

We refer to $H_{m,m',k,k'}$ for $k = k'$ as a band-to-band filter and for $k \neq k'$ as a crossband filter. In order to simplify the expression in (4.6) we propose approximate representations which employ only part or non of the crossband filters. For an approximation which uses $2K_{max}$ crossbands, the output signal is given by

$$Y_{m,k} = \frac{1}{K}\sum_{k'=k-K_{max}}^{k+K_{max}}\sum_{m'} H_{m,m',k,k'}X_{m',k'}\,. \tag{4.8}$$

For $K_{max} = 0$ the approximate representation uses only the band-to-band filter.

4.3 System Identification in the SSB Domain

In this section, we consider system identification in the SSB domain using the band-to-band approximation and an LS optimization criterion. The input signal $x(n)$ passes through an unknown system characterized by its impulse response $h(n)$, resulting in the

desired signal $d(n)$. Together with the background white noise $v(n)$, the output signal is given by

$$y(n) = d(n) + v(n) = h(n) * x(n) + v(n)\,. \tag{4.9}$$

From (4.9) and (4.5), the SSB representation of $y(n)$ may be written as

$$Y_{m,k} = D_{m,k} + V_{m,k} = \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'} H_{m,m',k,k'}X_{m',k'} + V_{m,k} \tag{4.10}$$

where $V_{m,k}$ is the SSB transform of $v(n)$.

Let us define N_{xh} and N_x as the number of time and crosstime samples, respectively, of the filter $H_{m,m',k,k'}$. Let $\mathbf{H}^{\mathbf{bb}}_{m,k}$ denote the band-to-band filter for time sample m and frequency band k.

$$\mathbf{H}^{\mathbf{bb}}_{m,k} = \left[\begin{array}{cccc} H^{\mathbf{bb}}_{m,0,k} & H^{\mathbf{bb}}_{m,1,k} & \cdots & H^{\mathbf{bb}}_{m,N_x-1,k} \end{array}\right]^T \tag{4.11}$$

and let $\mathbf{H}^{\mathbf{bb}}_{k}$ denote a column-stack concatenation of the above band-to-band filter $\left\{\mathbf{H}^{\mathbf{bb}}_{m,k}\right\}_{m=0}^{N_{xh}-1}$ for all the time samples m

$$\mathbf{H}^{\mathbf{bb}}_{\mathbf{k}} = \left[\begin{array}{cccc} \mathbf{H}^{\mathbf{bb}^T}_{0,k} & \mathbf{H}^{\mathbf{bb}^T}_{1,k} & \cdots & \cdots \quad \mathbf{H}^{\mathbf{bb}}_{N_{xh}-1,k}{}^T \end{array}\right]^T . \tag{4.12}$$

The dimensions of $\mathbf{H}^{\mathbf{bb}}_{\mathbf{k}}$ are $N_{xh} \times N_x$. Let $\mathbf{X}_k$ be the signal X at band k and let

$$\mathbf{\Delta}_k = \left[\begin{array}{ccccc} \mathbf{X}_k & 0 & \cdots & \cdots & 0 \\ 0 & \mathbf{X}_k & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & \mathbf{X}_k \end{array}\right] \tag{4.13}$$

represent a sparse matrix constructed from the input signal SSB coefficients of the k-th frequency-band, replicated N_{xh} times, where each replication is shifted by N_x columns with respect to the previous line. Now we can write the band-to-band estimate of the desired signal $\mathbf{D}_k$ in a vector form as

$$\mathbf{D}^{\mathbf{bb}}_{k} = \mathbf{\Delta}_k \mathbf{H}^{\mathbf{bb}}_{\mathbf{k}}\,. \tag{4.14}$$

This represents the SSB coefficients of the output signal at the k-th frequency-band, resulting from only the band-to-band filter $\mathbf{H}_k$.

Using the above notations, the LS optimization problem can be expressed as

$$\hat{\mathbf{H}}_{\mathbf{k}}^{\mathbf{bb}} = \arg\min_{\mathbf{H}_{\mathbf{k}}^{\mathbf{bb}}} \left\| \mathbf{Y}_k - \mathbf{\Delta}_k \mathbf{H}_{\mathbf{k}}^{\mathbf{bb}} \right\|^2 . \tag{4.15}$$

The solution to (4.15) is given by

$$\hat{\mathbf{H}}_k^{\mathbf{bb}} = \left(\mathbf{\Delta}_k^H \mathbf{\Delta}_k \right)^{-1} \mathbf{\Delta}_k^H \mathbf{Y}_k \tag{4.16}$$

where we assumed that $\mathbf{\Delta}_k^H \mathbf{\Delta}_k$ is not singular (otherwise, some regularization is included). Substituting (4.16) into (4.14), we obtain

$$\hat{\mathbf{D}}_k^{\mathbf{bb}} = \mathbf{\Delta}_k \hat{\mathbf{H}}_k^{\mathbf{bb}} \tag{4.17}$$

which is the estimate of the desired signal in the SSB domain at the k-th frequency-band using a band-to-band filter.

4.3.1 MSE computation

After calculating the estimated signal, we can analyse the mean-squared error (MSE) from two aspects:

1. An estimation error - derived by calculating the MSE between the estimated signal, $\hat{\mathbf{D}}_k^{\mathbf{bb}}$, and the real signal $D_{m,k}$ as defined in (4.10)

$$\epsilon_{\mathbf{estimate}} = \frac{E\left\{ \left\| \mathbf{D}_k - \hat{\mathbf{D}}_k^{\mathbf{bb}} \right\|^2 \right\}}{E\left\{ \left\| \mathbf{D}_k \right\|^2 \right\}} . \tag{4.18}$$

2. A theoretical error - derived by calculating the MSE between the estimated signal, $\hat{\mathbf{D}}_k$, and the signal $\mathbf{D}_k^{\mathbf{bb}}$ as defined in (4.14)

$$\epsilon_{\mathbf{theory}} = \frac{E\left\{ \left\| \mathbf{D}_k^{\mathbf{bb}} - \hat{\mathbf{D}}_k^{\mathbf{bb}} \right\|^2 \right\}}{E\left\{ \left\| \mathbf{D}_k^{\mathbf{bb}} \right\|^2 \right\}} . \tag{4.19}$$

4.4 Dereverberation in the SSB domain

In a reverberant environment, the AIR model in the time domain is given by [44]

$$h(n) = \begin{cases} b_d(n) & \text{if } 0 \le n < T_s \\ b_r(n) e^{-\delta(k) n} & \text{if } n \ge T_s \end{cases} \tag{4.20}$$

where $\delta(k)$ denotes the decay rate related to the reverberation time, $b_d(n)$ and $b_r(n)$ are zero-mean mutually independent and identically distributed (i.i.d.) Gaussian random variables, and T_s is the time when the early reflections end.

Assuming that the path from the source to the microphone can be treated as an LTI system, and using (4.5) and (4.20), we can express the reverberant signal $y(n)$ in the SSB domain as:

$$\begin{aligned} Y_{m,k} &= \frac{1}{K}\sum_{k'=0}^{K-1}\sum_{m'} H_{m,m',k,k'}X_{m',k'} = \\ &= \begin{cases} \frac{1}{K}\sum\limits_{k'=0}^{K-1}\sum\limits_{m'}\left(\sum\limits_{n}\vartheta_{1_{m,k,n}}\sum\limits_{l=0}^{Q-1} b_d(l)\vartheta_{2_{m',k',n-l}}X_{m',k'}\right), \\ \qquad\qquad\qquad\qquad \text{if } 0 \le m < N_e, \\ \frac{1}{K}\sum\limits_{k'=0}^{K-1}\sum\limits_{m'}\left(\sum\limits_{n}\vartheta_{1_{m,k,n}}\sum\limits_{l=0}^{Q-1} b_r(l)e^{-\delta(k)l}\vartheta_{2_{m',k',n-l}}X_{m',k'}\right), \\ \qquad\qquad\qquad\qquad \text{if } m \ge N_e. \end{cases} \end{aligned} \tag{4.21}$$

The parameter N_e specifies the portion of the AIR that is considered as late reverberations, and is related to T_s in the time domain.

Assuming that the SSB coefficients of the speech signal can be modelled as zero-mean i.i.d real random variables with a certain distribution and variance $\lambda_x(m,k)$, the expression for the reverberant component as presented in [44] is:

$$\begin{aligned} \lambda_r(m,k) = [1-\kappa(k)]\, e^{-2\delta(k)R}\lambda_r(m-1,k) + \\ +\kappa(k)e^{-2\delta(k)R}\lambda_y(m-1,k) \end{aligned} \tag{4.22}$$

where $\lambda_y(m,k) = E\left\{|Y(m,k)|^2\right\}$ and $\kappa(k)$ denotes the ratio between the energy of the reverberant and the direct path. The LRSV [43] is then given by

$$\lambda_l(m,k) = e^{-2\delta(k)R(N_e-1)}\lambda_r(m-N_e+1,k)\,. \tag{4.23}$$

4.5 Experimental Results

The signals used in the simulations include synthetic white Gaussian noise as well as real speech signals. Throughout this section, the AIR was simulated according to the method proposed in [63], with room dimensions of $6 \times 8 \times 5$ m, and a reverberation time of 500 ms. The SSB was implemented using $K = 32$ frequency bands, Kaiser synthesis window

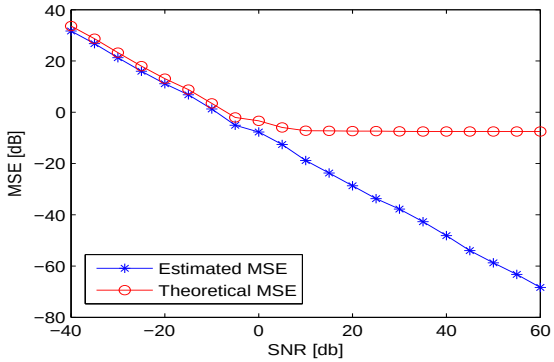


Figure 4.1: Theoretical and estimated MSE curves for the band-to-band identification system, as a function of SNR for a white Gaussian noise input signal.

of $4N + 1 = 129$ samples, and the related bi-orthogonal analysis window. The overlap between two successive frames was 50%.

