

# APPLIED METHODS FOR BLIND SPEECH ENHANCEMENT

Benny Sällberg

Blekinge Institute of Technology  
Doctoral Dissertation Series No. 2008:15  
School of Engineering





# **Applied Methods for Blind Speech Enhancement**

Benny Sällberg



Blekinge Institute of Technology Doctoral Dissertation Series

No 2008:15

ISSN 1653-2090

ISBN 978-91-7295-154-9

# **Applied Methods for Blind Speech Enhancement**

**Benny Sällberg**



Department of Signal Processing  
School of Engineering  
Blekinge Institute of Technology  
SWEDEN

© 2008 Benny Sällberg  
Department of Signal Processing  
School of Engineering  
Publisher: Blekinge Institute of Technology  
Printed by Printfabriken, Karlskrona, Sweden 2008  
ISBN 978-91-7295-154-9





## Abstract

Acoustic disturbances influence human speech communication by interfering with the communication process. In the worst case, it is impossible to communicate at all due to these disturbances. Methods that reduce the influence of the disturbances while preserving speech intelligibility are often desired. This thesis proposes real-world solutions for applied speech enhancement using autonomous and robust methods. Most of the work of the thesis concerns solutions to the problem of reducing acoustic disturbances within the framework of Blind Speech Enhancement (BSE). Notably, the term “blind” is assigned a positive attribute as it implies that the speech enhancement is carried out without any explicit references required. Instead, an assumption about the statistical independence between the sources coupled with an assumption regarding distinguishing statistical properties of the sources underpin the proposed methods. The unifying theory is Independent Component Analysis (ICA), which is performed by means of spatial filtering.

Two of the methods that are proposed in this thesis are shown, both in a theoretical and an empirical framework, to be robust in a real application while preserving stability even for Gaussian-only sources. Existing methods cannot guarantee stability in this scenario and Gaussian-only source mixtures may be the case in a real environment. The difference between the two methods lies in the different optimization strategies and the introduced approximations. The idea of injecting a single-channel method into the control loop of a blind beamformer is also proposed. In particular, two approaches are derived that aim at improving the blind beamformer in the case of disturbing noise and maintaining the same performance for different signal input levels. Finally, implementation aspects of a single-channel speech enhancer are discussed. The implementation aspects deal with the implementation of a speech enhancer in several different platforms such as analogue hardware, digital hardware, as well as hybrid analogue and digital hardware.



## Preface

This doctoral thesis summarizes my work at the Department of Signal Processing at Blekinge Institute of Technology beginning the 1<sup>st</sup> of February 2004. The thesis comprises six parts:

### Part

- I** Complex-valued Independent Component Analysis for Online Blind Speech Extraction
- II** Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction
- III** Online Maximization of Subband Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction
- IV** An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Robust Realtime Blind Speech Extraction
- V** Online Blind Speech Extraction Based on a Local Quadratic Kurtosis Criterion and a Preprocessing Automatic Gain Controller
- VI** Implementation Aspects of the Adaptive Gain Equalizer



## Acknowledgments

This work was made possible thanks to the great amount of support and encouragement I have received over the last five years from colleagues, friends and family. All present and former colleagues at the Department of Signal Processing are very special to me and they are all acknowledged here. I am particularly thankful to my supervisor Professor Ingvar Claesson. Thank you for making my PhD work possible, and thank you for your inspiration and guidance during these years. Also, I extend my gratitude to my dear friend and colleague Associate Professor Nedelko Grbić who deserves my everlasting gratefulness for all the hard work being my mentor. I am grateful because you have always given me your time when I needed it, and you have always supported, encouraged and aided me in my work. Nedo, I sincerely appreciate that you taught me how to water ski.

I also want to take this opportunity to thank all my friends and my family that have contributed to the life outside office hours. A special thank you goes to my dear friends Johan, Johnny and Martin. Thanks to my family Gunvor, Håkan, Eleonor, Tove, Noel, Lennart and Daniela for your support, love and tolerance towards me and my mission. It is probably not an easy task to enjoy life with a person whose mind is filled with signal processing. Endless thanks to you Dani for coping with my momentarily severe absent-mindedness and for all of your love, understanding and support.

Last but definitely not least I would like to acknowledge all those whose works I have read, scrutinized, welcomed and sometimes rejected. My work would not have been possible without all of your inspiring efforts - thank you!

*Benny Sällberg  
Karlskrona, November 13, 2008*



---

## Contents

<b>Publication List</b>	<b>1</b>
<b>Nomenclature</b>	<b>4</b>
<b>1 Motivation</b>	<b>5</b>
<b>2 System and Data Model</b>	<b>7</b>
2.1 Source and Sensor Model . . . . .	8
2.2 A Time-Frequency Domain Model . . . . .	12
2.3 Filter Bank Theory . . . . .	13
<b>3 Single-channel Speech Enhancement</b>	<b>15</b>
3.1 Spectral Subtraction . . . . .	16
3.2 Adaptive Gain Equalizer . . . . .	17
3.3 Motivation for Multi-channel Speech Enhancement . . .	19
<b>4 Multi-channel Speech Enhancement</b>	<b>21</b>
4.1 Optimal Beamforming . . . . .	24
4.1.1 Multi-Channel Wiener Filter . . . . .	24
4.1.2 Eigenvector Beamforming . . . . .	25
4.1.3 Linearly Constraint Minimum Variance Beam- former . . . . .	26
4.1.4 Minimum Variance Distortionless Response Beam- former . . . . .	27
4.1.5 Motivation for Adaptive Beamforming . . . . .	28
4.2 Adaptive Beamforming . . . . .	28
4.2.1 Adaptive Wiener Filter . . . . .	28
4.2.2 Generalized Sidelobe Canceler . . . . .	29
4.2.3 Motivation for Blind Adaptive Beamforming . .	30
4.3 Blind Adaptive Beamforming . . . . .	31
4.3.1 A Brief Introduction to ICA . . . . .	31
4.3.2 Kurtosis Measure of a Complex-valued Signal . .	34
4.3.3 An ICA Model using Beamforming Notation . .	41

---

4.3.4	Blind Beamforming by Kurtosis Maximization . . . . .	42
4.3.5	Post-processing versus In-the-Loop Processing . . . . .	45
<b>5</b>	<b>Realtime Speech Enhancement</b>	<b>47</b>
5.1	Realtime Speech Enhancement using DSP . . . . .	47
5.1.1	Fixed-Point Representation . . . . .	48
5.1.2	Floating-Point Representation . . . . .	48
5.2	Realtime Speech Enhancement in MATLAB . . . . .	49
5.3	Achieving High Realtime Performance . . . . .	50
5.4	Multi-rate Filter Bank . . . . .	51
5.4.1	Input-Output Signal Assembly . . . . .	52
5.4.2	Analysis Prototype Filter Polyphase Implementation . . . . .	53
5.4.3	Synthesis Prototype Filter Polyphase Implementation . . . . .	56
5.4.4	Efficient Filter Bank Implementation . . . . .	58
<b>6</b>	<b>Summary</b>	<b>59</b>
6.1	Main Contributions . . . . .	59
6.2	Future Research . . . . .	63
<b>Part I</b>		<b>77</b>
<b>Part II</b>		<b>111</b>
<b>Part III</b>		<b>129</b>
<b>Part IV</b>		<b>147</b>
<b>Part V</b>		<b>167</b>
<b>Part VI</b>		<b>183</b>

## Publication List

### Publications included in this thesis:

#### Part I is published as

B. Sällberg, N. Grbić, and I. Claesson, “Complex-valued Independent Component Analysis for Online Blind Speech Extraction”, *IEEE Transaction on Audio, Speech and Language Processing*, 16(8):1624–1632, November 2008

#### Part II is published as

B. Sällberg, N. Grbić, and I. Claesson, “Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction”, accepted for publication in *IEEE Signal Processing Letters*, November 2008.

#### Part III is published as

B. Sällberg, N. Grbić, and I. Claesson, “Online Maximization of Sub-band Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction”, *IEEE 15th International Conference on Digital Signal Processing*, pp. 603-605, July 2007.

#### Part IV is published as

B. Sällberg, N. Grbić, and I. Claesson, “An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Robust Realtime Blind Speech Extraction”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 309–312, March 2008.

#### Part V is published as

B. Sällberg, N. Grbić, and I. Claesson, “Online Blind Speech Extraction based on a Locally Quadratic Kurtosis Criteria and a Preprocessing

Automatic Gain Controller”, *49th International Symposium ELMAR-2007 focused on Multimedia Signal Processing and Communications*, pp. 139–142, Sept. 2007.

### **Part VI is published as**

B. Sällberg, N. Grbić, and I. Claesson, “Implementation Aspects of the Adaptive Gain Equalizer”, *Research Report Blekinge Tekniska Högskola 2006:4*, May 2006.

Parts of this report have been published as:

B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson, “A Mixed Analog-Digital Hybrid for Speech Enhancement Purposes”, *IEEE International Symposium on Circuits and Systems*, pp. 852–855, Vol. 2, May 2005.

B. Sällberg, and M. Dahl, “Speech Enhancement Implementations in the Digital, Analog, and Hybrid Domain”, *IEEE Swedish System on Chip Conference*, April 2005.

B. Sällberg, H. Åkesson, N. Westerlund, M. Dahl, and I. Claesson, “Analog Circuit Implementation for Speech Enhancement Purposes”, *IEEE 38th Asilomar Conference on Signals Systems, and Computers*, pp. 2285–2289, Vol. 2, Nov. 2004.

### **Other publications**

B. Sällberg, F. Sattar, and I. Claesson, “On a Method for Improving Impulsive Sounds Localization in Hearing Defenders”, *EURASIP Journal on Audio, Speech, and Music Processing special issue on Intelligent Audio, Speech, and Music Processing Applications*, Volume 2008, 2008.

M. Swartling, B. Sällberg, and N. Grbić, “Direction of arrival estimation for speech sources using fourth order cross cumulants”, *IEEE In-*

*ternational Symposium on Circuits and Systems*, pp. 1696–1699, May 2008.

Z. Yermeche, B. Sällberg, N. Grbić, and I. Claesson, “Real-Time DSP Implementation of a Subband Beamforming Algorithm for Dual Microphone Speech Enhancement”, *IEEE International Symposium on Circuits and Systems*, pp. 353–356, May 2007.

B. Sällberg, N. Grbić, and I. Claesson, “Blind Beamforming Using Parallel Single-channel Speech Enhancers”, *48th International Symposium ELMAR-2006 focused on Multimedia Signal Processing and Communications*, pp. 123–126, June 2006.

B. Sällberg, M. Swartling, N. Grbić, and I. Claesson, “Real time Implementation of a Blind Beamformer for Subband Speech Enhancement using Kurtosis Maximization”, *International Workshop on Acoustic Echo and Noise Control*, pp. 1–4, Sept. 2006.

B. Sällberg, “Applied Methods to Combat Noise in Human Communication”, *Licentiate Dissertation Blekinge Tekniska Högskola 2006:6*, ISBN 91-7295-087-0, June 2006.

B. Sällberg, L. Håkansson, and I. Claesson, “Active Noise Control for Hearing Protection using a Low Power Fixed point Digital Signal Processor”, *International Workshop on Acoustic Echo and Noise Control*, pp. 65–68, Sept. 2005.

## Nomenclature

$\mathbb{N}$	Natural numbers, including zero
$\mathbb{R}$	Real numbers
$\mathbb{C}$	Complex numbers
$j$	Imaginary unit, $j = \sqrt{-1}$
$M, m$	Number of microphones, $m : m \in \mathbb{N}, m < M$
$K, k$	Number of subbands, $k : k \in \mathbb{N}, k < K$
$(\cdot)_m$	Sub-script $m$ denotes microphone index
$(\cdot)^{[k]}$	Super-script $k$ denotes subband index
$I, i$	Number of independent sources, $i : i \in \mathbb{N}, i < I$
$t$	Sample index in time domain
$n$	Sample index in time-frequency domain
$F_S$	Sampling frequency in Hz
$x[t]$	Discrete time signal
$X[z]$	$\mathcal{Z}$ -transform of $x[t]$
$x^{[k]}[n]$	Time-frequency signal for subband $k$
$(\cdot)^*$	Complex conjugate
$\text{Re}\{\cdot\}$	Real part
$\text{Im}\{\cdot\}$	Imaginary part
$\text{E}\{x[t]\}$	Expectation of the signal $x[t]$ at time instant $t$
$\text{Var}\{x[t]\}$	Variance of the signal $x[t]$ at time instant $t$
$\mathbf{x}$	A vector, bold letter
$\mathbf{X}$	A matrix, capital bold letter
$\mathbf{I}_p$	Identity matrix of size $p \times p$
$\mathbf{0}_{p,q}$	Null-matrix of size $p \times q$
$\det \mathbf{X}$	Determinant of matrix $\mathbf{X}$
$\ \cdot\ _p$	$p$ -norm
$(\cdot)^T$	Matrix transpose
$(\cdot)^H$	Matrix complex conjugate transpose
$\nabla_{\mathbf{w}^*}$	Gradient operator with respect to $\mathbf{w}^*$
$\nabla_{\mathbf{w}^*}^2$	Laplace operator with respect to $\mathbf{w}^*$ , $\nabla_{\mathbf{w}^*}^2 \equiv \nabla_{\mathbf{w}^*} \nabla_{\mathbf{w}^*}^T$

## Introduction

### 1 Motivation

Speech communication is today an integral part of daily life for many people. While remote speech communication is carried out over some link, e.g., a telephone channel, microphone sensors are used to capture the speech signal to be transmitted. However, the microphones capture not only the intended speech signal but all acoustic sounds that are in the vicinity of the microphones. Sounds that are not wanted in the communication process are commonly referred to as disturbances, which can be both noise and interferences. Such unwanted sounds have a negative influence on communication because they may disturb the conversation and may even make it impossible to communicate at all. The level of the speech in relation to the level of the disturbances generally decreases when the distance between the speaker and the microphone increases, and this signifies the problem with disturbing sounds. An example where disturbances have a particularly bad influence on a conversation is in hands-free telephony where the user is typically at an arms-length distance from the microphone.

Signal processing methods have been shown to be an effective way of aiding the speech communication by reducing the level of the disturbances in relation to the level of the speech. Such signal processing methods are referred to as speech enhancement methods, or simply as speech enhancers (see, e.g., [1, 2, 3]). The speech enhancement has traditionally been carried out using one microphone, where temporal and statistical information about the speech signal and the disturbing signals has been exploited in order to perform the speech enhancement [4]. Several methods using multiple microphones exist today (see, e.g., [2, 5]), but many of them need to be re-designed or re-calibrated when the electroacoustical environment changes, e.g., when sources move. Modern advances in computer technology allow for the development of

elaborate speech enhancement methods using many microphones that are robust to real world variations (see, e.g., [6, 7, 8, 9, 10]).

This thesis deals with novel methods that aim to neutralize the negative influence of disturbing sounds by providing sophisticated speech enhancement and noise reduction. The developments in the thesis span theoretical analyzes of expected statistical performance of the methods, as well as practical experiments that validate the performance of the methods. The proposed methods are suitable for realtime application, and all methods have been implemented on digital signal processing systems and verified in real-time.

It is worth noting that even if the underlying theories for speech enhancement are based on theoretical and conceptual models, the actual speech enhancement effect is carried out by filtering the real observed data in any of the available physical domains (e.g., time and space). Hence, it is important to connect the physical domain processing with the models that describe the underlying processes. In relation to this, Section 2 provides a system model and a data model that are used to describe acoustic sources and the propagation of sound from an acoustic source to a receiving microphone. In addition, a time-frequency domain model is introduced in this section. An exemplary embodiment of how a time-frequency transformation can be carried out by using filter banks is presented in Section 5.4. A model and a set of common assumptions that underlie modern single-channel speech enhancement are provided in Section 3 together with a brief introduction of two exemplary single-channel speech enhancers. Techniques using multiple microphones, also known as array speech enhancement or beamforming, are outlined in Section 4. Theory and examples of optimal beamforming, adaptive beamforming, and blind beamforming are described in this section. A pronounced focus is on elaborate methods that rely only on statistical assumptions about the signals as this approach is closely related to the main work in the thesis. A short discussion about realtime implementation aspects are given under Section 5. A summary is given in Section 6 where the main contributions of each part are listed together with suggestions for future research.

## 2 System and Data Model

A mathematical framework is required in order to describe and to perform signal processing. The mathematical framework typically comprises a model together with a set of *a priori* assumptions that describe a set of selected features in the physical domain. This section presents a general model of how acoustic sound waves are propagated and received by a set of microphones. This model is used as a common ground for all parts of the thesis. The models that are used in the various parts of the thesis may, at first sight, appear to be different. However, it should be noted that the models are all sub-sets of the general model that is outlined next.

All signals are assumed to be electrical representations of physical quantities. This assumption implies that a sensor has been used to translate a physical quantity (e.g., air pressure level) into an electrical signal so that a certain amount of electrical voltage (or current) in the electrical sensor signal corresponds to a certain amount of a physical quantity. Unless otherwise stated, it is henceforth assumed that a continuous time electrical sensor signal has been sampled with the sampling frequency  $F_S$  [Hz] and that it is being correctly represented by a corresponding discrete-time signal.

Acoustic sound is described by the wave theory as traveling local variations in air pressure levels (see, e.g., [11]). A microphone essentially measures variations in the air pressure level around a nominal point and translates this into a representative electrical voltage. Microphones are therefore used to sense acoustic sound. The system model and the data model that have been used in the thesis focuses solely upon acoustic signals that have been captured by microphones. However, the proposed methods are not limited to the applications in the thesis, and the methods can be used for other applications as well.