4.5.1 System Identification

System identification results are shown under the assumption of band-to-band filtering, for SNR conditions ranging from -40 to 60 dB. Both the signal and the noise were white Gaussian noise of 2000 samples. In this subsection the source-microphone distance was 1 m, the length of the AIR was truncated to $Q = 700$, and the sampling rate was 8 kHz.

Figure 4.1 shows the graph of theoretical and estimated MSE for different SNR conditions. The estimated-MSE, is getting smaller as the SNR increases in spite of the fact that the model neglects all the cross-band filters. On the other hand, the theoretical-MSE remains almost constant after a certain SNR. This is due to the fact that the LS optimization was performed using the real output full-band signal. In other words, the identified model is closer to the representation of the full system, even though it lacks one dimension.

4.5.2 Dereverberation

In this subsection, we present and discuss results of dereverberation obtained using the SSB representation. The simulated AIR was of length $Q = 4096$ and the source-microphone distance varied between 0.5m and 3m. The parameter T_s was set to 48ms.

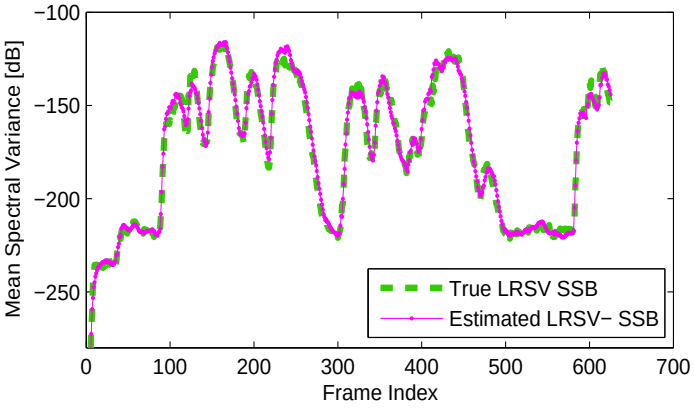


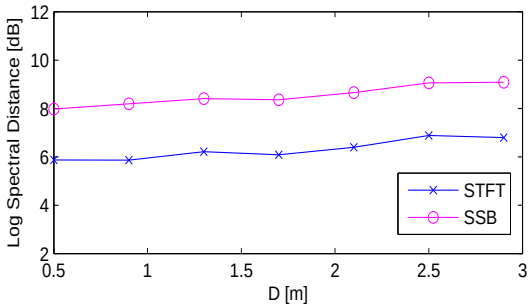
Figure 4.2: Mean Spectral Variance of true and estimated LRSVs of speech signal in the SSB Domain.

For qualitative evaluation of the LRSV estimation we used the mean spectral variance of the LRSV over all the frequency bins, which is given by

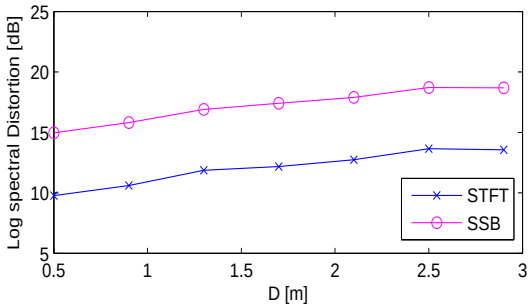
$$\text{Mean Spectral Variance [dB]} = 10\log\left(\text{mean}_k\left\{\lambda_l(m,k)\right\}\right) \tag{4.24}$$

The Mean Spectral Variance of the estimated LRSV was compared to the "true" LRSV, known from the AIR simulation [63]. The quantitative evaluation of the LRSV estimator was determined by the Log Spectral Distance measure. The dereverberation performance was evaluated using the mean segmental Signal to Reverberation Ratio (SRR) and the mean Log Spectral Distortion (LSD). Figure 4.2 shows the resulting true and estimated mean LRSVs of speech signals in the SSB domain, for a source-microphone distance of 1.3 m.

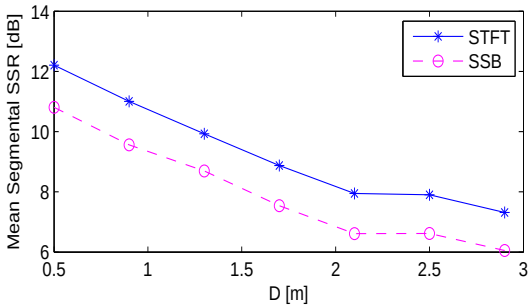
Figure 4.3 shows the dereverberation evaluation curves for a speech signal as a function of source microphone distance for the SSB and STFT representations. Clearly, the performance using the STFT representation is higher, which implies that real-valued representations are less suitable for dereverberation. This is associated with the fact that real-valued representations combine the phase information into the amplitude representation. Consequently, in estimating the LRSV we have to use a larger smoothing factor to compensate for multiple reflections with different delays, and this degrades the performance.



(a)



(b)



(c)

Figure 4.3: Dereverberation evaluation in the SSB domain in comparison to the STFT domain. (a) Log Spectral Distance; (b) Mean LSD; (c) Mean SRR.

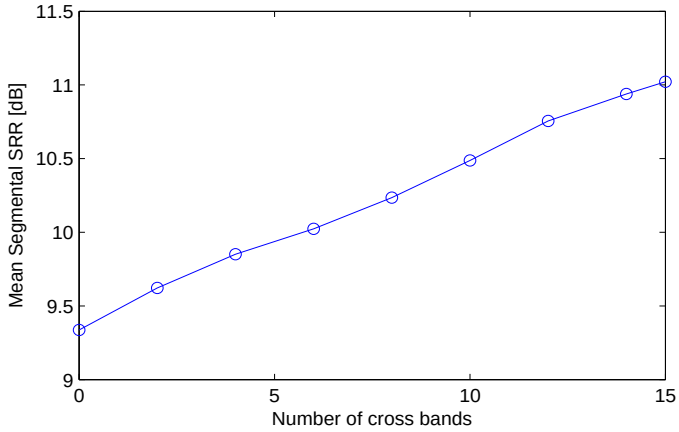


Figure 4.4: Mean SRR of the derverberation in the SSB Domain using various numbers of crossbands.

4.5.3 Cross-band analysis

Here, we analyse the dereverberation performance when using an increasing number of crossbands such that $0 \leq K_{max} \leq 15$. The input signal is white Gaussian noise of 2000 samples. The sampling rate is 4 kHz, and the length of the AIR is 1000 taps.

As can be seen from Figure 4.4, unlike the STFT case [50], the contribution of the crossband filters is distributed almost equally along all the crossbands. Nevertheless, as was shown in the system identification procedure, the band-to-band representation sufficiently describes the system and thus yields satisfying results with a low computational complexity.

4.6 Conclusions

We have investigated the SSB transform as a time-frequency domain representation for processing of speech signals. First, we developed a formulation of LTI systems in the SSB domain. Then we proposed system identification using a band-to-band filter approximation. We showed that as SNR improves, the identified band-to-band system becomes closer to the real system, even though it lacks the crossband dimension. This implies that the band-to-band approximation can sufficiently describe the system.

We also investigated the performance of dereverberation in the SSB transform domain, compared to dereverberation in the STFT domain. The evaluation measures show that

the STFT enables better results due to the fact it separates the spectral magnitude and phase representations, and thus facilitates the LRSV estimation. Finally, we examined the relationship between the AIR model complexity and the dereverberation performance, and showed that although the band-to-band representation gives sufficient results, each additional crossband contributes to further improvement.

Chapter 5

Dereverberation Using System Identification in the STFT domain with Crossband Filtering

5.1 Introduction

Acoustic signals radiated within a room are linearly distorted by reflections from walls and other objects. This type of distortion is commonly known as reverberation and it degrades the fidelity and intelligibility of speech. The resulting distorted output signal is a superposition of the direct path, which is an attenuated version of the original signal, and of the reverberant path, that consists of all the reflections.

Many methods have been proposed to address the dereverberation problem. One of the major groups, called blind deconvolution, identifies the acoustic impulse response (AIR) function, and based on that, designs an inverse filter to reduce the reverberation effect. For example, AIR identification performed via the Least Squares (LS) method was proposed by Xu et al. [2]. Identification by eigen-decomposition was proposed by Gurelli and Nikias [3], where the algorithm estimates the orders and root locations of the channel transfer functions. Using multichannel LMS and Newton adaptive filters, both in the time and frequency domains, was proposed by Huang and Benesty [4, 5], and was later used more specifically for the dereverberation application [16, 36].