## 2.1 Source and Sensor Model

It is assumed that a number of  $M$  microphones are being used. An electrical microphone signal is denoted as  $x_m[t]$ , where  $t : t \in \mathbb{N}$  denotes a sample index, and the index  $m : m \in \mathbb{N}, m < M$  is used to designate the  $m^{\text{th}}$  microphone signal. If  $M = 1$ , a short notation is used:  $x[t] \equiv x_0[t]$ . The microphone signals  $x_m[t]$  comprise a superposition of several signal components as described next. Some key definitions follow:

- A *source* refers to an entity that is capable of generating an acoustic sound that is unique in some sense, whether it be temporal, spatial, or statistical, from the sound of any other source.
- Only a source that is capable of producing acoustic sounds at levels that can be captured with a prescribed sensitivity of a sensor or a digital system is acknowledged here.
- A *clean source signal* refers to the recorded signal that has been produced by a source in a non-dispersive, lossless, and homogeneous medium where no other sounds are present.
- The received source signal of a *spatial source* carries spatial information that is unique from other spatial sources.
- Two main source classes are acknowledged in this work: *desired sources* and *disturbing sources*.
- The class of disturbing sources is further categorized into *noise sources* and *interfering sources*.

The desired sources together with the interfering sources are assumed to be spatial sources. It is furthermore assumed that there are a number of  $I$  spatial sources present and the corresponding clean source signals are denoted as  $s_i[t]$  for  $i : i \in \mathbb{N}, i < I$ . It is assumed that  $M$  noise signals are present, i.e., one noise signal  $v_m[t]$  per microphone element  $m : m \in \mathbb{N}, m < M$ . The noise signals may, for instance, comprise ambient noise or spatially incoherent noise [11].

The received signal for microphone  $m$ , due to the spatial propagation of a source signal related to the  $i^{\text{th}}$  spatial source, is denoted as  $x_{m,i}[t]$ . Due to the linearity of the wave equation [11], the propagation of a spatial source signal to a microphone is modeled according to a linear causal convolution:

$$x_{m,i}[t] = \sum_{\tau=0}^{T-1} a_{m,i}[\tau]s_i[t - \tau]. \quad (1)$$

In this model,  $a_{m,i}[\tau]$  is an impulse response function that describes the acoustical propagation path between the spatial source number  $i$  and microphone  $m$ . The length of a propagation path is here assumed to be restricted to  $T$  samples, or  $\frac{T}{F_S}$  seconds. A propagation model is illustrated in Figure 1 for a number of three propagation path components, one line-of-sight component, and two reflection components. It is acknowledged that the propagation model in (1) is restricted to linear propagation channels since it only captures linear dynamics of the wave equation. Nonlinear acoustic or electric effects that can arise in the wave propagation or the signal acquisition, e.g., due to a dispersive propagation [11] or sensor signal saturation, are not modeled here.

The received microphone signal for a set of  $I$  spatial sources, with an additive noise signal  $v_m[t]$ , is constructed by a linear superposition of all signals:

$$x_m[t] = \sum_{i=0}^{I-1} x_{m,i}[t] + v_m[t] = \sum_{i=0}^{I-1} \sum_{\tau=0}^{T-1} a_{m,i}[\tau]s_i[t - \tau] + v_m[t]. \quad (2)$$

The total propagation model according to (2) is illustrated in Figure 2.

The Signal to Noise Ratio (SNR) measures the power of the desired signals in relation to the power of the noise signals, and the Signal to Interference Ratio (SIR) measures the power of the desired signals in relation to the power of the interfering signals. The Signal to Noise and Interference Ratio (SNIR) measures the power of the desired signals in

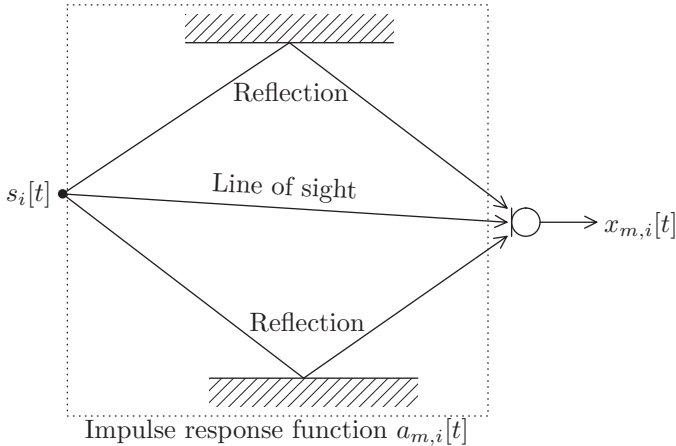


Figure 1: A propagation model from a spatial source  $i$ , with source signal  $s_i[t]$ , to a microphone  $m$  with signal  $x_{m,i}[t]$  according to a linear filter  $a_{m,i}[t]$  for three propagation path components, one line-of-sight component, and two reflection components.

relation to the combined power of the noise signals and the power of the interfering signals.

**Motivation for time-frequency processing** Algorithms that operate in the time domain may suffer from a heavy computational load. This problem is significant even for a moderately advanced task such as computing a matrix multiplication between a square matrix and a vector, which is the case in, e.g., the Recursive Least Squares (RLS) algorithm. The number of operations required for this task is proportionally quadratic to the number of filter coefficients. In addition, the rate of convergence for adaptive filters is generally reduced for long filters since the step-size is often inversely proportional to the number of filter taps [12]. A popular approach in modern signal processing taken in order to circumvent the drawbacks associated with time domain processing, is to introduce a time-frequency representation of the

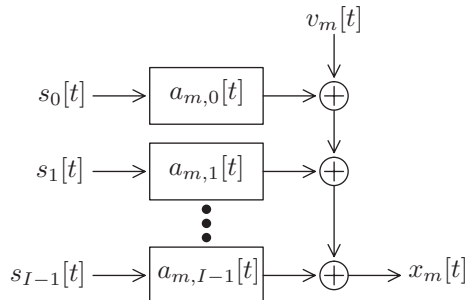


Figure 2: The propagation model for a total of  $I$  spatial source signals to microphone  $m$  including an additive noise signal  $v_m[t]$ .

time signals [13, 14]. A complex time domain problem can then be divided into a set of several low-complexity problems in the frequency domain.

A frequency domain representation of the model in (2), using  $K$  frequency bands, has propagation paths that are limited to  $T_K \propto \frac{T}{K}$  samples<sup>1</sup>, where  $T$  is the number of samples in the time domain propagation path (1). The relationship between the length of a frequency domain propagation path and the length of a time domain propagation path is directly applicable to filter theory as a  $L$ -tap long time domain filter corresponds to a number of  $K$  frequency domain filters, each of length  $L_K \propto \frac{L}{K}$  taps. An example that highlights the computational benefits of frequency domain processing follows.

**Example 1:** Let us assume that a time domain operation on a filter of length  $L$  taps requires an order of  $L^2$  computations. A corresponding operation that is computed in parallel for  $K$  frequency bands requires essentially an order of  $K \cdot L_K^2 = \frac{L^2}{K}$  operations instead of  $L^2$  operations.

<sup>1</sup>This discussion relates to a time-frequency signal representation that adopts a critical down-sampling of the time-frequency data, i.e., where a decimation factor  $D$  is equal to the number of time-frequency components  $K$ . An appropriate relationship is  $T_K \propto \frac{T}{D}$  if the decimation ratio  $D < K$ .

It should be noted that a frequency domain transformation requires a number of operations to be computed. If, for instance, a Fast Fourier Transform (FFT) [15] is used to compute a frequency domain representation, then an order of  $K \log_2 K$  operations are required for the FFT computation. The reduction in the number of required operations between a time domain approach and a frequency domain approach is essentially a factor  $K$  for the rudimentary example above. There are many cases where frequency domain processing renders a vast reduction in the number of operations (see, e.g., [8, 16, 17]). The very essence of why frequency domain processing is so much more efficient than corresponding time domain processing lies in the amount of parallelism that is provided by a frequency domain representation.

## 2.2 A Time-Frequency Domain Model

The actual transformation of a time signal into the time-frequency domain can be carried out in numerous ways [13, 14]. The time-frequency transformation is performed throughout the thesis by using a *filter bank*. A filter bank - as the name suggests - comprises a bank of  $K$  bandpass filters that selects various frequency bands of a signal. Each frequency band is shortly denoted as a *subband*. A subband signal for a microphone  $m$  is denoted as  $x_m^{[k]}[n]$ , where  $k : k \in \mathbb{N}, k < K$  represents the subband index and  $n : n \in \mathbb{N}$  is a subband sample index which may or may not be identical to  $t$ . The linear causal convolution model that was developed in the time domain (2) is represented in the time-frequency domain<sup>2</sup> as

$$x_m^{[k]}[n] \approx \sum_{i=0}^{I-1} \sum_{\tau=0}^{T_K-1} a_{m,i}^{[k]}[\tau] s_i^{[k]}[n - \tau] + v_m^{[k]}[n]. \quad (3)$$

---

<sup>2</sup>It is acknowledged that the time-frequency representation (3) of the model in (2) is an approximation as there may be signals leaking between subbands. However, it is assumed that the time-frequency transformation is suitably configured so as to minimize the signal leakage between subbands.

---

The signals  $a_{m,i}^{[k]}[n]$ ,  $s_i^{[k]}[n]$ , and  $v_m^{[k]}[n]$  are the time-frequency representations of  $a_{m,i}[t]$ ,  $s_i[t]$ , and  $v_m[t]$ , respectively. Due to the computational benefits and the versatility, it is the time-frequency model in (3) that has been used in the various parts of this work.

### 2.3 Filter Bank Theory

It is customary to divide a filter bank into an analysis part and a synthesis part. The analysis part computes the time-frequency subband signals from a time signal, and the synthesis part reconstructs an output time signal from the subband signals. If the filter bank filters are properly selected, the analysis part is followed by a decimator, and the synthesis part is preceded by an interpolator, then the efficiency of the filter bank is increased. The number of subbands  $K$  over the decimation ratio<sup>3</sup>  $D$ , with  $D \leq K$ , yields an *over-sampling ratio*  $O = K/D$ . Due to the particular filter bank structure, the over-sampling ratio is an integer value in this work, see Section 5. If  $O = 1$ , the filter bank is critically sampled, and if  $O > 1$ , the filter bank is non-critically sampled or over-sampled. A filter bank with a perfect reconstruction property renders a zero reconstruction error of the time signal after the synthesis part (provided that the subband signals are not altered) [18]. However, some applications, such as subband adaptive filtering, alter the subband signals, and these alterations may cause degradations due to aliasing distortion in the reconstructed time signal. Increasing the over-sampling ratio, which is equivalent to decreasing the decimation ratio, renders a non-critically sampled filter bank, and the frequency point around which aliasing-distortion occurs is thereby moved away from the subband frequency range. An increased over-sampling ratio is therefore a way to circumvent degradations due to aliasing distortion, and it makes the solution more robust in that aspect. In addition, the use of the Noble identities allows for the bandpass filtering and the decimation/interpolation to switch places in a polyphase structure,

---

<sup>3</sup>The filter bank employed in this work uses an interpolation ratio that is identical to the decimation ratio.

which render a further advancement in the efficiency of a filter bank [13, 14, 19].

In the discussion of filter banks, it is important to consider if the subbands are uniformly spaced or non-uniformly spaced. The word *spacing* refers to the distance between the center frequencies of two consecutive subbands in a filter bank. In some applications, for instance in perceptual audio coding, the filter bank is matched to a cochlear model having a non-uniform subband spacing [20]. However, it is sufficient in many applications to employ a uniform spacing of the filter bank, provided that the number of subbands is sufficiently large. This thesis focuses on uniformly spaced filter banks, and a particular filter bank structure is adopted here, namely the uniform Discrete Fourier Transform (DFT) modulated filter bank which is derived in Section 5.

### 3 Single-channel Speech Enhancement

Single-channel speech enhancement refers to methods that perform signal processing using only the data available in a single channel provided by one microphone, i.e.,  $M = 1$  (see, e.g., [1, 4]). The single-channel speech enhancement model that has been used in this work assumes one speech source signal (i.e.,  $I = 1$ ) that is disturbed by a noise signal:

$$x^{[k]}[n] = \sum_{\tau=0}^{T_K-1} a^{[k]}[\tau]s^{[k]}[n-\tau] + v^{[k]}[n]. \quad (4)$$

While the focus is on noise reduction, no attention is given to the propagation part of the model and it is therefore assumed that  $a^{[k]}[\tau] = \delta[\tau]$ , which renders a simplified single-channel noise reduction model:

$$x^{[k]}[n] = s^{[k]}[n] + v^{[k]}[n]. \quad (5)$$

It is assumed that the speech signal and the noise signal have zero mean and that the signals are uncorrelated, i.e.,

$$\mathbb{E} \left\{ s^{[k]}[n]v^{[k]}[n]^* \right\} = 0. \quad (6)$$

In addition, it is assumed that the speech signal is stationary over a shorter time than the disturbing noise signal. A real-valued gain function  $g^{[k]}[n]$  is used to impose a noise reduction effect as

$$\begin{aligned} y^{[k]}[n] &= g^{[k]}[n]x^{[k]}[n] = \\ &= g^{[k]}[n]s^{[k]}[n] + g^{[k]}[n]v^{[k]}[n]. \end{aligned} \quad (7)$$

The noise reduced output signal is denoted as  $y^{[k]}[n]$ . The signal model for single-channel noise reduction in (7) is widespread and used in most modern noise reduction methods (see, e.g., [1, 4, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32]).

At first glance, it would seem that the methods based on the model in (7) cannot contribute to speech enhancement in a better way than

the Wiener filter. However, by utilizing the fact that the speech signal is highly non-stationary with regard to the noise signal, these methods do reduce the overall amount of noise even further. In addition, when speech is active in one subband the human hearing system effectively masks the noise component in that subband as well as in surrounding subbands, and it makes the noise reduction effect perceptible [33, 34, 35].

### 3.1 Spectral Subtraction

One popular class of single-channel speech enhancement methods is based on the spectral subtraction technique which dates back to the late 1970s [4, 21, 22]. The spectral subtraction methods rely upon the assumption that the noise signal and the speech signal are uncorrelated (6). An estimate of the background noise spectrum  $\hat{P}_{v,\alpha}^{[k]}[n] \approx \text{E} \left\{ |v^{[k]}[n]|^\alpha \right\}$  is acquired during periods when speech is absent. In addition, an estimated spectrum of the observed signal mixture  $\hat{P}_{x,\alpha}^{[k]}[n] \approx \text{E} \left\{ |s^{[k]}[n]|^\alpha \right\} + \text{E} \left\{ |v^{[k]}[n]|^\alpha \right\}$  is constantly acquired. The subtraction of the background noise spectrum from the observed spectrum yields an estimate of the speech spectrum  $\hat{P}_{s,\alpha}^{[k]}[n] = \hat{P}_{x,\alpha}^{[k]}[n] - \hat{P}_{v,\alpha}^{[k]}[n] \approx \text{E} \left\{ |s^{[k]}[n]|^\alpha \right\}$ . The parameter  $\alpha$  controls the type of spectrum to be estimated, e.g., a magnitude spectrum ( $\alpha = 1$ ) or a power spectrum ( $\alpha = 2$ ). A real-valued spectral subtraction gain function  $g_{\text{SPEC-SUB}}^{[k]}[n]$  is computed as [24]

$$g_{\text{SPEC-SUB}}^{[k]}[n] = \left( \frac{\hat{P}_{x,\alpha}^{[k]}[n] - \beta \cdot \hat{P}_{v,\alpha}^{[k]}[n]}{\hat{P}_{x,\alpha}^{[k]}[n]} \right)^{\frac{1}{\alpha}}. \quad (8)$$

The parameter  $\beta$  is a subtraction factor, and it can be a function of the estimated SNR in order to improve performance [4]. In the case that  $\alpha = 2$  and  $\beta = 1$ , this gain function corresponds to an estimate of a time-varying noncausal Wiener filter. The spectral subtraction

method relies upon a supplementary structure, a Voice Activity Detector (VAD), that estimates when the received noisy signal contains speech. The performance of a spectral subtraction method goes therefore hand-in-hand with the performance of the VAD that fails in operation at SNR lower than 6 dB (see, e.g., [36]).

It is noted that this description of spectral subtraction only covers the underlying theory, and a straightforward implementation is typically flawed with undesired artifacts in the enhanced speech. Modern spectral subtraction focuses on reducing these artifacts by, e.g., incorporating smooth updates of the spectral estimates [23, 24]. Another approach uses a secondary microphone to acquire better estimates of the signal spectrums (see, e.g., [37]).

### 3.2 Adaptive Gain Equalizer

A fundamentally different method is the recently proposed Adaptive Gain Equalizer (AGE) [25, 26, 27] as it focuses on the enhancement of speech rather than focusing on reducing the noise, which is the case in, e.g., spectral subtraction. This alternative shift of focus renders a method that is free from a supplementary VAD structure, and this is advantageous in some cases. The AGE assumes the same signal model as the spectral subtraction (5) where the stationarity time of the speech signal is assumed to be significantly lower than the stationarity time of the noise signal. The AGE uses two averages with different tracking times. One slow average, denoted as  $a_{\text{slow},\alpha}^{[k]}[n]$ , tracks the noise signal level and one fast average, denoted as  $a_{\text{fast},\alpha}^{[k]}[n]$ , tracks the noise-plus-speech signal level according to:

$$a_{\text{slow},\alpha}^{[k]}[n] \approx \text{E} \left\{ \left| v^{[k]}[n] \right|^\alpha \right\}, \quad (9)$$

$$a_{\text{fast},\alpha}^{[k]}[n] \approx \text{E} \left\{ \left| s^{[k]}[n] \right|^\alpha \right\} + \text{E} \left\{ \left| v^{[k]}[n] \right|^\alpha \right\}. \quad (10)$$

The real-valued AGE gain function  $g_{\text{AGE}}^{[k]}[n]$  is then computed as a function  $f^{[k]}$  which has as input parameter the quotient of these two averages according to:

$$\begin{aligned} g_{\text{AGE}}^{[k]}[n] &= f^{[k]} \left( \left( \frac{a_{\text{fast},\alpha}^{[k]}[n]}{a_{\text{slow},\alpha}^{[k]}[n]} \right)^{\frac{1}{\alpha}} \right) \approx \\ &\approx f^{[k]} \left( \left( \frac{\mathbb{E} \left\{ |s^{[k]}[n]|^\alpha \right\}}{\mathbb{E} \left\{ |v^{[k]}[n]|^\alpha \right\}} + 1 \right)^{\frac{1}{\alpha}} \right). \end{aligned} \quad (11)$$

The function  $f^{[k]}$  is monotonically increasing and limited<sup>4</sup> as  $f^{[k]} : \mathbb{R} \rightarrow \mathbb{R} \in [1, G^{[k]}]$ . The term  $G^{[k]} > 1$  is a design parameter that controls the maximal allowed speech amplification in the method. The input parameter of this function is the quotient of the two averages, and the output value, i.e.,  $g_{\text{AGE}}^{[k]}[n]$ , is limited to an interval  $1 \leq g_{\text{AGE}}^{[k]}[n] \leq G^{[k]}$  either in a hard manner, where  $f^{[k]}$  is a hard clipper, or in a soft manner, where  $f^{[k]}$  is a smooth limiter [28, 29, 30, 31, 32]. During periods when speech is active,  $a_{\text{fast},\alpha}^{[k]}[n] \gg a_{\text{slow},\alpha}^{[k]}[n]$ , it means that  $g_{\text{AGE}}^{[k]}[n] = G^{[k]}$ . When speech is inactive,  $a_{\text{fast},\alpha}^{[k]}[n] \approx a_{\text{slow},\alpha}^{[k]}[n]$ , it means that  $g_{\text{AGE}}^{[k]}[n] = 1$ . A speech signal normally has a non-abrupt transition from the inactive mode to the active mode. This yields also a smooth transition of the gain function from the pass-through mode  $g_{\text{AGE}}^{[k]}[n] = 1$  to the maximum amplify mode  $g_{\text{AGE}}^{[k]}[n] = G^{[k]}$ . Clearly the gain function is significantly larger than unity only if speech is present. In all other cases, the gain function is close to unity and thereby leaves the noise essentially intact during speech inactivity. This behavior renders a particularly comfortable background noise quality

<sup>4</sup>Imposing a scaling  $\frac{1}{G^{[k]}}$  on the gain function in (11) yields a noise reduction method instead of a speech amplification method. This scaling may be useful in certain implementations that have a limited precision in the numerical representation.

of the enhanced sound. An example signal is presented in Figure 3 where a fast average and a slow average are illustrated together with an AGE gain function. The AGE has been verified in hardware (see, e.g., [28, 29, 30, 31, 32]), and a discussion on various implementation aspects of the AGE in different platforms such as in an analogue, a digital, and a hybrid electrical implementation are provided in Part VI. The AGE method is also used in Part IV where it complements a spatial blind beamformer method for speech enhancement.

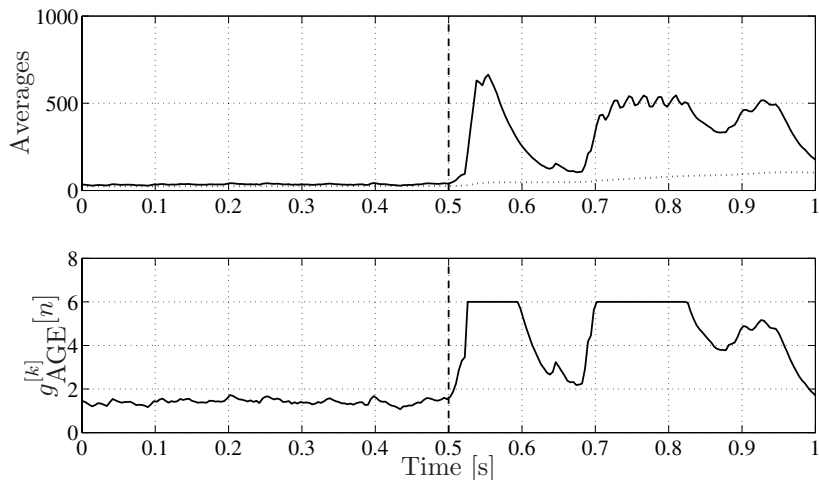


Figure 3: Upper: Estimates of the fast average (solid) and the slow average (dotted). Lower: The AGE gain function  $g_{\text{AGE}}^{[k]}[n]$ , which has been limited by a hard clipper to  $G^{[k]} = 6$ . Only noise is present during the first 0.5 s whereafter (marked by a vertical dashed line) both noise and speech are present.

### 3.3 Motivation for Multi-channel Speech Enhancement

Single-channel speech enhancement technologies are restricted to the available and estimated temporal information (time and frequency)

that is provided by the input signal. If the temporal information is insufficient, e.g., if the SNR is too low, the single-channel speech enhancers are destined to fail. A method that is able to also utilize the spatial domain has therefore the possibility of further advancing the speech enhancement performance. In order to operate in the spatial domain, it is required that several microphones be used, i.e., a microphone array. Speech enhancers that are based on microphone arrays provide a high degree of noise and interference reduction due to the spatial selectivity. In addition, it is possible, using a microphone array, to construct a distortionless speech enhancer, i.e., a speech enhancer that reduces disturbing signals without compromising the target speech signals. This is not possible in single-channel methods that distort the speech signals during the noise reduction process. However, it is noted that there is often a compromise between the level of speech distortion and the level of noise and interference reduction. If a small degree of speech distortion is allowed, the amount of noise and interference reduction is increased. Techniques for multi-channel speech enhancement are described next.

## 4 Multi-channel Speech Enhancement

The microphone array approach for speech enhancement harmonizes with the real world since physical sources are generally spatially disjoint (see, e.g., [11]). Due to the spatiotemporal filtering ability of multi-microphones technologies, such methods are favorable when it comes to applications with extremely low SNR or SIR for which the performance of single-channel speech enhancers generally fail. An example of such an application is hands-free telephony in high noise environments.

Techniques using multiple microphones refer to spatiotemporal filtering, or beamforming. The output signal of a beamforming in the time domain that is using Finite Impulse Response (FIR) filters  $w_m[t]$  of length  $L$ -taps is given by

$$y[t] = \sum_{m=0}^{M-1} \sum_{l=0}^{L-1} w_m[l] x_m[t-l], \quad (12)$$

where  $x_m[t]$  is defined in (2). If  $w_m[t] = \alpha_m \delta(t - \tau_m)$  for an attenuation factor  $\alpha_m$  and delay  $\tau_m$ , then this model corresponds to a *delay-and-sum* beamformer [11]. However, if there are several non-zero elements in  $w_m[t]$ , then (12) corresponds to a *filter-and-sum* beamformer. The time domain beamformer in (12) is illustrated in Figure 4. A cor-

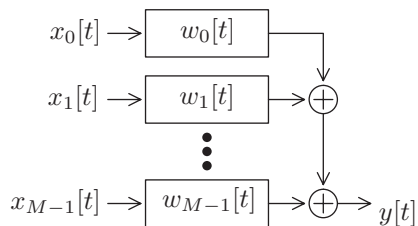


Figure 4: Time domain beamformer with beamformer filters  $w_m[t]$ .

responding input-output signal relation in the time-frequency domain

using FIR beamformer filters  $w_m^{[k]}[n]$  of length  $L_K \propto \frac{L}{K}$  taps is

$$y^{[k]}[n] = \sum_{m=0}^{M-1} \sum_{l=0}^{L_K-1} w_m^{[k]}[l]^* x_m^{[k]}[n-l], \quad (13)$$

where  $x_m^{[k]}[n]$  is defined in (3) and  $w_m^{[k]}[n]$  is the frequency representation of  $w_m[t]$ . A frequency domain beamformer is illustrated in Figure 5. A

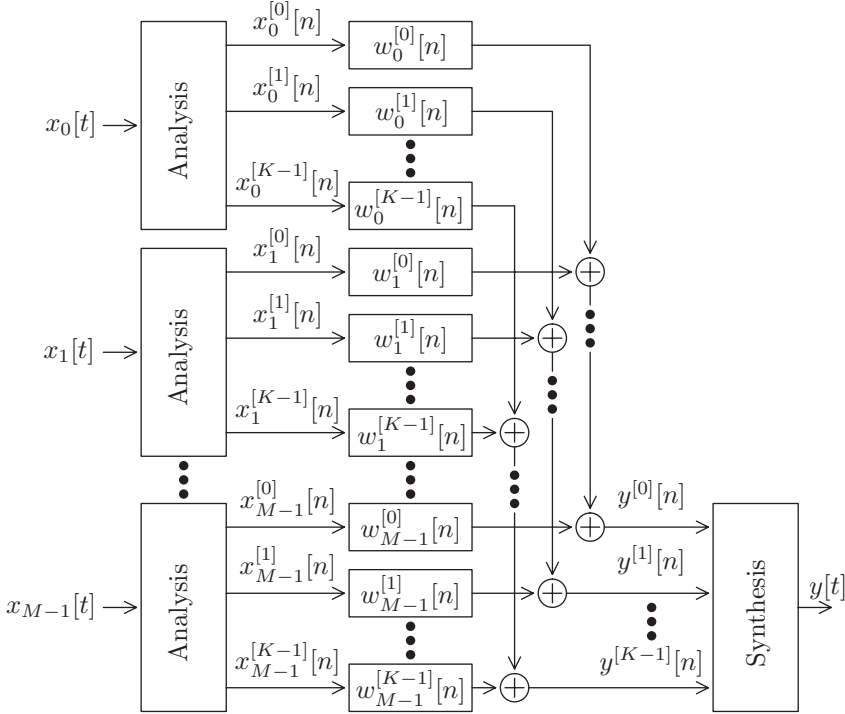


Figure 5: Frequency domain beamformer with beamformer filters  $w_m^{[k]}[n]$ .

compact notation is achieved by using a vector representation:

$$y^{[k]}[n] = \sum_{m=0}^{M-1} \mathbf{w}_m^{[k]H} \mathbf{x}_m^{[k]}[n] = \mathbf{w}^{[k]H} \mathbf{x}^{[k]}[n], \quad (14)$$

where

$$\mathbf{w}_m^{[k]} = \left( w_m^{[k]}[0], w_m^{[k]}[1], \dots, w_m^{[k]}[L_K - 1] \right)^T, \quad (15)$$

$$\mathbf{w}^{[k]} = \left( \mathbf{w}_0^{[k]T}, \mathbf{w}_1^{[k]T}, \dots, \mathbf{w}_{M-1}^{[k]T} \right)^T, \quad (16)$$

$$\mathbf{x}_m^{[k]}[n] = \left( x_m^{[k]}[n], x_m^{[k]}[n-1], \dots, x_m^{[k]}[n-L_K+1] \right)^T, \quad (17)$$

$$\mathbf{x}^{[k]}[n] = \left( \mathbf{x}_0^{[k]}[n]^T, \mathbf{x}_1^{[k]}[n]^T, \dots, \mathbf{x}_{M-1}^{[k]}[n]^T \right)^T. \quad (18)$$

A frequency domain beamformer that uses vector notation is illustrated in Figure 6.

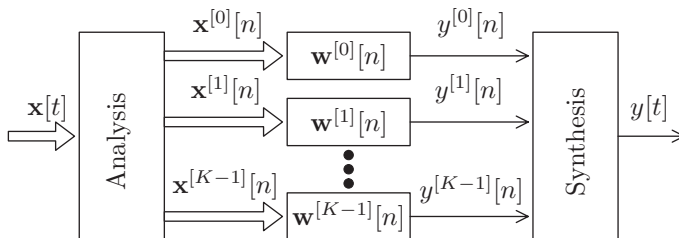


Figure 6: Frequency domain beamformer using vector notation and beamformer filters  $\mathbf{w}^{[k]}[n]$ .

It should be noted that a frequency domain beamformer always (provided that  $K > 1$ ) corresponds to a time domain filter-and-sum beamformer. A time domain filter-and-sum structure of length  $L$  taps is implied due to the relationship  $L_K \propto \frac{L}{K}$ , which means that  $L > 1$ . This work uses predominantly the beamformer given by (14), and it is often the case that  $L_K = 1$ .

The filter coefficients  $\mathbf{w}^{[k]}$  that are used in the digital filter network that comprises the beamformer (14) can be designed in many ways. The following sections explain some of the many methods for designing a beamformer, including a description of the beamforming approaches developed in this work.

## 4.1 Optimal Beamforming

An optimal beamformer pre-computes an optimal set of filter weights based on a model of the array and sources, or alternatively it can be based on calibration information. Under this class of beamformers, we find the multi-channel Wiener filter, the eigenvector beamformer, the Linearly Constrained Minimum Variance (LCMV) beamformer, and the Minimum Variance Distortion-less Response (MVDR) beamformer (see, e.g., [11, 38, 39, 40, 41]).

### 4.1.1 Multi-Channel Wiener Filter

A reference signal  $d^{[k]}[n]$  is used in order to formulate a Wiener filter [12]. An error signal is then defined as the difference between the reference signal and the beamformer output signal:  $d^{[k]}[n] - \mathbf{w}^{[k]H} \mathbf{x}^{[k]}[n]$ . The principle of orthogonality states that the error signal should be orthogonal to the input signal at an optimal point [15], i.e.,

$$\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \left( d^{[k]}[n]^* - \mathbf{x}^{[k]}[n]^H \mathbf{w}^{[k]} \right) \right\} = \mathbf{0}_{M \cdot L_K \times 1}, \quad (19)$$

where  $\mathbf{0}_{M \cdot L_K \times 1}$  is a null-vector of size  $M \cdot L_K \times 1$ . Well-established definition for an auto-correlation matrix and a cross-correlation vector are  $\mathbf{R}_{\mathbf{x}}^{[k]} = \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}$  and  $\mathbf{r}_{d\mathbf{x}}^{[k]} = \mathbb{E} \left\{ d^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\}$ , respectively. In a practical implementation, these correlations require being estimated based on the observed data or computed according to a model of the array and sources [42]. The Wiener-Hopf equations (also denoted as the normal equations)  $\mathbf{R}_{\mathbf{x}}^{[k]} \mathbf{w}^{[k]} = \mathbf{r}_{d\mathbf{x}}^{[k]}$  have an optimal solution, which is the Wiener filter,  $\mathbf{w}_{\text{WNR}}^{[k]}$ :

$$\mathbf{w}_{\text{WNR}}^{[k]} = \mathbf{R}_{\mathbf{x}}^{[k]-1} \mathbf{r}_{d\mathbf{x}}^{[k]}. \quad (20)$$

It was suggested in the standard Wiener formulation (see, e.g., [12]) that the reference signal  $d^{[k]}[n]$  should be one of the clean source signals  $s_i^{[k]}[n]$ . However, in that case, the Wiener filter has to tackle

not only the reduction of disturbances, but it also has to perform an inverse filtering of the propagation paths from the  $i^{\text{th}}$  source to the microphones. A simplified noise reduction problem is found by using one of the received speech signal components  $\sum_{\tau=0}^{T_K-1} a_{m,i}^{[k]}[\tau]s_i^{[k]}[n-\tau]$  as a reference (see, e.g., [5, 39]). By using a received speech component as a reference, it implies that the Wiener filter can focus solely on disturbance reduction as it does not have to perform an inverse filtering of the propagation path.