While the proposed methods enable one to identify the channel blindly, without any

prior knowledge of its properties, they suffer from several limitations. They usually require a multi-channel system, where the channels do not contain common zeros. High sensitivity to additive noise is another large drawback of those methods. Finally, formulation of the identification problem is done in the time domain, and thus addressing the frequency dependence of the function requires performing the entire procedure in each band separately.

To overcome some of the aforementioned drawbacks, the group of spectral enhancement methods was developed. These methods use the common methods for speech enhancement [42, 45, 57], and apply them to reverberant speech [43, 44]. The estimation of the short-term Power Spectral Density (PSD) of late reverberation is based on a statistical model for the room impulse response. Such statistical models usually exploit the fact that the envelope of the AIR decays exponentially and depends mainly on the reverberation time of the room [55, 56]. This parameter, also referred to as T_{60}, expresses the reflection properties of the objects in the room, and is therefore frequency dependent [51, 52]. However, the current statistical methods rely on an AIR model formulated in the time domain, making it difficult to incorporate the frequency dependence of the reverberation phenomenon.

Another disadvantage of the time domain formulation is the inaccurate transition between the time and the STFT domains. Using a time domain AIR model implies the time-domain derivation of the reverberant component, where the transition to the STFT domain is performed only at the spectral enhancement stage. This transition neglects the crossband filters contribution which tends to be significant in the STFT representation, as was shown in [50]. In [62] the model was formulated in the STFT domain but used only the band-to-band filter to derive the reverberant component.

The last drawback of the current routine is that *a-priori* knowledge of T_{60} of the room is required for a reliable model. T_{60} can be measured as a preprocessing step from the energy decay curve (EDC) [52], or blindly estimated by one of the methods proposed in [64–66]. In [67], the estimation of T_{60} is performed via warped filter banks, in order to address frequency dependence of this parameter.

In this chapter we combine the system identification approach with the spectral enhancement method, to achieve an AIR representation in which both time and frequency

dependencies are addressed, and with which prior knowledge of system parameters is not required. The AIR and the reverberant signal are directly represented in the Short Time Fourier Transform (STFT) domain, and an estimate of the reverberant component of the signal is derived on that basis. Accurate and simplified formulations are presented, with their corresponding effect on performance.

The proposed algorithm is comprised of several stages. First, a "calibration" stage is performed. A known white noise signal is played in the reverberant environment, and a filter is extracted relating the recorded output signal to its reverberant component. Afterwards, the identified filter is applied to the recoded signal, in order to calculate its reverberant component. We approximate the signal representation in the STFT domain, by using only some or none of the crossband filters. The accuracy of these approximations is analysed in terms of the mean squared error (MSE) between the actual reverberant component and its estimated versions. In the last stage, this distorting component is suppressed by the spectral enhancement algorithm.

This chapter is organized as follows. In Section 5.2 we formulate the reverberation problem in the STFT domain. In Section 5.3, we address the problem of the reverberant path estimation, by using system approximation routine, with band-to-band and cross-band filters. Evaluation criteria are presented in Section 5.4. Experimental results of the reverberant component estimation and the dereverberation are shown in Section 5.5.

5.2 Reverberant Signal Representation in the STFT Domain

In this section we formulate the dereverberation problem directly in the STFT domain. Specifically, we derive the STFT domain expressions for the AIR and the recorded reverberant signal. Throughout this chapter, time domain variables are denoted by $x(n)$ and STFT domain variables are indicated by $X(l,k)$. $\mathbf{X}(k)$ notates a vector of a STFT domain signal in the k-th band, and $\mathbf{X}$ notates a matrix form.

The reverberant signal resulting from the convolution of the anechoic speech signal $x(n)$

and a causal AIR $h(n)$, is given by

$$y(n) = \sum_{n=0}^{\infty} h(n')x(n-n')\,. \tag{5.1}$$

We assume that $h(n)$ is time invariant and that its length is infinite. In the STFT domain the signal $x(n)$ is given by

$$X(l,k) = \sum_{n=0}^{\infty} x(n)\tilde{\psi}(n-lR)e^{-\frac{j2\pi}{N}k(n-lR)}\,, \tag{5.2}$$

where l is the frame index, k is the frequency band index, R is the discrete time shift, and $\tilde{\psi}$ denotes the analysis window of length N. According to [50] we can express $y(n)$ in the STFT domain as

$$Y(l,k) = \sum_{k'=0}^{N-1}\sum_{l'=-\infty}^{\infty} H(l',k,k')X(l-l',k')\,, \tag{5.3}$$

where k and k' denote the band-to-band and crossband frequency bin indices, respectively. The STFT response $H(l',k,k')$ is related to the impulse response $h(n)$ by

$$H(l,k,k') = \left[h(n) * \vartheta(n,k,k')\right]|_{n=lR} \tag{5.4}$$

and $\vartheta(n,k,k')$ is a function of the analysis and synthesis windows:

$$\vartheta(n,k,k') = e^{\frac{j2\pi}{N}k'n}\sum_{n'=-\infty}^{\infty}\tilde{\psi}(n')\psi(n'+n)e^{-\frac{j2\pi}{N}n'(k-k')}\,. \tag{5.5}$$

Let us describe the AIR, $h(n)$, as a summation of the direct path and the reverberant path:

$$h(n) = h_d(n) + h_r(n)\,, \tag{5.6}$$

where the direct path is given by

$$h_d(n) = \begin{cases} h(n) & \text{for } 0 \le n < n_d \\ 0 & \text{otherwise} \end{cases}\,, \tag{5.7}$$

and the reverberant path is given by

$$h_r(n) = \begin{cases} h(n) & \text{for } n \ge n_d \\ 0 & \text{otherwise} \end{cases}\,. \tag{5.8}$$

Substituting the expressions of the direct and reverberant paths in (5.4) yields

$$H_d(l,k,k') = [h_d(n) * \vartheta(n,k,k')]\,|_{n=lR}$$
$$H_r(l,k,k') = [h_r(n) * \vartheta(n,k,k')]\,|_{n=lR}\,. \tag{5.9}$$

From (5.6) and (5.9), the expression of $h(n)$ in the STFT domain can be written as

$$H(l,k,k') = H_d(l,k,k') + H_r(l,k,k')\,. \tag{5.10}$$

Using (5.10) and (5.3), the reverberant recorded signal, $Y(l,k)$ can be expressed as

$$Y(l,k) = Y_d(l,k)+Y_r(l,k) = \sum_{k'=0}^{N-1}\sum_{l'=-\infty}^{\infty} H_d(l',k,k')X(l-l',k')+\sum_{k'=0}^{N-1}\sum_{l'=-\infty}^{\infty} H_r(l',k,k')X(l-l',k')\,. \tag{5.11}$$

5.3 Reverberant Component Estimation

In this section we propose a method for estimation of the reverberant component by using the output signal. Prior knowledge of the enclosed space parameters- e.g. reverberation time and DRR- is not required. The estimation is carried out in two stages: On the first stage we approximate the system that relates the output signal and its reverberant component. The procedure is performed for a white noise source signal, to receive the most accurate expression for the system. On the second stage, the reverberant component is estimated by convolving the extracted system with a given recoded signal, under the assumption that the expression of the approximated system is independent of the recorded signal.

5.3.1 An expression for the reverberant component

The input white Gaussian noise signal $x(n)$ passes through an acoustic path described by an impulse response $h(n)$, whose reverberant part is $h_r(n)$, and results in a recorded output signal $y(n)$. The reverberant component of the output signal is given by

$$y_r(n) = h_r(n) * x(n)\,. \tag{5.12}$$

According to (5.11), the STFT representation of $y_r(n)$ may be written as

$$Y_r(l,k) = \sum_{k'=0}^{N-1} \sum_{l'} H_r(l',k,k') X(l-l',k') \,. \tag{5.13}$$

Let us assume that the direct path of the output signal is an attenuated version of the source signal, and that $H_r(l,k,k') = 0$ for $l \leq 0$. Then:

$$Y_d(l,k) = B_d(k) X(l,k) \,, \tag{5.14}$$

where $B_d(k)$ is zero-mean, identically distributed (i.i.d.) Gaussian random noise. From (5.11), and (5.14), the reverberant component can be expressed as:

$$\begin{aligned} Y_r(l,k) &= \sum_{k'=0}^{N-1} \sum_{l'=1}^{N_h-1} H_r(l',k,k') X(l-l',k') = \\ &= \sum_{k'=0}^{N-1} \sum_{l'=1}^{\infty} H_r(l',k,k') \frac{Y_d(l-l',k')}{B_d(k')} = \\ &= \sum_{k'=0}^{N-1} \sum_{l'=1}^{\infty} H_r(l',k,k') \frac{Y(l-l',k') - Y_r(l-l',k')}{B_d(k')} \,. \end{aligned} \tag{5.15}$$

5.3.2 Band-to-band system approximation

In this subsection, we find the expression of the band-to-band filter that relates the output signal and its reverberant component. For that matter, we assume that the output recorded signal, $y(n)$, and its reverberant component $y_r(n)$, are known. In practice $h(n)$ can be determined by recording the response to an impulse signal, played for a very short time. The reverberant component h_r, is then extracted from $h(n)$, using the relation in (5.8). Alternatively, h_r can be estimated using a statistical model, as was proposed in [43, 44].