#### 4.1.2 Eigenvector Beamforming

Eigenvector beamforming refers to a class of methods that uses eigenanalysis to compute a beamformer filter (see, e.g., [43]). The maximum SNIR beamformer is one type of an eigenvector approach. The SNIR beamformer relies upon the partitioning of the received auto-correlation matrix that, due to the uncorrelated source signals, is:

$$\mathbf{R}^{[k]} = \text{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} = \sum_{i=0}^{I-1} \mathbf{R}_{\mathbf{x}_i}^{[k]} + \mathbf{R}_{\mathbf{v}}^{[k]}, \quad (21)$$

$$\mathbf{R}_{\mathbf{x}_i}^{[k]} = \text{E} \left\{ \mathbf{x}_i^{[k]}[n] \mathbf{x}_i^{[k]}[n]^H \right\}, \quad (22)$$

$$\mathbf{R}_{\mathbf{v}}^{[k]} = \text{E} \left\{ \mathbf{v}^{[k]}[n] \mathbf{v}^{[k]}[n]^H \right\}. \quad (23)$$

The matrix  $\mathbf{R}_{\mathbf{x}_i}^{[k]}$  corresponds to the auto-correlation matrix related to each received source signal  $\mathbf{x}_i^{[k]}[n]$ . A received source signal is defined by arranging elements in (3) related to source  $i$  as

$$\mathbf{x}_i^{[k]}[n] = \sum_{\tau=0}^{T_K-1} \mathbf{a}_i^{[k]}[\tau] s_i^{[k]}[n-\tau], \quad (24)$$

$$\mathbf{a}_i^{[k]}[\tau] = \left( a_{0,i}^{[k]}[\tau], a_{1,i}^{[k]}[\tau], \dots, a_{M-1,i}^{[k]}[\tau] \right)^T. \quad (25)$$

Hence,  $\mathbf{a}_i^{[k]}[\tau]$  contains the propagation paths from source  $i$  to all microphones. For illustration purposes, it is assumed that the first source  $i =$

0 is the desired source and that all other sources (for  $i \in \{1, 2, \dots, I-1\}$ ) are interferences. The SNIR beamformer,  $\mathbf{w}_{\text{SNIR}}^{[k]}$ , maximizes the beamformer output power for the desired source while minimizing the beamformer output power for all disturbances. The SNIR beamformer uses the following design criterion:

$$\mathbf{w}_{\text{SNIR}}^{[k]} = \arg \max_{\mathbf{w}^{[k]}} \left( \frac{\mathbf{w}^{[k]H} \mathbf{R}_{\mathbf{x}_0}^{[k]} \mathbf{w}^{[k]}}{\mathbf{w}^{[k]H} \left( \mathbf{R}_{\mathbf{v}}^{[k]} + \sum_{i=1}^{I-1} \mathbf{R}_{\mathbf{x}_i}^{[k]} \right) \mathbf{w}^{[k]}} \right). \quad (26)$$

The SNIR beamformer vector is a solution to a generalized eigenvalue problem:

$$\left( \mathbf{R}_{\mathbf{v}}^{[k]} + \sum_{i=1}^{I-1} \mathbf{R}_{\mathbf{x}_i}^{[k]} \right) \mathbf{w}^{[k]} \lambda = \mathbf{R}_{\mathbf{x}_0}^{[k]} \mathbf{w}^{[k]}, \quad (27)$$

where an eigenvalue is denoted by  $\lambda$  and a corresponding eigenvector is  $\mathbf{w}^{[k]}$ . The generalized eigenvalue problem can also be arranged as an ordinary eigenvalue problem (provided that the inverse exists):

$$\mathbf{w}^{[k]} \lambda = \left( \mathbf{R}_{\mathbf{v}}^{[k]} + \sum_{i=1}^{I-1} \mathbf{R}_{\mathbf{x}_i}^{[k]} \right)^{-1} \mathbf{R}_{\mathbf{x}_0}^{[k]} \mathbf{w}^{[k]}. \quad (28)$$

The eigenvector that corresponds to the largest eigenvalue associated to the eigenvalue problems above solves the SNIR beamformer criterion (26), and it corresponds therefore to the SNIR-optimal beamformer filter vector. Note that the SNIR beamformer criterion is scale invariant, i.e., any scale applied to the weight vector does not alter the SNIR-optimum.

### 4.1.3 Linearly Constraint Minimum Variance Beamformer

The LCMV beamformer minimizes the variance, i.e., power, of the beamformer output signal  $E \left\{ |y^{[k]}[n]|^2 \right\} = \mathbf{w}^{[k]H} \mathbf{R}_x^{[k]} \mathbf{w}^{[k]}$  subject to a

set of predetermined linear constraints  $\mathbf{w}^{[k]H} \mathbf{C}^{[k]} = \mathbf{c}^{[k]H}$ , or equivalently,  $\mathbf{C}^{[k]H} \mathbf{w}^{[k]} = \mathbf{c}^{[k]}$ . The LCMV optimization criterion is

$$\min_{\mathbf{w}^{[k]}} \mathbf{w}^{[k]H} \mathbf{R}_x^{[k]} \mathbf{w}^{[k]}, \text{ subject to } \mathbf{w}^{[k]H} \mathbf{C}^{[k]} = \mathbf{c}^{[k]H}. \quad (29)$$

Lagrangian multipliers  $\mathbf{\Lambda}^{[k]}$  are used to construct a cost function

$$J(\mathbf{w}^{[k]}, \mathbf{\Lambda}^{[k]}) = \mathbf{w}^{[k]H} \mathbf{R}_x^{[k]} \mathbf{w}^{[k]} - \left( \mathbf{w}^{[k]H} \mathbf{C}^{[k]} - \mathbf{c}^{[k]H} \right) \mathbf{\Lambda}^{[k]}. \quad (30)$$

The optimal solution of the Lagrange cost function is found at the point where all partial derivatives are equal to zero:

$$\frac{\partial J(\mathbf{w}^{[k]}, \mathbf{\Lambda}^{[k]})}{\partial \mathbf{w}^{[k]*}} = \mathbf{R}_x^{[k]} \mathbf{w}^{[k]} - \mathbf{C}^{[k]} \mathbf{\Lambda}^{[k]} = \mathbf{0}, \quad (31)$$

$$\begin{aligned} \frac{\partial J(\mathbf{w}^{[k]}, \mathbf{\Lambda}^{[k]})}{\partial \mathbf{\Lambda}^{[k]}} &= \mathbf{w}^{[k]H} \mathbf{C}^{[k]} - \mathbf{c}^{[k]H} = \\ &= \mathbf{C}^{[k]H} \mathbf{w}^{[k]} - \mathbf{c}^{[k]} = \mathbf{0}. \end{aligned} \quad (32)$$

Solving for  $\mathbf{w}^{[k]}$  in (31) yields  $\mathbf{w}^{[k]} = \mathbf{R}_x^{[k]-1} \mathbf{C}^{[k]} \mathbf{\Lambda}^{[k]}$ , and inserting this expression into (32) and solving for  $\mathbf{\Lambda}^{[k]}$  yields the optimal Lagrange multipliers:  $\mathbf{\Lambda}_{\text{opt}}^{[k]} = \left( \mathbf{C}^{[k]H} \mathbf{R}_x^{[k]-1} \mathbf{C}^{[k]} \right)^{-1} \mathbf{c}^{[k]}$ . Inserting the optimal Lagrange multipliers into (31) yields the optimal LCMV solution

$$\mathbf{w}_{\text{LCMV}}^{[k]} = \mathbf{R}_x^{[k]-1} \mathbf{C}^{[k]} \left( \mathbf{C}^{[k]H} \mathbf{R}_x^{[k]-1} \mathbf{C}^{[k]} \right)^{-1} \mathbf{c}^{[k]}. \quad (33)$$

#### 4.1.4 Minimum Variance Distortionless Response Beamformer

The MVDR beamformer is a special case of a LCMV beamformer as it uses a specific linear constraint. The MVDR beamformer uses an array steering vector  $\mathbf{a}^{[k]}[\vec{\mathbf{p}}]$ , where  $\vec{\mathbf{p}}$  is a spatial point where the distortion should be zero. The LCMV constraint matrices are set to

$\mathbf{C}^{[k]} = \mathbf{a}^{[k]}[\vec{\mathbf{p}}]$  and  $\mathbf{c}^{[k]} = 1$  for the MVDR beamformer, and this yields the linear constraint  $\mathbf{a}^{[k]}[\vec{\mathbf{p}}]^H \mathbf{w}^{[k]} = 1$ . This constraint ensures that the MVDR beamformer yields a unit array gain in the point  $\vec{\mathbf{p}}$ , hence, a distortionless response in that point. The MVDR-optimal solution is

$$\mathbf{w}_{\text{MVDR}}^{[k]} = \frac{\mathbf{R}_x^{[k]-1} \mathbf{a}^{[k]}[\vec{\mathbf{p}}]}{\mathbf{a}^{[k]}[\vec{\mathbf{p}}]^H \mathbf{R}_x^{[k]-1} \mathbf{a}^{[k]}[\vec{\mathbf{p}}]}. \quad (34)$$

#### 4.1.5 Motivation for Adaptive Beamforming

The optimal beamformers that are outlined here are non-changing, and the methods are therefore sensitive to any changes that may occur in the electroacoustical environment due to, for instance, moving sources or changing sensor characteristics. An alternative solution is to continuously correct the beamformer filter vector in order to adjust for possible changes in the electroacoustic environment.

## 4.2 Adaptive Beamforming

Adaptive beamforming performs continuous adjustments of the beamformer filter weights (see, e.g., [12, 44]), where a change of notation is incorporated by introducing a time-varying filter vector  $\mathbf{w}^{[k]}[n]$ . The filter vector adaptation is typically accommodated by a recursive equation in the form of:

$$\mathbf{w}^{[k]}[n] = \mathbf{w}^{[k]}[n-1] + \Delta_{\mathbf{w}}^{[k]}[n], \quad (35)$$

where  $\Delta_{\mathbf{w}}^{[k]}[n]$  denotes the update vector which corrects the filter vector from the previous iteration, i.e.,  $\mathbf{w}^{[k]}[n-1]$ . The structure of the update vector  $\Delta_{\mathbf{w}}^{[k]}[n]$  depends on the adaptive method used.

### 4.2.1 Adaptive Wiener Filter

The Wiener filter in (20) can normally not be solved directly while a desired signal is not accessible. A solution to this problem is to calibrate

for the desired source signal and to store and reuse the calibration information while the desired source signal is not available [45]. It is assumed that the first source ( $i = 0$ ) is the desired source and that this source is spatially stationary. The corresponding auto-correlation matrix for the first source is estimated during an acquisition phase where only this source is active as:

$$\hat{\mathbf{R}}_{x_0}^{[k]} = \frac{1}{N} \sum_{k=0}^{N-1} \mathbf{x}_0^{[k]}[k] \mathbf{x}_0^{[k]}[k]^H, \quad (36)$$

where  $\mathbf{x}_0^{[k]}[n]$  is defined in (24). In this case, the cross-correlation vector  $\mathbf{r}_{dx}^{[k]}$  that is used in the Wiener formulation is estimated by the first column of  $\hat{\mathbf{R}}_{x_0}^{[k]}$ , denoted as  $\hat{\mathbf{r}}_{dx}^{[k]} = \left[ \hat{\mathbf{R}}_{x_0}^{[k]} \right]_{:,1}$ . Combining the acquired desired auto-correlation matrix and cross-correlation vector with a continuously estimated received auto-correlation matrix  $\hat{\mathbf{R}}_x^{[k]}[n]$  yields the soft-constrained adaptive Wiener filter solution,  $\mathbf{w}_{\text{S.C.}}^{[k]}[n]$ , as

$$\mathbf{w}_{\text{S.C.}}^{[k]}[n] = \left( \hat{\mathbf{R}}_x^{[k]}[n] + \hat{\mathbf{R}}_{x_0}^{[k]} \right)^{-1} \hat{\mathbf{r}}_{dx}^{[k]}. \quad (37)$$

The approach of reusing calibration data corresponds to a soft constraint as it ensures a spatial passband towards the desired source, and all other sources and disturbing noise are regarded as undesired and are therefore attenuated. A practical realization of the adaptive Wiener filter is the Soft-constrained Recursive Least Squares (SC-RLS) beamformer [42, 46]. The SC-RLS structure is sensitive to movements amongst the calibrated desired sources, and an additional source tracking structure is employed in [7] in order to track and to compensate for these movements.

#### 4.2.2 Generalized Sidelobe Canceler

The Generalized Sidelobe Canceler (GSC) uses a beamformer matrix  $\mathbf{B}^{[k]}$  to compute an intermediate signal vector  $\mathbf{u}^{[k]}[n] = \mathbf{B}^{[k]H} \mathbf{x}^{[k]}[n]$ .

The intermediate signal vector  $\mathbf{u}^{[k]}[n]$  is partitioned into two parts:

$$\mathbf{u}^{[k]}[n] = \begin{pmatrix} u_0^{[k]}[n] \\ \mathbf{u}_{\text{block}}^{[k]}[n] \end{pmatrix}. \quad (38)$$

The sub-matrix of  $\mathbf{B}^{[k]}$  that is used to compute the part  $\mathbf{u}_{\text{block}}^{[k]}[n]$  of the intermediate vector is denoted as a *blocking matrix*. The GSC applies an adaptive filter  $\mathbf{w}_{\text{GSC}}^{[k]}[n]$  (size  $M - 1 \times 1$ ) to the blocking part of the intermediate vector according to  $\mathbf{w}_{\text{GSC}}^{[k]}[n]^H \mathbf{u}_{\text{block}}^{[k]}[n]$ , and the adaptive filter aims at minimizing the Mean Square Error (MSE), defined as  $\text{MSE} = \text{E} \left\{ \left| u_0^{[k]}[n] - \mathbf{w}_{\text{GSC}}^{[k]}[n]^H \mathbf{u}_{\text{block}}^{[k]}[n] \right|^2 \right\}$ . Hence, the GSC aims at

cancelling the sidelobes related to  $u_0^{[k]}[n]$  by adapting and subtracting the mainlobes related to  $\mathbf{u}_{\text{block}}^{[k]}[n]$  using the adaptive filter. The GSC can be implemented using, for instance, a Least Mean Squares (LMS) approach or a Recursive Least Squares (RLS) approach. The GSC is a practical realization of the LCMV beamformer, provided that it is implemented by employing an LMS or an RLS algorithm [47].

### 4.2.3 Motivation for Blind Adaptive Beamforming

While the aforementioned structures are adaptive in suppressing interferences and noise, the methods are sensitive to model or calibration errors for the desired sources. In addition, these structures require some explicit reference of the desired source which is impossible to acquire in some applications. As an alternative to these referenced structures are unsupervised or blind structures that do not require such explicit references. The notation *blind* implies that the beamforming is carried out without any required reference signals or any *a priori* knowledge of the acoustic environment. Instead, various statistical assumptions about the source signals are made in blind methods. One class of assumptions is described by the theory of Independent Component Analysis (ICA)

[17, 48, 49, 50] which assumes that the clean source signals are independent. ICA is a rather new theory that has grown in popularity due to its many areas of applicability including speech processing, medical processing, telecommunication processing, etc. (see, e.g., [48, 51, 52]). A brief introduction to ICA is provided next.

### 4.3 Blind Adaptive Beamforming

This section deals with a specific approach of performing array speech enhancement that is based on blind adaptive beamforming. First, a brief introduction to ICA is given. A discussion about the Kurtosis measure of a complex-valued signal is then given. An ICA model based on beamforming notation and a framework for blind beamforming based on Kurtosis maximization are provided. Finally, a discussion on how a blind beamformer is additionally improved by a single-channel method is given.

#### 4.3.1 A Brief Introduction to ICA

A linear ICA data model<sup>5</sup> assumes that a number of  $I$  original and independent source signals  $\mathbf{s}[t] = (s_0[t], s_1[t], \dots, s_{I-1}[t])^T$  are being observed through an invertible source mixture model as described by a *mixing matrix*  $\mathbf{A}$ . The observed signal mixture is  $\mathbf{x}[t] = \mathbf{A}\mathbf{s}[t] \in \mathbb{R}^M$ . Due to the mixing matrix, there is likely dependence between the  $M$  observed signals  $x_m[t]$  in the vector  $\mathbf{x}[t] = (x_0[t], x_1[t], \dots, x_{M-1}[t])^T$ . The ICA model uses an *unmixing solution*  $\mathbf{W} = (\mathbf{w}_0, \mathbf{w}_1, \dots, \mathbf{w}_{M-1})^T \in \mathbb{R}^{M \times M}$  to yield a set of output signals  $\mathbf{y}[t] = \mathbf{W}\mathbf{x}[t] = \mathbf{W}\mathbf{A}\mathbf{s}[t]$ . The ICA task is to find an unmixing solution  $\mathbf{W}$  based on the observed data  $\mathbf{x}[t]$  that makes the unmixed signals  $\mathbf{y}[t] = (y_0[t], y_1[t], \dots, y_{M-1}[t])^T$  statistically independent which implies uncorrelated signals.

<sup>5</sup>In order to simplify the discussion and to illustrate the fundamentals of ICA, it is assumed, only in this section, that the ICA problem applies to real-valued signals. However, the ICA problem is easily extended to the complex-valued case.

In order to preserve readability in the sequel of this section, let  $\mathbf{s}$  denote a vector of stochastic variables  $s_i$ , and a random vector  $\mathbf{x} = \mathbf{A}\mathbf{s}$  is observed. The ICA output signal is  $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{A}\mathbf{s}$ . The ICA task can be formulated by using information theory, described next.

**Mutual information** The *mutual information* of a random vector  $\mathbf{y} = \mathbf{W}\mathbf{x}$  is [49, 53, 54]

$$I(\mathbf{y}) = \sum_{m=0}^{M-1} H(y_m) - H(\mathbf{y}), \quad (39)$$

where the *entropy*  $H(y_m)$  and *joint entropy*  $H(\mathbf{y})$  are defined next. Mutual information can be seen as a measure of how much information each component  $y_m$  has about the other components in the vector  $\mathbf{y}$ . Clearly, if  $I(\mathbf{y}) = 0$ , there is no information between the components in the set, and the components are independent.

**Entropy** Entropy is a measure of structure or randomness of a random variable or vector. The joint entropy of a continuous-valued random vector  $\mathbf{x}$  with joint probability density function  $p_{\mathbf{x}}(\mathbf{x})$  is sometimes referred to as *differential joint entropy*, and it is defined as:

$$H(\mathbf{x}) = - \int p_{\mathbf{x}}(\mathbf{u}) \log p_{\mathbf{x}}(\mathbf{u}) d\mathbf{u}. \quad (40)$$

The joint density of a linear and invertible transformation  $\mathbf{y} = \mathbf{W}\mathbf{x}$ , where  $|\det \mathbf{W}| > 0$  and  $\mathbf{x} = \mathbf{W}^{-1}\mathbf{y}$  is:

$$p_{\mathbf{y}}(\mathbf{y}) = p_{\mathbf{x}}(\mathbf{x}) |\det \mathbf{W}|^{-1}. \quad (41)$$

The joint entropy of the transformation  $\mathbf{y}$  is therefore

$$H(\mathbf{y}) = H(\mathbf{x}) + \log |\det \mathbf{W}|. \quad (42)$$

If it is ensured that the unmixing matrix is unitary, which means that  $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ , then  $|\det \mathbf{W}| = 1$ , i.e.,  $\log |\det \mathbf{W}| = 0$  and the joint entropy is

$$H(\mathbf{y}) = H(\mathbf{x}). \quad (43)$$

**Negentropy** An alternative measure of the structure or randomness of a random variable  $y_m$  is *negentropy*. Negentropy is a measure of Gaussianity, and it is defined for a random variable  $y_m$  in relation to a normalized Gaussian variable  $v$  as

$$J(y_m) = H(v) - H(y_m) \Rightarrow H(y_m) = H(v) - J(y_m). \quad (44)$$

**ICA optimization criterion** Minimizing the mutual information  $I(\mathbf{y})$  in (39) is a way to perform ICA, and it is formulated as an optimization criterion:

$$\mathbf{W}_{\text{ICA}} = \arg \min_{\mathbf{W}} I(\mathbf{y}). \quad (45)$$

When  $I(\mathbf{y})$  is minimal, then the corresponding output components  $y_m$  are made as independent as possible, and the solution that minimizes the mutual information corresponds to the ICA solution  $\mathbf{W}_{\text{ICA}}$ . However, by using the fact that  $\mathbf{W}^T \mathbf{W} = \mathbf{I}$ , and by inserting (43) and (44) into (39), the mutual information of  $\mathbf{y}$  is rendered according to

$$I(\mathbf{y}) = -C - \sum_{m=0}^{M-1} J(y_m), \quad (46)$$

where  $C = H(\mathbf{x}) - MH(v)$ . The minimization of mutual information can simply omit the term  $C$  as it is a constant term that is not dependent on  $\mathbf{W}$ . Hence, an alternative and equivalent optimization criterion for ICA is therefore

$$\mathbf{W}_{\text{ICA}} = \arg \max_{\mathbf{W}} \sum_{m=0}^{M-1} J(y_m), \text{ subject to } \mathbf{W}^T \mathbf{W} = \mathbf{I}. \quad (47)$$

Under these circumstances, the minimization of mutual information is equivalent to the maximization of negentropy, and a solution to the ICA task is to make the variables  $y_m$  maximally non-Gaussian under the unitary constraint for  $\mathbf{W}$ .

**Partial ICA optimization criterion** A partial ICA solution is found if the unmixing solution is one column vector  $\mathbf{w}_m$  of the full unmixing matrix, i.e.,  $y_m = \mathbf{w}_m^T \mathbf{x}$ , in which case the partial ICA optimization criterion is

$$\mathbf{w}_{\text{ICA}} = \arg \max_{\mathbf{w}_m} J(y_m), \text{ subject to } \mathbf{w}_m^T \mathbf{w}_m = 1. \quad (48)$$

**Approximation of negentropy** Negentropy can be approximated by higher-order moments (cumulants) such as the Kurtosis value of the variable [49]. The Kurtosis measure of a real-valued variable  $y_m$  is defined as

$$\kappa(y_m) = \text{E} \{y_m^4\} - 3\text{E} \{y_m^2\}^2. \quad (49)$$

While  $\kappa(v) = 0$  for a Gaussian variable  $v$ , it stands clear that maximization of the Kurtosis measure is therefore one approach to solving the ICA problem, and the partial ICA optimization criterion is

$$\mathbf{w}_{\text{ICA}} = \arg \max_{\mathbf{w}_m} \kappa(y_m), \text{ subject to } \mathbf{w}_m^T \mathbf{w}_m = 1. \quad (50)$$

The approach to maximize the Kurtosis measure in order to perform ICA has found a variety of applications related to speech enhancement including source separation [55], speech separation [56], and speech de-reverberation [57]. However, the Kurtosis maximization approach used in this work regards speech enhancement by adaptive blind beamforming [58].

### 4.3.2 Kurtosis Measure of a Complex-valued Signal

From the standpoint of the theory of ICA, it is clear that the Kurtosis measure can be used as a discriminative measure for assessing the Gaussianity of a signal. The definition of the Kurtosis measure of a complex-valued stochastic process has 16 variants [59], and one particular variant of the Kurtosis measure for a complex-valued signal  $s_i^{[k]}[n]$

has been used extensively:

$$\kappa \left\{ s_i^{[k]}[n] \right\} = \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^4 \right\} - 2 \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\}^2 - \left| \mathbb{E} \left\{ s_i^{[k]}[n]^2 \right\} \right|^2. \quad (51)$$

If the real and imaginary parts of  $s_i^{[k]}[n]$  are uncorrelated and have the same variance, then  $\mathbb{E} \left\{ s_i^{[k]}[n]^2 \right\} = 0$  and the signal is said to have a circular distribution. The Kurtosis value for a signal that has a circular distribution is

$$\kappa \left\{ s_i^{[k]}[n] \right\} = \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^4 \right\} - 2 \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\}^2. \quad (52)$$

**Source signal circularity assumption** In order to assert the circularity assumption, the Kurtosis measure in (51) is normalized by the square subband power and averaged over the subbands. The Kurtosis measure that assumes a circular distribution in (52) is also normalized by the square subband power and averaged over the subbands. If the circularity assumption holds for a test signal, then the two normalized and averaged Kurtosis measures should be equal. The two normalized Kurtosis measures are averaged for 20 different speech signals, 20 white Gaussian noise signals, and 20 mixtures of speech and noise. The sampling frequency is 8 kHz. The measures are presented in Figure 7 for various filter bank configurations, i.e., number of subbands, over-sampling ratio, and number of taps in the prototype filter (see Section 5 for details regarding the filter bank). The circularity assumption is valid for signals with a high Kurtosis value, i.e., the target speech signals and the speech plus noise mixtures. The circularity assumption influences only the noise signals that have a low Kurtosis value. Either of the Kurtosis measures (51) and (52) provide a distinguishing measure between the target signals, i.e., speech, and the undesired signals, i.e., noise. Based on the results presented here, it can be assumed that the evaluated subband signals have circular distributions.

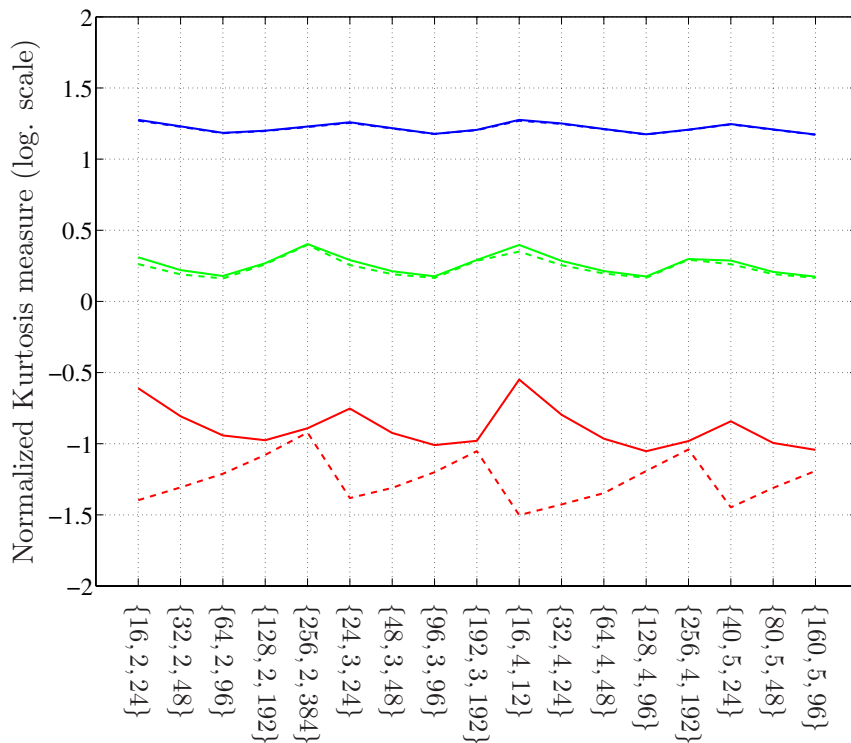


Figure 7: Normalized Kurtosis measure (solid) and normalized Kurtosis measure that assumes a circular distribution (dashed). The measures are averaged for 20 speech signals (blue), 20 white noise signals (red), and 20 mixtures of speech and noise (green). X-axis:  $\{K, O, N\}$ , where  $K$  is the number of subbands,  $O$  is the over-sampling ratio, and  $N$  is the length of the prototype filter. Observe that the influence of a circularity assumption is low except for the signals with a low Kurtosis value.

**Kurtosis values of speech and noise** The use of the Kurtosis measure in ICA and BSE is validated due to the fact that a speech signal has a higher Kurtosis value than a Gaussian noise signal [60]. The Kurtosis measure in (52), normalized by the square subband power, is averaged for 20 speech source signals, 20 white Gaussian noise signals, and 20 mixtures of speech and noise. The sampling frequency is 8 kHz. The normalized Kurtosis measure is presented in Figure 8. The filter bank configuration uses  $K = 128$  subbands, a ratio  $O = 2$  over-sampling, and it uses  $N = 192$  taps in the prototype filter. A Gaussian signal has a very low Kurtosis value (theoretically equal to zero) while a speech signal has a much higher Kurtosis value. The mixture of speech and noise has a Kurtosis value that lies between that of the speech and that of the noise. Hence, these results practically validate the use of the Kurtosis measure as a discriminating function in ICA.

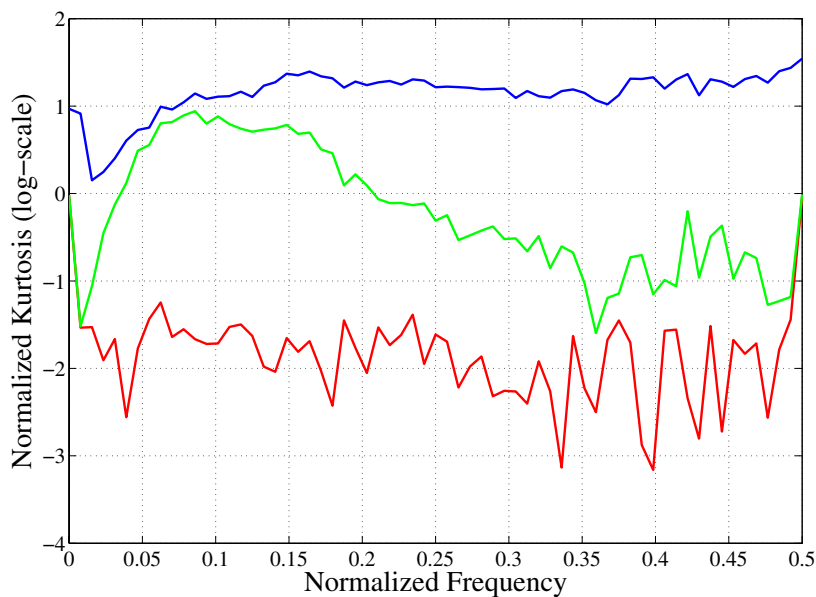


Figure 8: The normalized Kurtosis measure, defined for a signal  $x^{[k]}[n]$  as  $E\{|x^{[k]}[n]|^4\}/E\{|x^{[k]}[n]|^2\}^2 - 2$ , is averaged for 20 speech signals (**blue**), 20 white Gaussian noise signals (**red**), and 20 mixtures of speech and noise (**green**). Observe that the Kurtosis value of the speech is higher than the Kurtosis value of the noise. The Kurtosis value of the mixture of speech and noise lies between the Kurtosis values of the speech and the noise. The signals are around 4.5 seconds in duration.

**Critique towards the Kurtosis measure and a practical solution** A common critique of the Kurtosis measure is that it is not viable to be used as a contrast function in ICA as it is sensitive towards data outliers. There is a practical solution to circumvent the reported sensitivity of the Kurtosis measure. In a practical application, microphones are used to capture sounds in the acoustic environment. The low-level microphone signals are amplified by a signal amplifier in the capturing hardware before the signals are converted to a digital representation. A rule of thumb when designing a digital system is to adjust the microphone signal amplifier so that as much as possible of the digital bits in a digital representation are utilized at normal signal levels. This is sometimes referred to as a good utilization of the system's dynamic range. Practically, any extreme microphone signal value that lies outside the voltage range of the sound capturing system is limited to the operating voltage of the system. Hence, a practical solution to the outlier problem is to utilize a good dynamic range in which case the sound capturing hardware limits any outlier and therefore rectifies the critique towards the Kurtosis measure. The practical solution to the outlier problem is illustrated in Figure 9 where the outliers are limited by the sound capturing hardware.

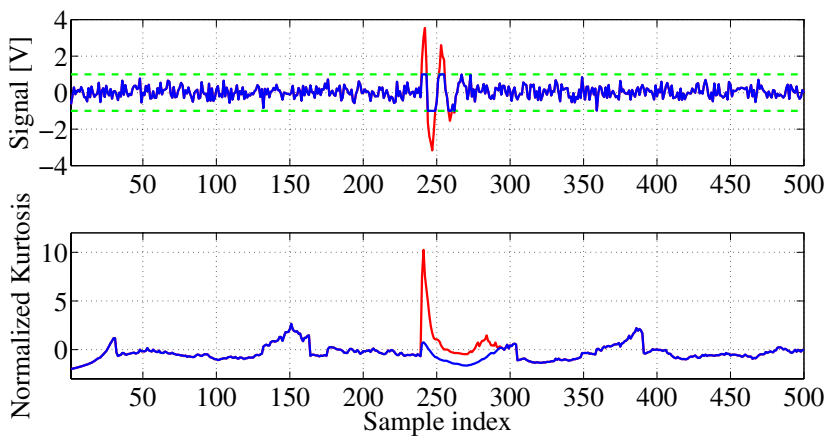


Figure 9: Upper: A white Gaussian signal (**blue**) added with a short-duration outlying impulse (**red**) that have been limited to the voltage range of the sound capturing system (**green**). The lower plot shows the normalized Kurtosis value for the signal with suppressed outliers (**blue**) and the signal with outliers (**red**).

### 4.3.3 An ICA Model using Beamforming Notation

The source signal model that is used in this work (3) is also used as a mixing model for the complex-valued ICA problem by incorporating a vector notation:

$$\begin{aligned} x_m^{[k]}[n] &= \sum_{i=0}^{I-1} \sum_{\tau=0}^{T_K-1} a_{m,i}^{[k]} s_i^{[k]}[n-\tau] + v_m^{[k]}[n] = \\ &= \sum_{i=0}^{I-1} \mathbf{a}_{m,i}^{[k]} \mathbf{s}_i^{[k]}[n] + v_m^{[k]}[n], \end{aligned} \quad (53)$$

$$\mathbf{x}^{[k]}[n] = \begin{pmatrix} x_0^{[k]}[n] \\ x_1^{[k]}[n] \\ \vdots \\ x_{M-1}^{[k]}[n] \end{pmatrix} = \mathbf{A}^{[k]} \mathbf{s}^{[k]}[n] + \mathbf{v}^{[k]}[n], \quad (54)$$

where  $\mathbf{v}^{[k]}[n] = (v_0^{[k]}[n], v_1^{[k]}[n], \dots, v_{M-1}^{[k]}[n])^T$ . The source-mixing matrix  $\mathbf{A}^{[k]}$  (size  $M \times I \cdot T_K$ ) and source signal vector  $\mathbf{s}^{[k]}[n]$  (size  $I \cdot T_K \times 1$ ) are defined as

$$\mathbf{A}^{[k]} = \begin{pmatrix} \mathbf{a}_{0,0}^{[k]} & \mathbf{a}_{0,1}^{[k]} & \cdots & \mathbf{a}_{0,I-1}^{[k]} \\ \mathbf{a}_{1,0}^{[k]} & \mathbf{a}_{1,1}^{[k]} & \cdots & \mathbf{a}_{1,I-1}^{[k]} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{a}_{M-1,0}^{[k]} & \mathbf{a}_{M-1,1}^{[k]} & \cdots & \mathbf{a}_{M-1,I-1}^{[k]} \end{pmatrix}, \quad (55)$$

$$\mathbf{a}_{m,i}^{[k]} = (a_{m,i}^{[k]}[0], a_{m,i}^{[k]}[1], \dots, a_{m,i}^{[k]}[T_K - 1]), \quad (56)$$

$$\mathbf{s}^{[k]}[n] = \left( \mathbf{s}_0^{[k]}[n]^T, \mathbf{s}_1^{[k]}[n]^T, \dots, \mathbf{s}_{I-1}^{[k]}[n]^T \right)^T, \quad (57)$$

$$\mathbf{s}_i^{[k]}[n] = \left( s_i^{[k]}[n], s_i^{[k]}[n-1], \dots, s_i^{[k]}[n-T_K+1] \right)^T. \quad (58)$$

The  $I$  source signals  $s_i^{[k]}[n]$  are assumed to be independent in the ICA framework, and the ICA-task is to find an unmixing filter  $\mathbf{w}^{[k]}$  so that

one of the  $I$  independent components is extracted, i.e.,

$$\begin{aligned} y^{[k]}[n] &= \mathbf{w}^{[k]H} \mathbf{x}^{[k]}[n] = \mathbf{w}^{[k]H} \mathbf{A}^{[k]} \mathbf{s}^{[k]}[n] + \mathbf{w}^{[k]H} \mathbf{v}^{[k]}[n] = \\ &= \alpha_i s_i^{[k]}[n - \tau_i] + \mathbf{w}^{[k]H} \mathbf{v}^{[k]}[n], \end{aligned} \quad (59)$$

where  $\alpha_i$  and  $\tau_i$  represent a scaling ambiguity and a delay due to the unmixing process. The ICA mixing and separation model of independent sources is illustrated in Figure 10. The ICA methods as described in this thesis cannot directly suppress the disturbing noise  $\mathbf{v}^{[k]}[n]$ . However, an indirect method is proposed in Part IV where an additional single-channel speech enhancer is combined with the beamformer. It stands clear that the ICA formulation of unmixing an observed mixture of signals in (59) is identical to the beamforming model in (14), and this model has been the core of most of the work presented in the thesis.