Let $X(l,k)$ be the STFT transform of the input white Gaussian noise signal, and let $Y_r^{\mathbf{bb}}(l,k)$ be the approximation of $Y_r(l,k)$ using only the band-to-band filter, $H_r^{\mathbf{bb}}(l,k) \equiv H_r(l,k,k)$, with $H_r(l,k,k) = 0$ for $k' \neq k$. The expression in (5.15) can then be written as:

$$Y_r^{\mathbf{bb}}(l,k) = \frac{H_r(l,k)}{B_d(k)} * \left(Y(l,k) - Y_r^{\mathbf{bb}}(l,k) \right) \,. \tag{5.16}$$

Hence,

$$Y_r^{\mathbf{bb}}(l,k) * \left(\delta(l) + \frac{H_r^{\mathbf{bb}}(l,k)}{B_d(k)} \right) = \frac{H_r^{\mathbf{bb}}(l,k)}{B_d(k)} * Y(l,k) \,. \tag{5.17}$$

Let us define $\text{inv}\{F(l)\}$ as the inverse filter of $F(l)$, i.e.,

$$\text{inv}\{F(l)\} * F(l) = \delta(l)\,. \tag{5.18}$$

Using the above notation, (5.17) can be expressed as

$$Y_r^{\mathbf{bb}}(l,k) = \text{inv}\left\{B_d(k)\delta(l) + H_r^{\mathbf{bb}}(l,k)\right\} * H_r^{\mathbf{bb}}(l,k) * Y(l,k)\,. \tag{5.19}$$

Let us denote the relation between the output signal $Y(l,k)$ and its approximate reverberant component $Y_r^{\mathbf{bb}}(l,k)$ as $\tilde{H}_r^{\mathbf{bb}}(l,k)$, where

$$\tilde{H}_r^{\mathbf{bb}}(l,k) = \text{inv}\left\{B_d(k)\delta(l) + H_r^{\mathbf{bb}}(l,k)\right\} * H_r^{\mathbf{bb}}(l,k)\,. \tag{5.20}$$

The resultant band-to-band approximation of reverberant component in terms of the above definition is given by

$$Y_r^{\mathbf{bb}}(l,k) = \tilde{H}_r^{\mathbf{bb}}(l,k) * Y(l,k) = \sum_{l'=0}^{N_h-1} \tilde{H}_r^{\mathbf{bb}}(l',k) Y^{\mathbf{bb}}(l-l',k)\,, \tag{5.21}$$

where we assume that $\tilde{H}_r^{\mathbf{bb}}(l,k)$ is a causal filter of length N_h. It should be noted, that theoretically the filter $\tilde{H}_r^{\mathbf{bb}}$ can be of infinite length, since the filter $H_r^{\mathbf{bb}}$ might be infinite. However, since $H_r^{\mathbf{bb}}$ approximately describes an exponential decay [55,56], after a certain length, the addition coefficients can be neglected. Therefore, one can assume that $H_r^{\mathbf{bb}}$ and, consequently, $H_r^{\mathbf{bb}}$ are of finite length.

Next we approximate the system $\tilde{H}_r^{\mathbf{bb}}(l,k)$, by using the system identification routine in the STFT domain. Since the theoretical value of the band-to-band approximation of the reverberant component, $Y_r^{\mathbf{bb}}$, is not available, we replace it with the real, full-band, value, Y_r. Let $\tilde{H}_r^{\mathbf{bb}}(k)$ denote the band-to-band filter at frequency band k

$$\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{bb}}(k) = \left[\begin{array}{cccc} \tilde{H}_r^{\mathbf{bb}}(0,k) & \tilde{H}_r^{\mathbf{bb}}(1,k) & \cdots & \tilde{H}_r^{\mathbf{bb}}(N_h-1,k) \end{array}\right]^T \tag{5.22}$$

and let

$$\mathbf{Y_k} = \left[\begin{array}{ccccc} Y(0,k) & 0 & \cdots & \cdots & 0 \\ Y(1,k) & Y(0,k) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y(N_y-1,k) & \cdots & \cdots & \cdots & Y(N_y+N_h-2,k) \end{array}\right] \tag{5.23}$$

represent a $N_y \times N_h$ Toeplitz matrix constructed from the output signal STFT coefficients of the k-th frequency-band, where N_y is the length of the output signal $Y(l,k)$ in frequency

band k. Notice that in order to receive the above matrix, $Y(l,k)$ should be padded with zeros from the index $Y(N_y,k)$ up to the index $Y(N_y+N_h-2,k)$. Using the above notations, the LS optimization problem can be expressed as

$$\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{bb}}(k) = \arg\min_{\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{bb}}(k)} \left\| \mathbf{Y_r}(k) - \mathbf{Y_k}\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{bb}}(k) \right\|^2 . \tag{5.24}$$

where

$$\mathbf{Y_r}(k) = \left[\begin{array}{ccccc} Y_r(0,k) & Y_r(1,k) & Y_r(2,k) & \cdots & Y_r(N_y-1,k) \end{array} \right]^T . \tag{5.25}$$

The solution to (5.24) is given by

$$\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{bb}}(k) = \left(\mathbf{Y_k}^H \mathbf{Y_k}\right)^{-1} \mathbf{Y_k}^H \mathbf{Y_r}(k) , \tag{5.26}$$

where we assume that $\mathbf{Y_k}^H\mathbf{Y_k}$ is not singular. In the ill-conditioned case, matrix regularization is required [68].

5.3.3 System approximation with 2K crossband filters

In this subsection we identify the filter that relates the output signal to its reverberant component, using $2K$ crossband filters around the main frequency band k. From (5.15), the expression of the reverberant component using $2K$ crossbands (in addition to the band-to-band filter) is given by

$$Y_r^{\mathbf{2K}}(l,k) = \sum_{k'=k-K}^{k+K} \sum_{l'=1}^{\infty} H_r^{\mathbf{2K}}(l',k,k') \frac{Y(l-l',k') - Y_r^{\mathbf{2K}}(l-l',k')}{B_d(k')} . \tag{5.27}$$

Similarly to (5.20), let us define the filter that relates the output signal to its approximate reverberant component using $2K$ crossbands as $\tilde{H}_r^{\mathbf{2K}}(l,k)$. Following this notation, (5.27) can be expressed as

$$Y_r^{\mathbf{2K}}(l,k) = \sum_{k'=k-K}^{k+K} \sum_{l'=1}^{\infty} \tilde{H}_r^{\mathbf{2K}}(l',k,k' \bmod N) Y(l-l',k' \bmod N) . \tag{5.28}$$

Writing $\tilde{H}_r^{\mathbf{2K}}(l',k,k' \bmod N)$ in vector form yields

$$\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{2K}}(k) = \left[\begin{array}{cccc} \tilde{\mathbf{H}}_r^T(k,(k-K) \bmod N) & \tilde{\mathbf{H}}_r^T(k,(k-K+1) \bmod N) & \cdots & \tilde{\mathbf{H}}_r^T(k,(k+K) \bmod N) \end{array} \right]^T . \tag{5.29}$$

Let $\mathbf{\Delta_k}$ be a concatenation of the matrices $\{\mathbf{Y_{k'}}\}_{k'=(k-K) \bmod N}^{(k+K) \bmod N}$ along the column dimension

$$\mathbf{\Delta_k} = \left[\begin{array}{cccc} \mathbf{Y}_{(\mathbf{k-K}) \bmod \mathbf{N}} & \mathbf{Y}_{(\mathbf{k-K+1}) \bmod \mathbf{N}} & \cdots & \mathbf{Y}_{(\mathbf{k+K}) \bmod \mathbf{N}} \end{array} \right] . \tag{5.30}$$

The LS optimization problem for the estimation of $\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{2K}}(k)$ from the given signals $\mathbf{Y_r(k)}$ and $\mathbf{Y(k)}$ can be expressed as

$$\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{2K}}(k) = \arg \min_{\tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{2K}}(k)} \left\| \mathbf{Y_r}(k) - \mathbf{\Delta_k} \tilde{\mathbf{H}}_{\mathbf{r}}^{\mathbf{2K}}(k) \right\|^2 . \tag{5.31}$$

The solution to (5.24) is given by

$$\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{2K}}(k) = \left(\mathbf{\Delta_k}^H \mathbf{\Delta_k} \right)^{-1} \mathbf{\Delta_k}^H \mathbf{Y_r}(k) . \tag{5.32}$$

5.3.4 Reverberant component estimation from a recorded signal

After the performing the procedure described above, we can use the approximated filter $\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}(k)$, to estimate the reverberant component of a new signal, recorded in the same environment that characterizes the identification stage. In the following discussion we refer only to the $2K$ approximation, since the band-to-band filter is its private case, with $K = 0$.