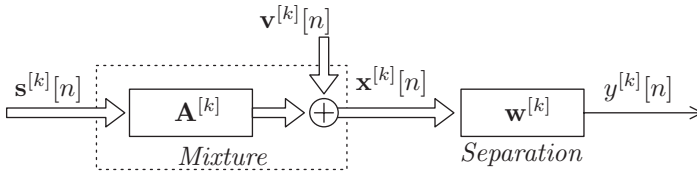


Figure 10: Noisy Independent Component Analysis (ICA) mixture and separation model.

#### 4.3.4 Blind Beamforming by Kurtosis Maximization

Performing ICA by maximization of the Kurtosis measure for real-valued mixtures can be directly extended to the case of complex-valued ICA (see, e.g., [61, 62, 63]).

Now, consider the ICA model in (59) which is equivalent to a beamformer model. Let us assume that one of two sources is a speech source and the other source is an interfering Gaussian noise source. In addition, also assume that the beamformer filter weights  $\mathbf{w}^{[k]}$  are adjusted so

that the Kurtosis value at the beamformer output  $\kappa \{y^{[k]}[n]\}$  is maximized. A speech source has generally a higher Kurtosis value than other signals (see Section 4.3.2). Hence, if  $\kappa \{y^{[k]}[n]\}$  is maximized, it implies that the desired speech source is extracted. The extraction of a source signal in BSE is practically achieved when the spatial nulls of the beamformer FIR-filter are in the directions of the interfering noise sources. This means that, at most,  $M - 1$  noise sources can be rejected unless the beamformer uses several taps per filter. The use of several taps in a beamformer filter provides an additional possibility to suppress interfering noise in the temporal domain, leading to an increase in the speech extraction performance in general.

The optimization problem for blind beamforming by Kurtosis maximization is formulated as

$$\max_{\mathbf{w}^{[k]}} \kappa \{y^{[k]}[n]\}, \text{ subject to } \|\mathbf{w}^{[k]}\|_2 = 1. \quad (60)$$

The normalization constraint, i.e.,  $\|\mathbf{w}^{[k]}\|_2 = 1$ , can be imposed directly onto the optimization approach by using, for instance, Kuhn-Tucker conditions (see, e.g., [61, 62]). However, a common approach in the iterative online optimization of this criterion is to perform a post-normalization of the weights after each iteration (see, e.g., [48]) as this enforces the unity norm onto the updated weight set. One approach in iterative search methods uses first and second order derivatives of the objective criterion. A discussion on the theory and necessary conditions for the existence of a gradient and a Hessian matrix for a real-valued function of complex valued variables is found in, e.g., [63, 64, 65].

**Two recursive optimization strategies** Two optimization strategies to solve the optimization in (60) recursively are proposed in this thesis. The underlying idea of the strategies is to render an optimization solution that can be robustly implemented and realized in hardware.

One of the most popular methods today for performing BSE is the FastICA method [48, 49, 61, 62]. The popularity of FastICA is due

to its flexibility in the choice of an activation function, i.e., nonlinearity, and its extremely fast convergence for most cases. In fact, for the Kurtosis nonlinearity, the FastICA provides a cubic and global convergence [66]. However, one fundamental problem with FastICA is that the method diverges for Gaussian-only signal mixtures. In some real applications, it cannot be guaranteed that the observed signal mixture is always non-Gaussian, and the FastICA is therefore not advisable in real applications. An alternative approach to solve the ICA problem is introduced in Part I of the thesis. The proposed method shares all elementary parts of FastICA. The proposed method is shown to not diverge for Gaussian-only source mixtures and is therefore viable to use in a real application. The proposed method shares important and advantageous benefits with FastICA, such as the fixed-point property and the global convergence. The method proposed in Part I and the FastICA method use an approximative Newton technique in which some statistical expectations related to the Hessian matrix are approximated and simplified.

An alternative approach is presented in Parts II to V of this thesis, where a local approximation of the Kurtosis measure is constructed and then solved. The local approximation regards a quadratic approximation of the Kurtosis measure. It is shown in Part II that this approximation technique does not diverge for Gaussian-only source mixtures, and it is therefore also viable to use in a real application. In fact, when the method in Part II has converged to a desired solution, it has a zero approximation error which is a strong benefit of this algorithm.

**Online processing versus data batch processing** The optimization problem in (60) can be solved either in a batch processing mode or in an online mode. Batch processing starts by capturing a certain amount of data. Required statistical expectations are then approximated and computed for the entire data batch, and the optimization is solved in an iterative manner until a stopping criterion is met. The FastICA method is one example of a method that is well-suited for a

batch processing approach [61]. The batch approach requires the storage of the entire data batch in memory during batch-iterations. This yields a high memory consumption and a large number of required computations, and this can be a limitation in a practical implementation that may be restrained with respect to the amount of memory and available processing power. An alternative is found in the online approach where statistical expectations are estimated continuously as new data is available and the beamformer filter is repeatedly updated. The lion's share of the work presented in the thesis (Parts I to V) focuses on BSE using the Kurtosis maximization approach in an online framework [6, 67, 68, 69, 70].

#### 4.3.5 Post-processing versus In-the-Loop Processing

Blind beamformers that are based on a unity norm-constraint for the beamformer weights operate only in the spatial domain, leaving the temporal domain essentially unprocessed. In addition to the pure beamforming task, some suggest applying a temporal post-processing method, such as spectral subtraction, to the output signals of the blind beamforming structure in order to improve the temporal speech enhancement performance (see, e.g., [71, 72]).

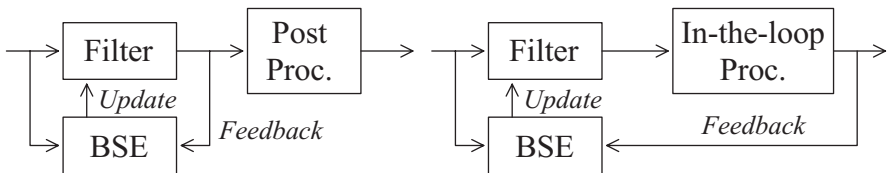


Figure 11: Post-processing (left) versus in-the-loop processing (right) as a means to improve Blind Signal Extraction (BSE) performance.

A fundamentally different approach is introduced in Part IV of this thesis where a speech enhancer is incorporated into the control loop of a blind beamformer, see Figure 11. The *in-the-loop* speech enhancer works in symbiosis with the blind beamformer while the spatial pro-

cessing of the beamformer aids the temporal processing of the speech enhancer, and vice versa. This approach has been shown to be advantageous over the normal post-processing approach, and it yields better speech enhancement performance, especially for a low SIR [70].

Due to the fact that nonlinearities are applied to the input data in a BSE method, the performance of a blind beamformer is connected to the magnitude of the incoming signal. Part V of this thesis elaborates on the introduction of an Automatic Gain Control (AGC) unit into the control loop of a blind beamformer. An AGC is an adaptive algorithm that ensures a uniform magnitude of the processed signal. The use of an AGC in Part V is novel, and it makes the blind beamformer invariant to various input signal magnitudes, i.e., the performance of the blind beamformer with an AGC is not connected to the input signal magnitude and this is a strong benefit in this approach.

## 5 Realtime Speech Enhancement

All parts of this thesis, except for Part II which deals with a theoretical analysis, have been implemented and verified using realtime signal processing hardware. The virtue of implementing a speech enhancer in realtime lies in the possibility of being able to test the algorithm and the implementation using a realistic end-user scenario. A realtime implementation also allows for industry partners, research colleagues, decision makers, and funding agency personnel to test the algorithm and implementation using their own subjective judgement. In addition, a realtime implementation allows the researcher to quickly evaluate and adjust new ideas before commencing costly and tedious subjective measurement campaigns.

This section deals with some aspects regarding the realtime implementation of a speech enhancer. The focus in this section is on discrete-time signal processing as discrete-time signal processing satisfies a majority of end-user applications. Although it is suitable in some cases to perform a continuous-time signal processing (see Part VI of this thesis), it is omitted from this section. Realtime aspects regarding the implementation of speech enhancement structures on a Digital Signal Processor (DSP) and on a realtime MATLAB<sup>6</sup> framework are given. While filter banks are central in all parts of this thesis, the uniform Discrete Fourier Transform (DFT) modulated filter bank is described at the end of this section together with a discussion regarding its implementation.

### 5.1 Realtime Speech Enhancement using DSP

A DSP is distinguished from other digital processors as it is optimized for performing a large number of arithmetic and logical processing operations on digital signals. Aside from the mainstream DSP series, there are Application Specific Integrated Circuits (ASIC) which allow a detailed level of control of the processing together with a massive

---

<sup>6</sup>MATLAB is a registered trademark for Mathworks, Inc.

parallelism. Mixed-signal processors and mixed-signal microcontrollers provide an entire system-on-chip with analogue signal processing that runs parallel to the digital processing core. System-on-chip solutions typically have a low footprint. In addition, there are multiple-core DSP structures that have slave-DSP cores that are optimized for certain specific tasks such as filter bank processing. A DSP system uses an analogue-to-digital converter in order to represent analogue signals in a digital format that can be processed in the DSP. The conversion system starts by sampling the analogue signal at specific sampling time points. A quantization thereafter assigns digital (binary) numbers to the sample values, resulting in a stream of digital values that need to be processed. There are two main methods used for representing a digital value: the fixed-point numerical format and the floating-point numerical format, described next.

### 5.1.1 Fixed-Point Representation

The notation *fixed-point* implies that the decimal point (also called the radix point) of a digital number is at a fixed location. Fixed-point arithmetic operations require constant supervision so that the numerical range is not violated. A speech enhancer implementation on fixed-point arithmetic requires therefore careful and constant supervision of the used dynamical range in order to sustain a high level of enhanced speech quality. If the dynamical range is violated, proper counter-measures, such as saturation handling, are required.

### 5.1.2 Floating-Point Representation

As opposed to fixed-point representation, *floating-point* representation covers a much larger numerical range through its versatile number format where the decimal point is variable. Floating-point arithmetic is preferable in many signal processing applications as the dynamical range in the arithmetic operations is preserved. Implementing a signal processing algorithm in floating-point arithmetic is therefore straight-

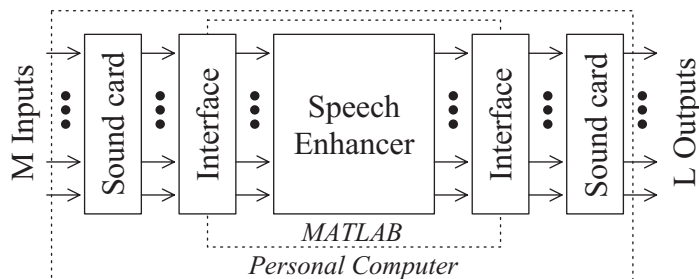


Figure 12: A MATLAB framework using a sound card that is a part of a personal computer for real-time speech enhancement.

forward. However, floating-point arithmetic operations require more control logic than corresponding fixed-point arithmetic operations in order to keep track of the varying mantissa and exponent. The power consumption is therefore higher in floating-point systems than in fixed-point systems. For this reason, fixed-point arithmetic can often be seen in industrially oriented signal processor systems where low power consumption is important.

## 5.2 Realtime Speech Enhancement in MATLAB

It is often beneficial in applied research as well as in education to implement speech enhancement algorithms using a high-level programming language such as MATLAB. MATLAB is well-known within engineering. MATLAB is also easy to get started with, and the dense and well-documented signal processing packages simplify the advanced processing of signals. It is possible to interface a computer sound card that is a part of an ordinary personal computer with MATLAB in order to construct a real-time speech enhancement system, see Figure 12. Here, the number of inputs  $M$  and outputs  $L$  in the MATLAB framework is determined by the number of inputs and outputs supported by the sound card. The MATLAB framework presents a block of input data (of size  $B \times M$ ) from the sound card driver to a user-defined MAT-

LAB script and a block of output data (of size  $B \times L$ ) is prepared for the sound card output driver. The block length  $B$  is determined by the sound card driver, and it may in many cases be adjusted to suit the delay requirements of the specific application. The task of the speech enhancement MATLAB script is to compute the output data block from the input data block. This computation must be carried out within the duration of one data block in order to maintain a constant flow of data to the output.

### 5.3 Achieving High Realtime Performance

Modern DSPs support effective computations through the Single Instruction Multiple Data (SIMD) mode which means that the same instruction is performed for different data in parallel. The SIMD mode is suitable for vector operations such as element-wise vector multiplications or vector inner products. The vector approach is also the approach used in MATLAB, a software that is optimized for vector and matrix manipulations. In order to facilitate the efficient vector-based processing approach, the data has to be stored in a transformed approach in the system memory. Operations on data that is stored in a transformed approach require a more special treatment compared to the ordinary case. In order to illustrate the transformed approach, consider the subband signal vectors  $\mathbf{x}^{[k]} = (x_0^{[k]}, x_1^{[k]}, \dots, x_{M-1}^{[k]})^T$  of size  $M \times 1$ , where  $k : k \in \mathbb{N}, k < K$ . The task is to compute the subband outer products  $\mathbf{R}^{[k]} = \mathbf{x}^{[k]} \mathbf{x}^{[k]H}$  of size  $M \times M$  for all subbands  $k$ . This operation requires a loop that runs over all  $K$  subbands and that computes an  $M \times M$  outer product per subband. Now consider a transformed storage of the data. The subband data is stored in the vectors  $\mathbf{x}'_m = (x_m^{[0]}, x_m^{[1]}, \dots, x_m^{[K-1]})^T$  of size  $K \times 1$ , where  $m : m \in \mathbb{N}, m < M$ . A similar outer product is then computed for the transformed data as  $\mathbf{R}'_{m,p} = \mathbf{x}'_m \bullet \mathbf{x}'_p^*$  of size  $K \times 1$  for  $p, m \in \{0, 1, \dots, M-1\}$  and where  $\bullet$  denotes an element-wise multiplication. This operation requires a loop that runs  $M^2$  iterations, and

one  $K$ -sized element-wise multiplication is computed per loop iteration. If  $K > M^2$ , it means that the transformed approach requires less loop iterations than the ordinary approach. In addition, the optimized vector-processing in a DSP and in MATLAB is utilized more efficiently in the transformed approach. The output data of the ordinary approach is identical to the output data of the transformed approach, i.e.,  $\mathbf{R}'_{m,p} = \left( [\mathbf{R}^{[0]}]_{m,p}, [\mathbf{R}^{[1]}]_{m,p}, \dots, [\mathbf{R}^{[K-1]}]_{m,p} \right)^T$ . This alternative way of managing the data was introduced in [67] for a MATLAB-based realtime implementation. The transformation approach reduces the number of computations in a high-performance floating point DSP platform by 7.5 times for a typical vector operation [46].

#### 5.4 Multi-rate Filter Bank

The adopted filter bank structure uses a multi-rate representation of the data, i.e., decimators and interpolators are used in the filter bank structure [13, 14], see Figure 13. The bandpass filters in the filter bank are constructed by modulating corresponding analysis and synthesis prototype filters. An efficient structure uses a polyphase representation of the modulated prototype filters.

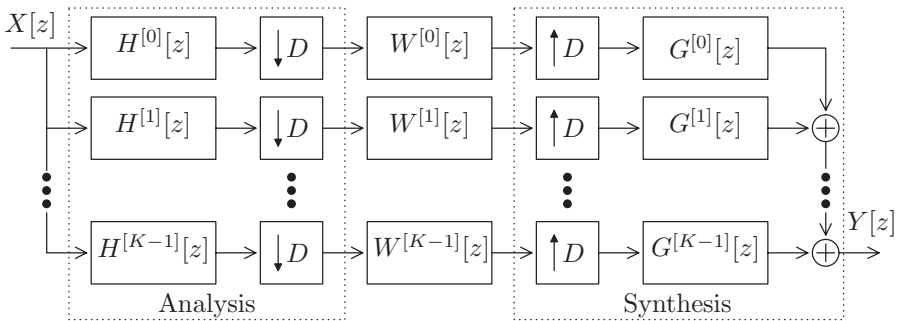


Figure 13: Multi-rate filter bank with  $K$  channels and subband filtering.