The procedure of estimating the reverberant component of a given recorded signal, and using it for dereverberation algorithm, can be described as follows:

1. Measure an impulse response $h(r)$ of a given environment
2. Create a white noise source signal $x(n)$ and record its output signal $y(n)$.
3. Compute the STFT domain expressions - $Y(l,k)$, and $Y_r(l,k)$, using h_r.
4. Define the number of crossbands K and the filter length N_h for the identification procedure.
5. Obtain the filter $\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{2K}}(k)$ that relates the output signal and its reverberant component, via the LS optimization problem.
6. Given a new recorded signal, apply on it the estimated filter $\hat{\tilde{\mathbf{H}}}_{\mathbf{r}}^{\mathbf{2K}}(k)$, to obtain its approximated reverberant component, $\hat{Y}_r^{\mathbf{2K}}$.

7. Compute the spectral variance of the reverberant component, given by

$$\lambda_r^{\mathbf{2K}}(l,k) = \eta|\hat{Y}_r^{\mathbf{2K}}(l,k)|^2 + (1-\eta)\lambda_r^{\mathbf{2K}}(l-1,k) \tag{5.33}$$

   where η is the smoothing factor.

8. Use the estimated reverberant spectral variance in a spectral enhancement algorithm, to obtain a dereverberated signal.

5.4 Performance Criteria

5.4.1 Estimation of the reverberant component

Mean Spectral Variance For a qualitative evaluation of the estimation, we average the power spectral density (PSD), also called spectral variance of the reverberant component at a given time frame, over all the frequency bins:

$$\bar{\lambda}_r(l) = 10\log_{10}\left(\text{mean}_k\left\{\lambda_r(l,k)\right\}\right) \tag{5.34}$$

This calculation is performed for the estimated signals, and is compared to the "true" reverberant component.

Mean Squared Error (MSE) For quantitative evaluation we calculated the MSE between the true and the estimated reverberant component, where the estimation was performed using a varying number of crossbands and different lengths of the filter $\tilde{H}_r$.

$$\varepsilon(K,N_h) = \frac{E\left\{\left\|Y_r - \hat{Y}_r^{\mathbf{2K}}\right\|^2\right\}}{E\left\{\|Y_r\|^2\right\}} \tag{5.35}$$

5.4.2 Dereverberation

The performance of the dereverberation algorithm is evaluated using the mean segmental signal-to-reverberation ratio (SRR) and the mean log spectral distortion (LSD) measures.

Signal to Reverberation Ratio (SRR) The instantaneous segmental SRR of the m-th frame is defined as

$$\mathrm{SRR}_{\mathrm{seg}}(l) = 10\log_{10}\left(\frac{\sum_{n=lR}^{lR+N-1} z_d^2(n)}{\sum_{n=lR}^{lR+N-1}\left(z_d(n)-\hat{z}_d(n)\right)^2}\right) \tag{5.36}$$

where N is the frame length, R is the frame, $\hat{z}_d^2(n)$ is the enhanced signal, and $z_d^2(n)$ is the normalized direct signal. The normalization is performed by applying on the original direct signal the full dereverberation algorithm- using the same set of parameters as was used for the reverberant signal. The mean segmental SRR is then obtained by averaging the segmental SRR over all frames.

Log Spectral Distortion (LSD) LSD is defined as the Lp norm of the difference between the STFT transforms of $y_d(n)$ and $\hat{y}_d(n)$, namely $Y_d(l,k)$ and $\hat{Y}_d(l,k)$, in the l-th frame:

$$\mathrm{LSD}(l) = \left(\frac{1}{K}\sum_{k=0}^{K-1}\left|\mathcal{L}\left\{\hat{Y}_d(l,k)\right\} - \mathcal{L}\left\{Y_d(l,k)\right\}\right|^2\right)^{\frac{1}{2}} \ \mathrm{dB} \tag{5.37}$$

where $\mathcal{L}\left\{X(l,k)\right\} = \max\left\{20\log_{10}\left(|X(l,k)|\right),\delta\right\}$ is the log spectrum confined to about 50 dB dynamic range ($\delta = \max_{l,k}\left\{20\log_{10}\left(|X(l,k)|\right)\right\} - 50$). The mean LSD is obtained by averaging (5.37) over all frames containing speech.

5.5 Experimental Results

The signals used in the simulations include synthetic white Gaussian noise as well as real speech signals, which are taken from the TIMIT database. The sample frequency is $f_s = 16$ [kHz]. The STFT synthesis window is a Hamming window and the analysis window is the related bi-orthogonal window. The windows are of length $N = 256$ and the overlap between two successive STFT frames is 50 percent. The AIR is simulated according to the method proposed in [63], with room dimensions of $6 \times 8 \times 5$ [m].

5.5.1 Reverberant component estimation

In this subsection we show the results of the reverberant component estimation using system identification with crossband filters. First we focus on the band-to-band identification

results. Figure 5.1 shows the real and the estimated band-to-band spectral variances of the reverberant component as a function of time frames, in three frequency bands, whose central frequencies are $f = 635$ [Hz], $f = 1375$ [Hz], and $f = 1875$ [Hz] . The spectral variances are found by smoothing the PSD along the time frames. The smoothing factor is equal to $\alpha = 0.9$. Figure 5.2 shows the real and estimated spectral variances, averaged over all the frequency bands. The two estimated curves denote two options for choosing the length of the identified filter $\tilde{H}_r$: one curve corresponds to $N_h = 1$ and the other curve corresponds to $N_h = 61$, which, for our set of parameters, is the number of coefficients to be used for a full STFT representation.
The results show that the estimation is relatively accurate even without using crossband filters. By using a full representation of the output signal and of its reverberant component in the identification stage, we "force" the system to be as similar to the full representation as possible. However, as we can see from figure 5.2, using too few coefficients in the filter results in underestimation. Thus the filter length should be sufficiently long.

Figure 5.4 shows the reverberant component estimation using $2K$ crossband filters, corresponding to using K crossband filters at each side of the main frequency band k. The figure exhibits several trends. First of all, it can be seen that using no crossbands at all yields the best results for any selection of filter length N_h. Secondly, the filter length has a different effect for various values of K: for small K's ($0 \leq K \lessapprox 2$), estimation improves if N_h increases, whereas, for $K \geq 2$, it is better to choose a shorter filter. A plausible explanation for the difference could be that the system identification assumes that the identified transfer function $\tilde{H}_r$ is independent of the input and output signals. As we have shown in the mathematical formulation, this assumption is most accurate for the band-to-band case, in which a long filter results in a more accurate estimation, as it allows us to describe the system with higher resolution. When we increase the number of crossbands, the model becomes less accurate, and thus, the long filter length adds higher error to the estimation.

5.5.2 Dereverberation results

In this subsection we show the results of dereverberation algorithm, where we incorporated our estimator of reverberant component. The spectral enhancement stage, was performed

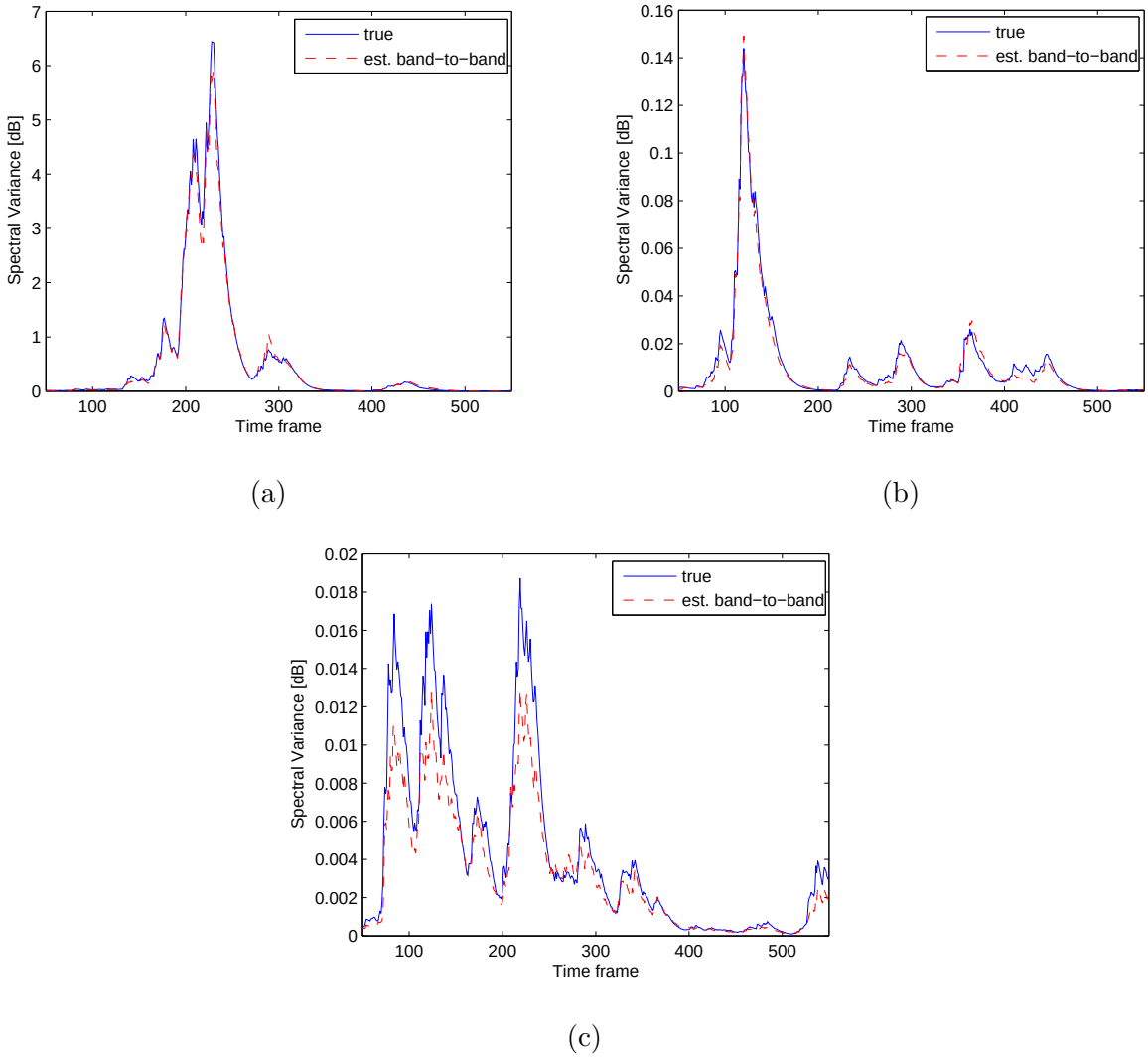


Figure 5.1: Spectral variances of the true reverberant component and its band-to-band estimation, at three frequency bands. (a) $f = 635$ [Hz], (b) $f = 1375$, (c) $f = 1875$ [Hz].

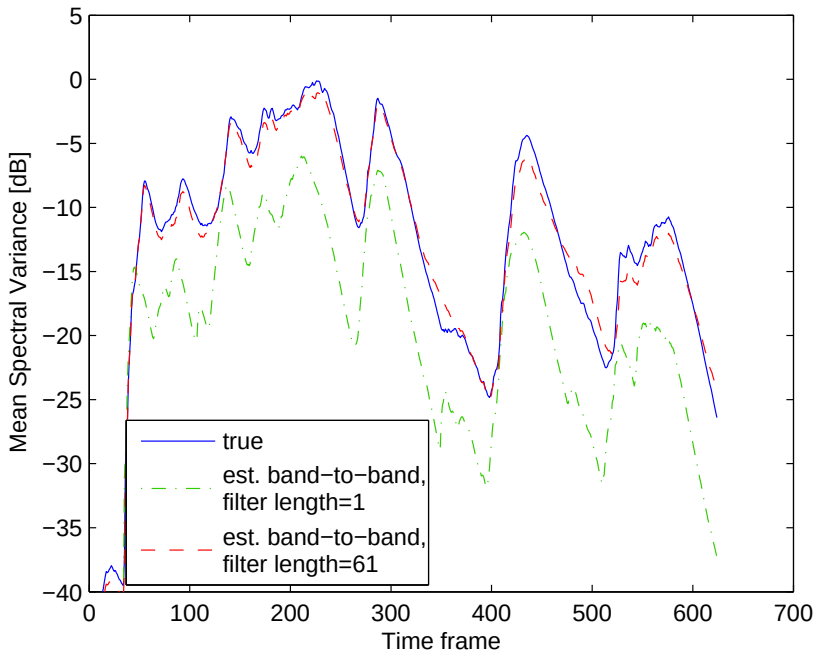


Figure 5.2: Mean spectral variance over all the frequency bands of the true reverberant component and its band-to-band estimation.

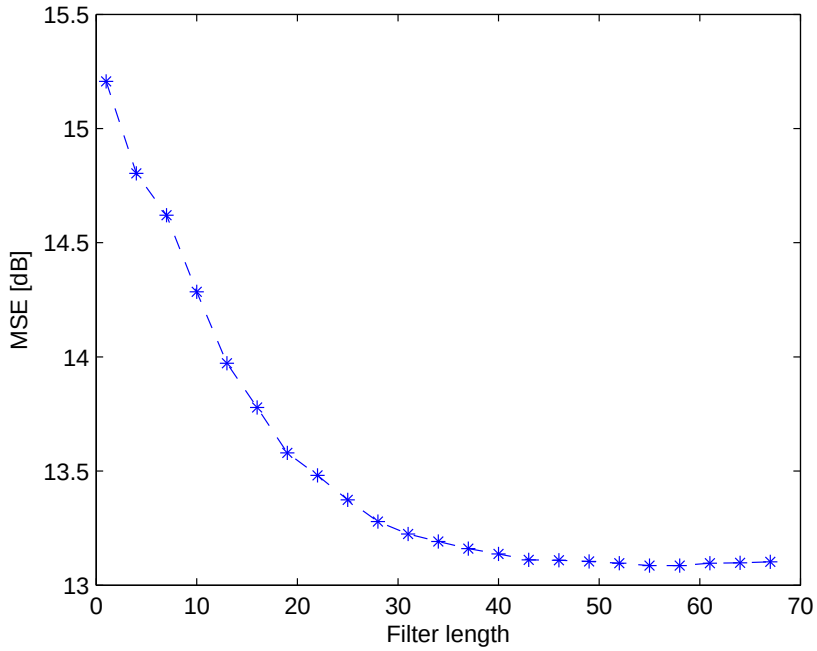


Figure 5.3: Mean square error between the true reverberant component and its band-to-band estimation, as a function of the filter length.

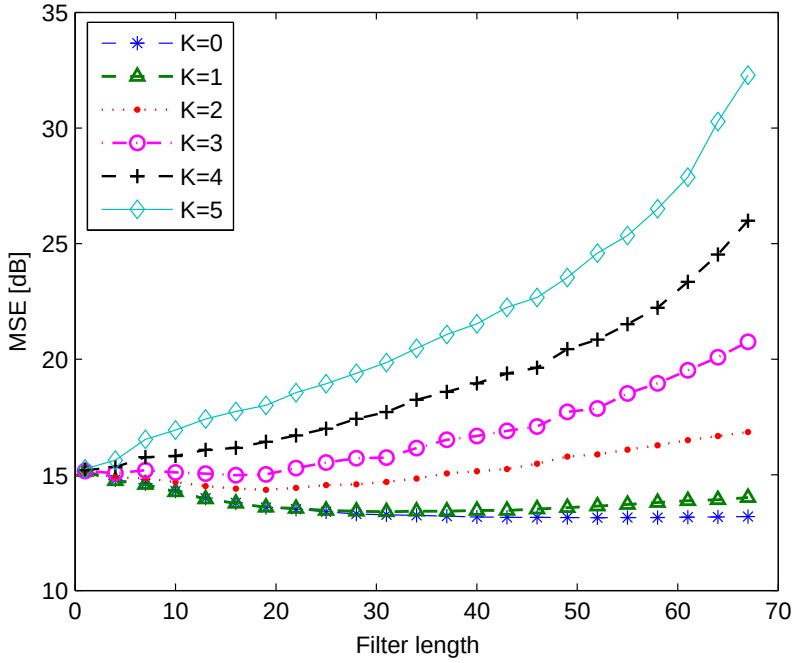
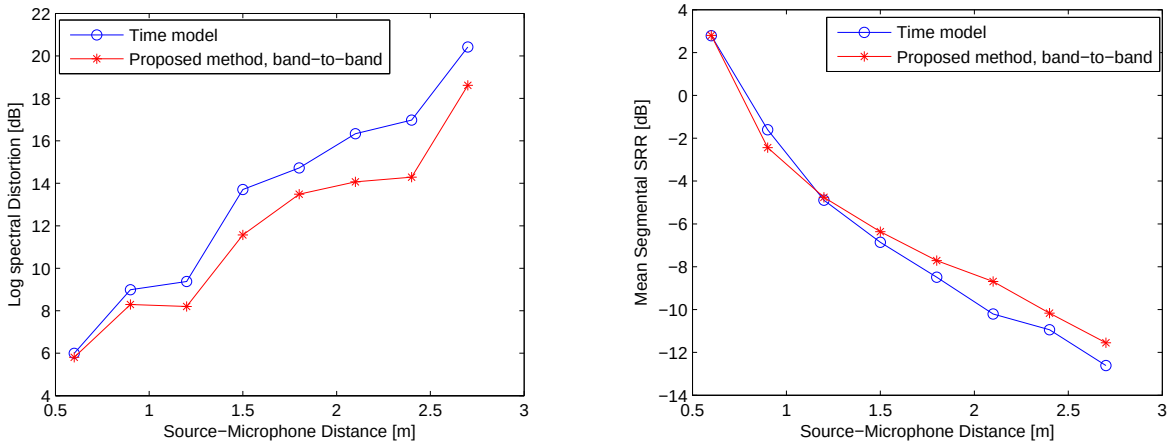


Figure 5.4: MSE between the real and estimated reverberant component as a function of the length of the filter $\tilde{H}_r(k,k')$. The six curves correspond to six different K's.

using the OM-LSA method [45], which receives as an input the recorded speech and the estimated reverberant component, and derives the post-filter that suppresses the reverberation effect. In both simulations described below, eight speakers were chosen from the TIMIT database, four males and four females. Each speaker was placed by the simulation in several positions in the room which corresponded to source-microphone distances of $0.6 - 2.7$ [m]. The evaluation measures- SRR and LSD, were averaged along all the received signals and depicted in Figures 5.5 and 5.6. The performance of the proposed estimator is compared to that of the generalized statistical model [44], termed here as "time model".