### 5.4.1 Input-Output Signal Assembly

The bank of bandpass filters for the analysis part and the synthesis part are denoted as  $H^{[k]}[z]$  and  $G^{[k]}[z]$ , respectively. The corresponding impulse response functions are  $h^{[k]}[t]$  and  $g^{[k]}[t]$ . The subband signals  $X^{[k]}[z]$  are constructed by filtering an input signal  $X[z]$  with each analysis bandpass filter  $H^{[k]}[z]$  and then decimating a factor  $D < K$ :

$$X^{[k]}[z] = \frac{1}{D} \sum_{d=0}^{D-1} H^{[k]}[W_D^d z^{\frac{1}{D}}] X[W_D^d z^{\frac{1}{D}}], \quad (61)$$

where  $W_D = e^{-j\frac{2\pi}{D}}$ . The sum in the expression above represents each of the  $D$  aliasing terms that arise in the decimation process. The subband output signals  $Y^{[k]}[z]$  are filtered versions of the subband input signals according to

$$\begin{aligned} Y^{[k]}[z] &= W^{[k]}[z] X^{[k]}[z] = \\ &= \frac{1}{D} W^{[k]}[z] \sum_{d=0}^{D-1} H^{[k]}[W_D^d z^{\frac{1}{D}}] X[W_D^d z^{\frac{1}{D}}]. \end{aligned} \quad (62)$$

The reconstructed output signal  $Y[z]$  corresponds to the sum of synthesis filtered subband output signals interpolated a factor  $D$ :

$$Y[z] = \sum_{d=0}^{D-1} \sum_{k=0}^{K-1} T_d^{[k]}[z] W^{[k]}[z^D] X[W_D^d z], \quad (63)$$

$$T_d^{[k]}[z] = G^{[k]}[z] H^{[k]}[W_D^d z]. \quad (64)$$

The filters  $T_d^{[k]}[z]$  correspond to the transfer functions related to the  $d^{\text{th}}$  aliasing term, and the term  $\sum_{k=0}^{K-1} T_0^{[k]}[z]$  gives the desired part of the output signal. The analysis and synthesis filters can be designed in various ways so as to optimize selected criteria regarding, for instance, minimum passband ripple, maximum stopband attenuation, minimum aliasing distortion, or minimum total group delay (see, e.g., [19]).

### 5.4.2 Analysis Prototype Filter Polyphase Implementation

This thesis uses analysis bandpass filters  $h^{[k]}[t]$  that are modulated versions of the lowpass prototype filter  $h^{[0]}[t]$ . The length of the prototype filter is  $T = PD$  taps, where  $P$  is an integer value. The analysis bandpass filters are constructed as

$$h^{[k]}[t] = W_K^{-kt} h^{[0]}[t] \Rightarrow H^{[k]}[z] = H^{[0]}[W_K^k z]. \quad (65)$$

The analysis prototype filter  $h^{[0]}[t]$  is represented by  $D$  polyphase components:

$$e_d[p] = h^{[0]}[d + pD], \quad (66)$$

for  $d : d \in \mathbb{N}, d < D$  and  $p : p \in \mathbb{N}, p < P$ . The  $\mathcal{Z}$ -transform of each polyphase component is

$$E_d[z] = \sum_{p=0}^{P-1} e_d[p] z^{-p} = \sum_{p=0}^{P-1} h^{[0]}[d + pD] z^{-p}. \quad (67)$$

The analysis prototype filter is therefore

$$H^{[0]}[z] = \sum_{d=0}^{D-1} z^{-d} E_d[z^D] = \sum_{d=0}^{D-1} z^{-d} \sum_{p=0}^{P-1} h^{[0]}[d + pD] z^{-pD}. \quad (68)$$

The modulated analysis filters are

$$H^{[k]}[z] = \sum_{d=0}^{D-1} W_K^{-dk} z^{-d} \sum_{p=0}^{P-1} h^{[0]}[d + pD] W_K^{-pkD} z^{-pD}. \quad (69)$$

The  $K$  subbands are grouped into  $O$  groups, where subband group  $o$  contains subbands with index  $k = o + iO$ , for  $o : o \in \mathbb{N}, o < O$  and  $i : i \in \mathbb{N}, i < D$ . Hence, the subband  $k = o + iO$  in group  $o$  and group

element  $i$  is

$$\begin{aligned}
 H^{[o+iO]}[z] &= \sum_{d=0}^{D-1} W_D^{-di} z^{-d} \underbrace{\sum_{p=0}^{P-1} \underbrace{W_K^{-o(d+pD)} h^{[0]}[d+pD]}_{=e_{d,o}[p]} z^{-pD}}_{=E_{d,o}[z^D]} = \\
 &= \sum_{d=0}^{D-1} W_D^{-di} z^{-d} E_{d,o}[z^D] = \\
 &= \left(1, W_D^{-i}, \dots, W_D^{-(D-1)i}\right) \begin{pmatrix} E_{0,o}[z^D] \\ z^{-1} E_{1,o}[z^D] \\ \vdots \\ z^{-(D-1)} E_{D-1,o}[z^D] \end{pmatrix}. \tag{70}
 \end{aligned}$$

The term  $e_{d,o}[p]$  corresponds to the  $d^{\text{th}}$  analysis modulated polyphase filter in group  $o$ , and  $E_{d,o}[z]$  is the corresponding  $Z$ -transform. All analysis filters in group  $o$  are stacked in the vector  $\mathbf{H}^{[o]}[z]$  as

$$\mathbf{H}^{[o]}[z] = \begin{pmatrix} H^{[o]}[z] \\ H^{[O+o]}[z] \\ \vdots \\ H^{[K-O+o]}[z] \end{pmatrix} = DD_D^{-1} \begin{pmatrix} E_{0,o}[z^D] \\ z^{-1} E_{1,o}[z^D] \\ \vdots \\ z^{-(D-1)} E_{D-1,o}[z^D] \end{pmatrix}, \tag{71}$$

where  $\mathbf{D}_D^{-1}$  is a  $D \times D$  inverse DFT matrix. The analysis part of a uniform DFT modulated filter bank with  $K$  channels and a factor  $D$  decimation is shown in Figure 14.

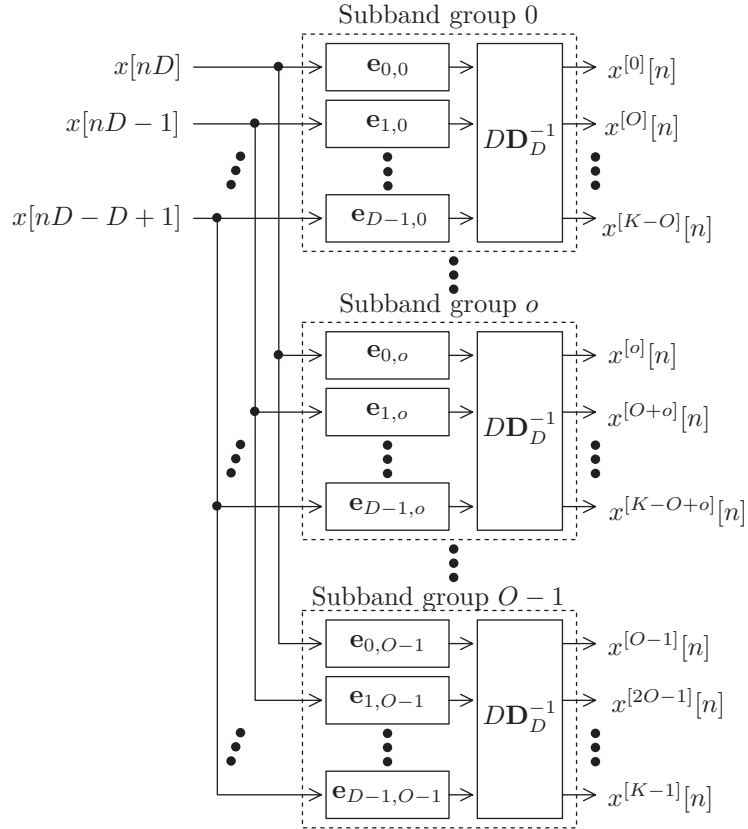


Figure 14: Analysis part of a uniform Discrete Fourier Transform (DFT) modulated filter bank with  $K$  channels and a factor  $D$  decimation. The polyphase filters are given by  $\mathbf{e}_{d,o} = (e_{d,o}[0], e_{d,o}[1], \dots, e_{d,o}[P-1])^T$  and  $\mathbf{D}_D^{-1}$  is the inverse DFT matrix.

### 5.4.3 Synthesis Prototype Filter Polyphase Implementation

The synthesis modulated polyphase filter of a uniform DFT modulated filter bank is in this work the time-reversed and conjugated corresponding analysis modulated polyphase filter. In this case, the synthesis polyphase components for subband group  $o$  and index  $d$  are

$$f_{d,o}[p] = e_{d,o}[P-1-p]^*. \quad (72)$$

The  $Z$ -transform of each modulated synthesis polyphase filter is

$$F_{d,o}[z] = \sum_{p=0}^{P-1} f_{d,o}[p]z^{-p}. \quad (73)$$

The corresponding modulated synthesis filters for subband index  $k = o + iO$  are

$$\begin{aligned} G^{[o+iO]}[z] &= \sum_{d=0}^{D-1} W_D^{di} z^{-(D-1-d)} \sum_{p=0}^{P-1} f_{d,o}[p]z^{-pD} = \\ &= \sum_{d=0}^{D-1} W_D^{di} z^{-(D-1-d)} F_{d,o}[z^D]. \end{aligned} \quad (74)$$

The synthesis filters in group  $o$  are stacked in the vector  $\mathbf{G}^{[o]}[z]$  as

$$\mathbf{G}^{[o]}[z] = \begin{pmatrix} G^{[o]}[z] \\ G^{[O+o]}[z] \\ \vdots \\ G^{[K-O+o]}[z] \end{pmatrix} = \mathbf{D}_D \begin{pmatrix} z^{-(D-1)} F_{0,o}[z^D] \\ z^{-(D-2)} F_{1,o}[z^D] \\ \vdots \\ F_{D-1,o}[z^D] \end{pmatrix}. \quad (75)$$

The synthesis part of a uniform DFT modulated filter bank with  $K$  channels and a factor  $D$  interpolation is shown in Figure 15.

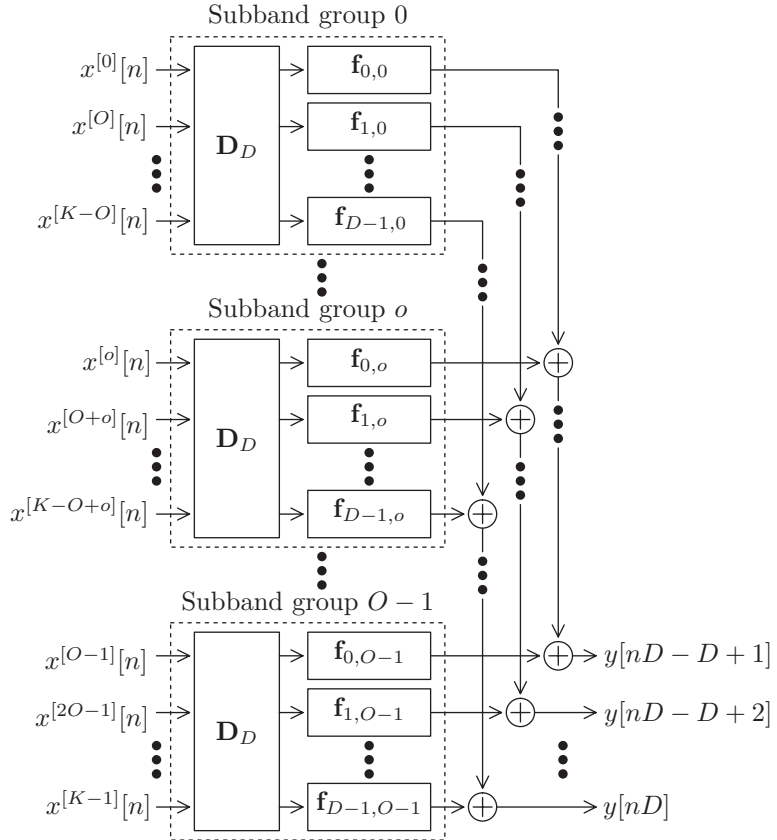


Figure 15: Synthesis part of a uniform Discrete Fourier Transform (DFT) modulated filter bank with  $K$  channels and a factor  $D$  decimation. The polyphase filters are given by  $\mathbf{f}_{d,o} = (f_{d,o}[0], f_{d,o}[1], \dots, f_{d,o}[P-1])^T$  and  $\mathbf{D}_D$  is the DFT matrix.

#### 5.4.4 Efficient Filter Bank Implementation

The proposed filter bank structure utilizes a set of subband index groups. The approach to group subbands allows a flexible configuration in the number of subbands  $K$  and the oversampling ratio  $O = K/D$ . The number of data samples that are needed before the filter bank conversion can commence is equal to the decimation ratio  $D$ . The size  $D$  also corresponds to the size of the DFT and IDFT operations used in the filter bank. If the size  $D$  is selected as a power of two, the efficient Fast Fourier Transform (FFT) and the inverse Fast Fourier Transform (IFFT) can be used instead of the DFT and IDFT computations. An efficient filter bank implementation uses the polyphase structure with down sampling of subband data together with the use of the efficient FFT and IFFT algorithms.

## 6 Summary

This thesis deals with applied methods for blind speech enhancement. Multi-microphone technologies have been developed together with a single-channel method. All developments have been verified in hardware and realtime experiments. A brief listing of the main contributions for each part of the thesis follows.

### 6.1 Main Contributions

#### **Part I - Complex-valued Independent Component Analysis for Online Blind Speech Extraction**

Part I of the thesis regards a new criterion for complex-valued ICA based on the maximum-Kurtosis approach [6]. The new method is put in relation to the established FastICA method (using the Kurtosis measure) which is an ICA method with high convergence speed (see, e.g., [61, 62]). A fundamental requirement of ICA is that the independent sources are non-Gaussian. Even though this is a requirement for ICA, it cannot always be guaranteed in some real applications. For instance, in a hands-free speech scenario, the speaker may be silent for a long period of time and only interfering noise or disturbances are active. In the case that the interfering noise source signals are Gaussian, the ICA requirement fails. It has previously been shown that the FastICA method diverges for Gaussian-only source signal mixtures (see, e.g., [61]). The proposed method is shown, in Part I, to share important statistical properties with the FastICA method while it circumvents the divergence drawback of FastICA. Further empirical analysis shows that the proposed method is preferable over the popular FastICA in an online configuration.

## **Part II - Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction**

Another method for Kurtosis-based complex-valued ICA that is based on a local quadratic approximation of the Kurtosis measure has been derived during the scope of this thesis. The method in Part II uses another approximation approach than the method in Part I. Specifically, in Part I, the approximation involves a simplification of the Hessian matrix inside the optimization routine. In Part II, the approximation regards a reformulation of the Kurtosis measure into an approximative quadratic form whose exact solution is used in the adaptive coefficient update. The method in Part II has a better extraction performance, in general, but the introduced approximation can make the Hessian matrix close to singular which may be of a problem in real applications.

Part II regards a theoretical analysis of the local quadratic Kurtosis criterion [73]. It is shown that the quadratic approximation technique possesses the desirable fixed-point property provided that the extracted source  $s_i^{[k]}[n]$  has a Kurtosis value that satisfies  $\kappa \{s_i^{[k]}[n]\} > -1$ . This is in many cases a relaxed criterion in relation to the commonly used FastICA method which requires that  $\kappa \{s_i^{[k]}[n]\} \neq 0$ . The fixed-point property implies that an optimization method stays at an optimal solution once it is reached. In addition, due to the fixed-point property, the approximation error in the proposed quadratic technique is zero at an optimal point. The method in Part II plays a central role also in Parts III to V.

## **Part III - Online Maximization of Subband Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction**

The idea of using a quadratic criterion to perform a local approximation of the Kurtosis measure is introduced in Part III [68]. The local approximation is inspired by the work in [74] where a similar approximation technique was used to derive a projection approximation subspace tracking technique. The underlying idea is to replace some parts of the

output signal  $y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n]$  in the Kurtosis optimization criterion (60) with an *a priori* output signal  $\tilde{y}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}^{[k]}[n]$ , where the previous weights  $\mathbf{w}^{[k]}[n-1]$  are seen as constants. At the end, the approximation of the Kurtosis measure renders an expression that is quadratic with respect to the current beamformer filter weights  $\mathbf{w}^{[k]}[n]$ . There are many ways to optimize a quadratic expression and the proposed quadratic approach is therefore directly accessible to most of them. The method is derived, and empirical results from a Digital Signal Processor (DSP) implementation show the high performance of the method.

#### **Part IV - An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Robust Realtime Blind Speech Extraction**

So far, the blind beamforming has been carried out based on an ICA model without any disturbing noise. Disturbing noise generally has a negative influence on the extraction performance of an ICA method. In Part IV, the ICA model includes disturbing noise and a novel solution to reduce the noise is presented. A single-channel speech enhancer is incorporated into the control loop of a blind beamformer [70]. The temporal processing provided by the speech enhancer supports the spatial processing of the beamformer, and vice versa - a successful symbiosis. The combined solution provides a high degree of spatiotemporal speech enhancement in adverse environments. A DSP implementation of the proposed method is presented in this part where empirical results highlight the beneficial behavior of the method.

#### **Part V - Online Blind Speech Extraction Based on a Local Quadratic Kurtosis Criterion and a Preprocessing Automatic Gain Controller**

The Kurtosis-based beamforming criterion in (60) varies with respect to the input signal magnitude. This can easily be verified by consid-

ering the Kurtosis measure while  $\kappa \{ \alpha y^{[k]}[n] \} = |\alpha|^4 \kappa \{ y^{[k]}[n] \}$  for an arbitrary factor  $\alpha$ . This variability of the Kurtosis measure towards input signal magnitudes has a potentially negative influence on the performance of the signal extraction method. A practical solution to this problem is presented in Part V by including an Automatic Gain Control (AGC) unit in the control loop of a blind beamformer [69]. It is shown that the beamformer with an integrated AGC unit is invariant with respect to different input signal magnitudes. A DSP implementation is provided together with empirical results that support the use of this proposed approach.