First, we consider a scenario where the parameters of the system are known. It should be noted here that the generalized model requires the knowledge of an additional parameter, κ which is related to the Direct to Reverberation Ratio (DRR) value. This parameter depends mainly on the source-to-microphone distance and on T_{60}, and is computed here from the output of the RIR generator [63]. The T_{60} value was set to 0.5 [sec]. The results of the simulation are depicted in figure 5.5. As can be seen, the performance of the



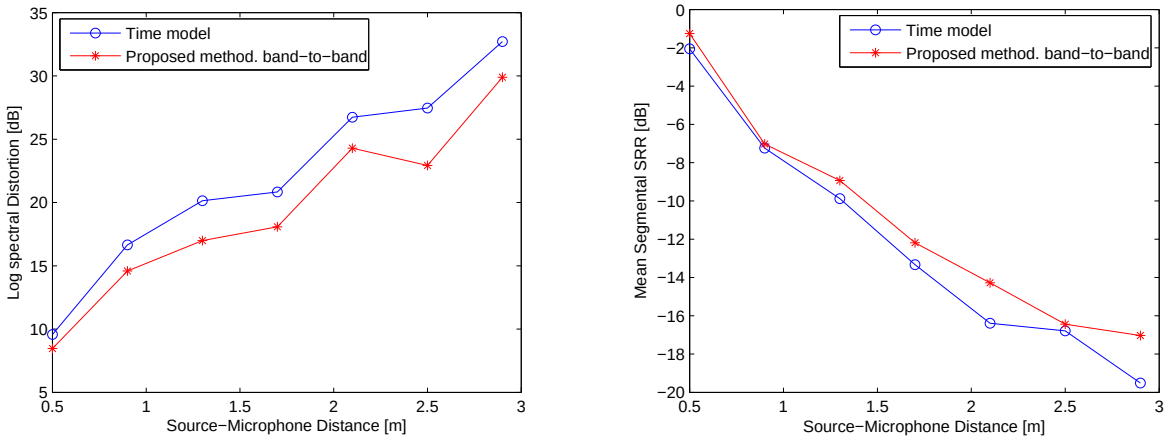
(a) (b)

Figure 5.5: The mean LSD (a) and SRR (b) as a function of source-microphone distance, under the assumption that the parameters of the model (T_{60} and κ) are known.

band-to-band version of the new method is slightly better than the existing time domain method for both SRR and LSD measures. This is in spite of our assumption that the parameters used in the generalized statistical model are known and accurate. However, this assumption might not hold for real systems, and thus our second simulation assumes that the parameters are frequency-dependent and unknown. The T_{60} parameter had seven values: [1.2 1 0.8 0.6 0.5 0.4 0.3] [sec] - each value corresponding to a different frequency region, equally divided at the range of $0-8$ [kHz]. For the generalized statistical model, we assumed that the T_{60} was equal to the average value over the frequency bands, $T_{60} = 0.69$ [sec]. The parameter κ was also an average of the real frequency-dependent κ. In both simulations of the generalized statistical method, the time when the late reflections begin was set to 48 [msec]. Figure 5.6 clearly demonstrates that for that case the difference in performance is much more significant. The identification routine of our method enables to follow the frequency changes and to find a more accurate estimator, while the performance of the time model is considerably lower.

5.6 Conclusions

In this chapter we have addressed the dereverberation problem, incorporating system identification in the STFT domain, in spectral enhancement method. The identification



(a) (b)

Figure 5.6: The mean LSD (a) and SRR (b) as a function of source-microphone distance, when the parameters of the model (T_{60} and κ) are unknown and frequency dependent.

stage allowed us to estimate the reverberant component of the recorded speech without any prior knowledge of system parameters, whereas, performing it in the STFT domain addressed the frequency dependency of the reflection coefficients in the derived estimator.

First, we derived an expression of the AIR and the recorded signal directly in the STFT domain. Based on that formulation, we performed system identification with crossband filters, in order to find the expression of the distorting filter, and of the reverberant component of the recorded speech. We tested the accuracy of the estimated reverberant component as a function of the number and length of crossband filters used in the identification step.

The performance of the system identification stage was analysed in terms of the mean-square error (MSE) between the actual reverberant component and its estimate by the approximate representations. It was shown that the smallest MSEs are achieved by using the band-to-band filter, due to the fact that the mathematical expression that relates the output signal to its reverberant component is relatively simple in those cases. Also, it was found that for a small number of crossbands, it is advantageous to increase the lengths of the filters.

In the last stage, we used the reverberant component estimate in the spectral enhancement algorithm for dereverberation. We measured the performance for different reverberant environments and for various distances between the source and the micro-

phone. We showed that our method achieved better results than the existing statistical method, especially for the cases where the models' parameters were unknown.

Chapter 6

Conclusion

6.1 Summary

In this thesis we addressed the problem of dereverberation of speech signals, acquired in an enclosed space by a single microphone. Dereverberation was approached with two time-frequency representations: The short-time Fourier transform (STFT), and the Single-Side-Band transform (SSB).

For the SSB transform, we formulated the SSB-domain filter that relates the input and output signals, and developed an identification routine for its band-to-band approximation. We showed that as the SNR improves, the identified system becomes almost similar to the real system, in spite of the fact that no crossbands were included. Then we performed dereverberation in the SSB domain, and concluded that the real-valued representation is less appropriate for the dereverberation application, due to a distortion in the signals' amplitude, caused by the non-uniform phase delay.

Following that, we concentrated on the STFT representation, where we proposed a method for estimating the reverberant component. The proposed approach requires no prior knowledge regarding system parameters, and allows one to take into account the frequency dependency of the acoustic properties of the environment.

First, we derived an expression of the AIR and the recorded signal directly in the STFT domain. Based on that formulation, we performed system identification with crossband filters, in order to find the expression of the distorting filter, and of the reverberant component of the recorded speech. We tested the accuracy of the estimated reverber-

ant component as a function of the number and length of crossband filters used in the identification step.

Our experimental results show, that using only the band-to-band filter, with the maximal filter length, yields the best results. Correspondingly, the theoretical derivation shows that the mathematical model is most accurate for that case. Finally, we use the derived estimator in the spectral enhancement method for dereverberation, and show that, our method gives better performance in comparison with the existing statistical method described in [44].

6.2 Future Research

The aspects that were discussed in this thesis can be further explored in the following aspects:

Identification in the SSB domain Currently, the identification was confined to the band-to-band filter. A full identification problem can be formed, which will allow changing the number of crossband filters, and exploring the relationship between the SNR conditions, the length of the input signal, and the number of optimal crossbands that should be taken. Moreover, in the current work we have not analysed the effect of the crosstime samples on the representation of the system. Reducing the amount of crosstime samples in a similar way to the crossbands, could significantly reduce the complexity of the model and of the identification problem.

SSB-adjusted post filtering As a part of the dereverberation procedure, we applied the OM-LSA speech enhancement post-filter, on the reverberated signal. This method was developed for the STFT representation, and thus performs the estimation of the a-priori SIR on the signals' log-amplitude, leaving the phase unmodified. Applying the same post-filtering to a real-valued signal, might distort its amplitude due to the information from the multiple reflections with different delays. Therefore, it would be interesting to explore other ways of conducting the post-filtering, which would be adjusted for the real-valued representation.

Estimation of frequency dependent T_{60} Reverberation time is a crucial parameter in the modelling of the acoustic transfer function, and its correct estimation can strongly affect the performance of the dereverberation algorithm. Future research can focus on exploring the typical frequency dependency of T_{60} in a given space. Namely, given the T_{60} of a specific frequency band, devising an ability to estimate its value in other frequency bands, based on known parameters of the enclosed space.

Bibliography

[1] H. Kuttruff, *Room Acoustics, 4th edn.* Taylor and Frances, 2000.

[2] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least-squares approach to blind channel identification," *Signal Processing, IEEE Transactions on*, vol. 43, no. 12, pp. 2982 –2993, dec 1995.

[3] M. Gurelli and C. Nikias, "Evam: an eigenvector-based algorithm for multichannel blind deconvolution of input colored signals," *Signal Processing, IEEE Transactions on*, vol. 43, no. 1, pp. 134 –149, jan 1995.

[4] Y. A. Huang and J. Benesty, "Adaptive multi-channel least mean square and newton algorithms for blind channel identification," *Signal Process.*, vol. 82, pp. 1127–1138, August 2002. [Online]. Available: http://dl.acm.org/citation.cfm?id=607012.607020

[5] Y. Huang and J. Benesty, "A class of frequency-domain adaptive approaches to blind multichannel identification," *Signal Processing, IEEE Transactions on*, vol. 51, no. 1, pp. 11 – 24, jan. 2003.

[6] S. Subramaniam, A. Petropulu, and C. Wendt, "Cepstrum-based deconvolution for speech dereverberation," *Speech and Audio Processing, IEEE Transactions on*, vol. 4, no. 5, pp. 392 –396, sep 1996.

[7] A. Petropulu and C. Nikias, "Blind convolution using signal reconstruction from partial higher order cepstral information," *Signal Processing, IEEE Transactions on*, vol. 41, no. 6, pp. 2088 –2095, jun 1993.

[8] M. Triki and D. Slock, "Delay and predict equalization for blind speech dereverberation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 5, may 2006, p. V.

[9] J. Hopgood and P. Rayner, "Blind single channel deconvolution using nonstationary signal processing," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, no. 5, pp. 476 – 488, sept. 2003.

[10] C. Evers and J. Hopgood, "Parametric modelling for single-channel blind dereverberation of speech from a moving speaker," *Signal Processing, IET*, vol. 2, no. 2, pp. 59 –74, june 2008.

[11] C. Evers, J. Hopgood, and J. Bell, "Acoustic models for online blind source dereverberation using sequential monte carlo methods," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 4597 –4600.

[12] S. T. Neely and J. B. Allen, "Invertibility of a room impulse response," *The Journal of the Acoustical Society of America*, vol. 66, no. 1, pp. 165–169, 1979. [Online]. Available: http://link.aip.org/link/?JAS/66/165/1

[13] J. Mourjopoulos, P. Clarkson, and J. Hammond, "A comparative study of least-squares and homomorphic techniques for the inversion of mixed phase signals," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82.*, vol. 7, may 1982, pp. 1858 – 1861.

[14] B. Radlovic and R. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 728 –737, nov 2000.

[15] Peter and Vary, "An adaptive filter-bank equalizer for speech enhancement," *Signal Processing*, vol. 86, no. 6, pp. 1206 – 1214, 2006, ¡ce:title¿Applied Speech and Audio Processing¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168405003221

[16] Y. Huang, J. Benesty, and J. Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 882–895, Sept. 2005.

[17] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 36, no. 2, pp. 145 –152, feb 1988.

[18] P. Nelson, F. Orduna-Bustamante, and H. Hamada, "Inverse filter design and equalization zones in multichannel sound reproduction," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 3, pp. 185 –192, may 1995.

[19] D. M. M. M. Hikichi, T., "On robust inverse filter design for room transfer function fluctuations." in *Proc. European Signal Processing Conf. (EUSIPCO)*, 2006.

[20] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *ASSP Magazine, IEEE*, vol. 5, no. 2, pp. 4 –24, april 1988.

[21] M. Brandstein and D. Wards, *Microphone Arrays, signal processing techniques and applications*. 1 edn. Springer, 2001.

[22] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 34, pp. 229 – 240, 1996, ¡ce:title¿Acoustic Echo Control and Speech Enhancement Techniques¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016763939600057X

[23] G. W. and Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol. 20, no. 3-4, pp. 229 – 240, 1996, ¡ce:title¿Acoustic Echo Control and Speech Enhancement Techniques¡/ce:title¿. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S016763939600057X

[24] S. Haykin, *Adaptive filter theory (2nd ed.)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1991.

[25] J. Flanagan, A. Surendran, and E. Jan, "Spatially selective sound capture for speech and audio processing," *Speech Communication*, vol. 13, no. 12, pp. 207

– 222, 1993. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167639393900072S

[26] T. Nishiura, S. Nakanura, and K. Shikano, "Speech enhancement by multiple beamforming with reflection signal equalization," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 189 –192 vol.1.

[27] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 5, pp. 425 –437, sep 1997.

[28] Y. Grenier and S. Affes, "Microphone array response to speaker movements," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 1, apr 1997, pp. 247 –250 vol.1.

[29] E. Jan and J. Flanagan, "Microphone arrays for speech processing," in *Signals, Systems, and Electronics, 1995. ISSSE '95, Proceedings., 1995 URSI International Symposium on*, oct 1995, pp. 373 –376.

[30] E. Jan, P. Svaizer, and J. Flanagan, "Matched-filter processing of microphone array for spatial volume selectivity," in *Circuits and Systems, 1995. ISCAS '95., 1995 IEEE International Symposium on*, vol. 2, apr-3 may 1995, pp. 1460 –1463 vol.2.

[31] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "On the use of linear prediction for dereverberation of speech," in *In Proceedings of the IEEE International Workshop on Acoustic Echo and Noise Control*, 2003, pp. 99–102.

[32] S. P. Yegnanarayana, B., "Enhancement of reverberant speech using lp residual signal." *IEEE Trans. Speech Audio Process.*, p. 267281, 2000.

[33] S. Griebel and M. Brandstein, "Microphone array speech dereverberation using coarse channel modeling," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, vol. 1, 2001, pp. 201 –204 vol.1.

[34] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, may 2002, pp. I–541 –I–544.

[35] B. Gillespie, H. Malvar, and D. Florencio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 6, pp. 3701–3704, 2001.

[36] M. Wu and D. Wang, "A two-stage algorithm for one microphone reverberant speech enhancement," 2006. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.125.3297

[37] T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, and M. Miyoshi, "Importance of energy and spectral features in gaussian source model for speech dereverberation," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, oct. 2007, pp. 299 –302.

[38] T. Nakatania, T. Yoshiokaa, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation in short time fourier transform domain with crossband effect compensation," in *Hands-Free Speech Communication and Microphone Arrays, 2008. HSCMA 2008*, may 2008, pp. 220 –223.

[39] N. Gaubitch and P. Naylor, "Spatiotemporal averaging method for enhancement of reverberant speech," in *Digital Signal Processing, 2007 15th International Conference on*, july 2007, pp. 607 –610.

[40] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," 2004. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.3111

[41] M. R. P. Thomas, N. D. Gaubitch, J. Gudnason, and P. A. Naylor, "A practical multichannel dereverberation algorithm using multichannel dypsa and spatiotemporal averaging," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, oct. 2007, pp. 50 –53.

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

[43] K. Lebart, J. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[44] E. A. P. Habets, "Single and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, 2007.

[45] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Processing Letters*, vol. 9, pp. 113–116, Apr. 2002.

[46] H. W. Löllmann and P. Vary, "Low delay noise reduction and dereverberation for hearing aids," *EURASIP J. Adv. Signal Process*, vol. 2009, pp. 1:1–1:9, January 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/437807

[47] M. Jeub, M. Schäfer, T. Esch, and P. Vary, "Model-based dereverberation preserving binaural cues," *Trans. Audio, Speech and Lang. Proc.*, vol. 18, no. 7, pp. 1732–1745, Sep. 2010. [Online]. Available: http://dx.doi.org/10.1109/TASL.2010.2052156

[48] E. Habets, N. Gaubitch, and P. Naylor, "Temporal selective dereverberation of noisy speech using one microphone," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-april 4 2008, pp. 4577 –4580.

[49] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing," *Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 2, pp. 253–266, Feb. 2009. [Online]. Available: http://dx.doi.org/10.1109/TASL.2008.2008039

[50] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[51] M. R. Schroeder and K. H. Kuttruff, "On frequency response curves in rooms. comparison of experimental, theoretical, and monte carlo results for the average frequency spacing between maxima," vol. 34, no. 1, pp. 76–80, 1962.

[52] M. R. Schroeder, "New method of measuring reverberation time," *Acoustical Society of America Journal*, vol. 37, pp. 1187–+, 1965.

[53] B. Radlovic and R. Kennedy, "Nonminimum-phase equalization and its subjective importance in room acoustics," *Speech and Audio Processing, IEEE Transactions on*, vol. 8, no. 6, pp. 728 –737, nov 2000.

[54] F. Talantzis and D. B. Ward, "Robustness of multichannel equalization in an acoustic reverberant environment," *The Journal of the Acoustical Society of America*, vol. 114, no. 2, pp. 833–841, 2003. [Online]. Available: http://link.aip.org/link/?JAS/114/833/1

[55] M. R. Schroeder, "Frequency-correlation functions of frequency responses in rooms," *The Journal of the Acoustical Society of America*, vol. 34, no. 12, pp. 1819–1823, 1962.

[56] J. Polack, "La transmission de lenergie sonore dans les salles," Ph.D. dissertation, 1988.

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

[58] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, pp. 12–15, Jan. 2002.

[59] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, May 2006.

[60] J. Benesty, T. Gnsler, D. R. Morgan, M. M. Sondhi, and G. S. L., *Advances in Network and Acoustic Echo Cancellation*. Springer, 2001.

[61] S. Gannot and I. Cohen, "Speech enhancement based on the general transfer function GSC and postfiltering," *IEEE Trans. Speech and Audio Processing*, vol. 12, no. 6, pp. 561–571, Nov. 2004.

[62] E. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, Sept. 2009.

[63] E. Habets, "Room impulse response (RIR) generator," May 2008.

[64] V. P. Lollmann, H.W., "Estimation of the reverberation time in noisy environments," in *Int. Workshop Acoust. Echo Noise Control (IWAENC)*, 2008.

[65] F. D. P. Cox, Trevor J.; Li, "Extracting room reverberation time from speech using artificial neural networks," *J. Audio Eng. Soc*, vol. 49, no. 4, pp. 219–230, 2001.

[66] R. Ratnam, D. L. Jones, B. C. Wheeler, J. William D. O'Brien, C. R. Lansing, and A. S. Feng, "Blind estimation of reverberation time," *The Journal of the Acoustical Society of America*, vol. 114, no. 5, pp. 2877–2892, 2003.

[67] H. Lollmann and P. Vary, "Estimation of the frequency dependent reverberation time by means of warped filter-banks," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, may 2011, pp. 309 –312.

[68] A. Neumaier, "Solving ill-conditioned and singular linear systems: A tutorial on regularization," *SIAM Rev.*, vol. 40, no. 3, pp. 636–666, Sep. 1998.