## **Part VI - Implementation Aspects of the Adaptive Gain Equalizer**

Previously, the work has been focusing on blind beamforming as a means to perform the speech enhancement. Part VI is dedicated to a single-channel speech enhancer denoted as the Adaptive Gain Equalizer (AGE) [29, 30, 31, 32]. Part VI focuses in particular on various implementations aspects of the AGE speech enhancer. It is shown that the AGE can be successfully and equivalently implemented using analogue electrical components (e.g., resistors, capacitors, diodes, operational amplifiers, and transistors) as well as by using a digital processor. A hybrid solution that uses a mixture of analogue and digital electronics shows promising results. An analogue filter bank is used to realize the speech enhancement effect, and a DSP steers the subband gain of the analogue filter bank. The sound path from input to output is entirely analogue, and the processed sound is therefore not constrained by sampling or quantization. In addition, the digital processing allows for other advanced speech processing to be performed. The AGE method that is presented in Part VI is also used as an integral component of Part IV in order to construct a spatiotemporal speech enhancer.

## 6.2 Future Research

This section suggests some directions for future research that are intimately related to the methods proposed in the thesis:

**Spectral Modulation Filtering** An interesting and alternative filtering approach used to perform the signal extraction is encompassed by spectral modulation filtering [75, 76]. Spectral modulation filtering can be seen as a special case of time domain processing. Spectral modulation models the speech as comprising a carrier signal that is modulated by a modulation signal. The idea behind spectral modulation filtering is that the modulation that is assumed to take place in a speech signal is distinguished from other modulations in other signals. Hence, this provides a unique domain where a speech signal can be extracted from a mixture of observed signals. The idea has already been proven for source separation of music sources [77], and it is suggested that it be combined with the BSE field.

**BSS** While BSE focuses on the extraction of one source signal, Blind Signal Separation (BSS) has as its task to extract a number of mixed signals [8, 9, 16, 17, 55, 56]. BSS can be shown as advantageous in applications where it is beneficial to extract several independent sources, for instance, in a meeting room. BSS can be performed by several BSE-methods run either in parallel using a symmetric orthogonalization of the filter weights or run sequentially using a deflation procedure to make the extraction filters orthogonal to each other [48, 78]. It is possible to extend the BSE-methods that are proposed in this thesis to a BSS configuration by utilizing an orthogonalization procedure. However, care has to be taken in the BSS applications as BSS renders that the extracted source signals may be ordered differently in different subbands; this is denoted as the permutation problem. The solution to the permutation problem is denoted as permutation alignment [79].

**Speech de-reverberation** The propagation of an acoustic source signal in an enclosure, e.g., a room, is modeled according to (2), and a microphone receives in this case a mixture of delayed and possibly attenuated components of the same source signal due to a multi-path propagation. In acoustics, the multi-path propagation is called *reverberation*, and if the reverberation is too significant it has a negative effect on the intelligibility and clearness of speech. The process of counteracting reverberation via so-called *de-reverberation* is important, and it has attracted some research during the last decades. The observed reverberant speech will have a distribution closer to the Gaussian distribution in comparison to the original clean source signal. Or, in terms of the Kurtosis measure, the observed reverberant speech will have a Kurtosis value that lies between that of the original clean signal and the Kurtosis value of a Gaussian signal (zero). Hence, one approach to counteract the reverberation is through the maximization of the Kurtosis measure. While this has been performed earlier (see, e.g., [57]), it would be interesting to elaborate on the de-reverberation performance based on the methods proposed in this thesis as the methods are shown to be suitable for a variety of real applications.

## References

- [1] J. Lim. Speech enhancement. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 11:3135–3142, April 1986.
- [2] C. Kyriakakis, P. Tsakalides, and T. Holman. Surrounded by sound. *IEEE Signal Processing Magazine*, 16(1):55–66, January 1999.
- [3] J. Ortega-Garcia and J. Gonzalez-Rodriguez. Overview of speech enhancement techniques for automatic speaker recognition. *Fourth International Conference on Spoken Language*, 2(2):926–932, October 1996.
- [4] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proceedings of the IEEE*, 80(10):1526–1555, October 1992.
- [5] Y. Huang, J. Benesty, and J. Chen. Analysis and comparison of multichannel noise reduction methods in a common framework. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):957–968, July 2008.
- [6] B. Sällberg, N. Grbić, and I. Claesson. Complex-valued independent component analysis for online blind speech extraction. *IEEE Transactions on Speech and Audio Processing*, 16(8):1624–1632, November 2008.
- [7] Z. Yermeche, N. Grbić, and I. Claesson. Blind subband beamforming with time-delay constraints for moving source speech enhancement. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2360–2372, November 2007.
- [8] R. Mukai, H. Sawada, S. Araki, and S. Makino. Blind source separation of many signals in the frequency domain. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:969–972, May 2006.

- [9] R. Mukai, H. Sawada, S. Araki, and S. Makino. Robust real-time blind source separation for moving speakers in a room. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:469–472, May 2003.
- [10] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of a dominant source signal from mixtures of many sources. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 3:61–64, March 2005.
- [11] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [12] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [13] R. E. Crochiere and L. R. Rabiner. *Multirate Digital Signal Processing*. Prentice Hall, 1983.
- [14] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [15] J. G. Proakis. *Digital Signal Processing*. Pearson Higher Education, 1995.
- [16] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [17] N. Gribić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.
- [18] M. Vetterli and D. le Gall. Perfect reconstruction fir filter banks: some properties and factorizations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(7):1057–1071, July 1989.

- [19] J. M. de Hahn, N. Grbić, I. Claesson, and S. Nordholm. Filter bank design for subband adaptive microphone arrays. *IEEE Transactions on Speech and Audio Processing*, 11(1):14–23, January 2003.
- [20] F. Baumgarte. A computationally efficient cochlear filter bank for perceptual audio coding. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:3265–3268, May 2001.
- [21] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, April 1979.
- [22] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121, December 1984.
- [23] O. Cappé. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349, April 1994.
- [24] H. Gustafsson, S. Nordholm, and I. Claesson. Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Transactions on Speech and Audio Processing*, 9(8):799–807, November 2001.
- [25] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer. *IEEE International Symposium on Digital Signal Processing and Communications Systems*, September 2003.
- [26] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings. *IEEE Vehicular Technology Conference*, 5:3718–3722, September 2004.

- [27] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement for personal communication using an adaptive gain equalizer. *Elsevier Signal Processing*, 85(6):1089–1101, 2005.
- [28] N. Westerlund, M. Dahl, and I. Claesson. Real-time implementation of an adaptive gain equalizer for speech enhancement purposes. *WSEAS International Conference on Electronics, Control and Signal Processing*, September 2003.
- [29] B. Sällberg, H. Åkesson, N. Westerlund, M. Dahl, and I. Claesson. Analog circuit implementation for speech enhancement purposes. *IEEE 38th Asilomar Conference on Circuits, Systems and Computers*, 2:2285–2289, November 2004.
- [30] B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson. A mixed analog - digital hybrid for speech enhancement purposes. *IEEE International Symposium on Circuits and Systems*, 2:852–855, May 2005.
- [31] B. Sällberg and M. Dahl. Speech enhancement implementations in the digital, analog, and hybrid domain. *IEEE Swedish System on Chip Conference*, April 2005.
- [32] B. Sällberg, N. Grbić, and I. Claesson. Implementation aspects of the adaptive gain equalizer. Blekinge tekniska högskola, research report, issn: 1103-1581, May 2006.
- [33] M. Shroeder. Models of hearing. *Proceedings of the IEEE*, 63(9):1332–1350, September 1975.
- [34] M. R. Flax and J. S. Jin. Hybrid auditory masking models. *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 1–4, May 2001.
- [35] B. C. J. Moore. *An introduction to the psychology of hearing*. Academic press, fourth edition, 1997.

- 
- [36] J. Ramírez, J. Segura, C. Benítez, Á. de la Torre, and A. Rubio. A new kullback-leibler vad for speech recognition in noise. *IEEE Signal Processing Letters*, 11(2):266–269, February 2004.
- [37] H. Gustafsson, U. Lindgren, I. Claesson, and S. Nordholm. System and method for dual microphone signal noise reduction using spectral subtraction. US Patent 6717991, April 6 2004.
- [38] S. Nordholm, I. Claesson, and N. Grbić. *Signal Processing Techniques and Applications in Microphone Arrays*, chapter Optimal and Adaptive Microphone Arrays for Speech Input in Automobiles, pages 307–330. Springer Verlag, 2001.
- [39] N. Grbić, S. Nordholm, and A. Cantoni. Optimal fir subband beamforming for speech enhancement in multipath environments. *IEEE Signal Processing Letters*, 10(11):335–338, November 2003.
- [40] J. Chen, J. Benesty, Y. Huang, and S. Doclo. New insights into the noise reduction wiener filter. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1218–1234, July 2006.
- [41] A. Spriet, S. Doclo, M. Moonen, and J. Wouters. A unification of adaptive multi-microphone noise reduction systems. *International Workshop on Acoustic Echo and Noise Control*, pages 1–4, September 2006.
- [42] N. Grbić and S. Nordholm. Soft constrained subband beamforming for hands-free speech enhancement. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 1:885–888, 2002.
- [43] S. Doclo and M. Moonen. Gsvd-based optimal filtering for single and multi-microphone speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 50(9):2230–2244, September 2002.

- [44] B. Widrow, P. E. Mantey, L. J. Griffiths, and B. B. Goode. Adaptive antenna systems. *Proceedings of the IEEE*, 55(12):2143–2159, December 1967.
- [45] Z. Yermeche. *Soft-Constrained Subband Beamforming for Speech Enhancement*. PhD thesis, Blekinge Tekniska Högskola, 2007:14, November, ISBN: 978-91-7295-121-1.
- [46] Z. Yermeche, B. Sällberg, N. Grbić, and I. Claesson. Real-time dsp implementation of a subband beamforming algorithm for dual microphone speech enhancement. *IEEE International Symposium on Circuits and Systems*, pages 353–356, May 2007.
- [47] S. Werner, J. Apolinario, and M. de Campos. On the equivalence of rls implementations of lcmv and gsc processors. *IEEE Signal Processing Letters*, 10(12):356–359, December 2003.
- [48] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [49] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [50] J. F. Cardoso. Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10):2009–2025, October 1998.
- [51] F. Wang, L. Hongwei, and L. Rui. Data mining with independent component analysis. *The 6th World Congress on Intelligent Control and Automation*, 2:6043–6047, June 2006.
- [52] M. Potter and W. Kinsner. Competing ica techniques in biomedical signal analysis. *Canadian Conference on Intelligent Sensing and Information Processing*, 2:987–992, May 2001.
- [53] S. Haykin. *Neural Networks: a comprehensive foundation*. Prentice Hall, 1999.

- [54] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.
- [55] J. F. Cardoso. Source separation using higher order moments. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 4:2109–2112, May 1989.
- [56] J. P. LeBlanc and P. L. de Léon. Speech separation by kurtosis maximization. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2:1029–1032, May 1998.
- [57] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 6:3701–3704, May 2001.
- [58] Z. Ding. A new algorithm for automatic beamforming. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2:689–693, November 1991.
- [59] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [60] W. Zhang and S. Gazor. Statistical modelling of speech signals. *IEEE International Conference on Signal Processing*, 1:480–483, August 2002.
- [61] E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, February 2000.
- [62] S. C. Douglas. Fixed-point fastica algorithms for the blind separation of complex-valued signal mixtures. *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1320–1325, October 2005.

- [63] H. Li and T. Adali. A class of complex ica algorithms based on the kurtosis cost function. *IEEE Transactions on Neural Networks*, 19(3):408–420, March 2008.
- [64] D. H. Brandwood. A complex gradient operator and its application in adaptive array theory. *Proceedings of the IEE*, 130(1):11–16, February 1983.
- [65] A. van den Bos. Complex gradient and hessian. *IEE Proceedings Vision, Image and Signal Processing*, 141(6):380–382, December 1994.
- [66] E. Oja and Z. Yuan. The fastica algorithm revisited: Convergence analysis. *IEEE Transactions on Neural Networks*, 17(6):1370–1381, November 2006.
- [67] B. Sällberg, M. Swartling, N. Grbić, and I. Claesson. Real-time implementation of a blind beamformer for subband speech enhancement using kurtosis maximization. *International Workshop on Acoustics, Echo and Noise Control*, pages 485–489, September 2006.
- [68] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of subband kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE 15th International Conference on Digital Signal Processing*, pages 603–606, July 2007.
- [69] B. Sällberg, N. Grbić, and I. Claesson. Online blind speech extraction based on a locally quadratic kurtosis criteria and a preprocessing automatic gain controller. *IEEE 49th International Symposium ELMAR*, pages 139–142, September 2007.
- [70] B. Sällberg, N. Grbić, and I. Claesson. An adaptive blind beamformer with an integrated single-channel noise reduction method for robust realtime blind speech extraction. *IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 309–312, March 2008.

- [71] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann. Post-processing for convolutive blind source separation. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:37–40, May 2006.
- [72] S. Y. Low, S. Nordholm, and R. Togneri. Convolutive blind signal separation with post-processing. *IEEE Transactions on Speech and Audio Processing*, 12(5):539–548, September 2004.
- [73] B. Sällberg, N. Grbić, and I. Claesson. Statistical analysis of a local quadratic criterion for blind speech extraction. *IEEE Signal Processing Letters*, accepted for publication November 2008.
- [74] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, January 1995.
- [75] L. Qin and L. Atlas. Coherent modulation filtering for speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4481–4484, March 2008.
- [76] L. Qin and L. Atlas. Properties for modulation spectral filtering. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4:521–524, March 2005.
- [77] L. Atlas and C. Janssen. Coherent modulation spectral filtering for single-channel music source separation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4:461–464, March 2005.
- [78] A. Cichocki, R. Thawonmas, and S. Amari. Sequential blind signal extraction in order specified by stochastic properties. *IEEE Electronic Letters*, 33(1):64–65, January 1997.
- [79] H. Sawada, S. Araki, R. Mukai, and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation.

*IEEE Transactions on Audio, Speech, and Language Processing*,  
15(5):1592–1604, July 2007.





PART I

**Complex-valued Independent Component Analysis for Online Blind Speech Extraction**

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson. Complex-valued Independent Component Analysis for Online Blind Speech Extraction. *IEEE Transactions on Audio Speech and Language Processing*, 16(8):1624-1632, November 2008.

© 2008 IEEE. Reprinted, with permission, from IEEE Transactions on Audio Speech and Language Processing.

**Modification to the original paper:**

The notations have been standardized so as to fit the other parts of this thesis.

# Complex-valued Independent Component Analysis for Online Blind Speech Extraction

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

## Abstract

This paper presents a theoretical analysis of a certain criterion for complex-valued Independent Component Analysis (ICA) with a focus on Blind Speech Extraction (BSE) of a spatiotemporally non-stationary speech source. In the paper, the proposed criteria denoted KSICA is related to the well-known FastICA method with the Kurtosis contrast function. The proposed method is shown to share the important fixed-point feature with the FastICA method, although an improvement with the proposed method is that it does not exhibit the divergent behavior for a mixture of Gaussian-only sources that the FastICA method tends to do, and it shows better performance in online implementations. Compared to the FastICA, the KSICA method provides a 10dB higher source extraction performance and a 10 dB lower standard deviation in a data batch approach when the data batch size is less than 100 samples. For larger batch sizes, the KSICA method performs equally well. In an online application with spatially stationary sources the KSICA method provides around 10 dB higher interference suppression, and 1 MOS-unit lower speech distortion compared to the FastICA for 0.15 s time constant in the algorithm update parameter. Thus, the FastICA performance matches the KSICA performance for a time constant above 1 s. Finally, in an online application with a moving speech source, the KSICA method provides 10 dB higher interference suppression, compared to the FastICA for the same algorithm settings. All in all, the proposed KSICA method is shown to be a viable alternative for online BSE of complex-valued signal mixtures.

## 1 Introduction

Blind extraction of signals is a reoccurring problem in a variety of signal processing applications including speech enhancement, extraction of biomedical signals, etc. The notation “blind” implies that the signal extraction is relying only on certain assumptions regarding the statistical properties of the source signals, such as an assumption regarding the independence of the source signals. Some of these blind approaches can be sorted under the field of Independent Component Analysis (ICA) [1, 2]. This paper focuses on a new method for performing Blind Speech Extraction (BSE) using an ICA approach. This approach is explored through a real-time speech enhancement application. It should be noted that in BSE, it is desirable to extract a dominant speech source (or a group of dominant sources) from an observed mixture of many sources [3, 4, 5, 6, 7].

To extract speech, a convolution model is adopted where a set of source signals are emitted in a room and received by an array of microphones. Such a convolution model in the time domain corresponds to a multiplicative model in the frequency domain [8, 9]. It is in many cases desirable to perform the ICA in the frequency (subband) domain [10, 11, 12], which in general yields a faster convergence rate and a lower computational load as opposed to a corresponding full-band time domain method. The BSE discussed in this paper was furthermore performed on complex-valued data generated with the help of a specific Fourier-transform-based time-frequency subband transformation. One popular method used to perform BSE is the FastICA method [2, 13]. The FastICA method has been reported to be a fast and efficient method for blind extraction of signals. Bingham et al. [13] derived the FastICA method for complex-valued signal compositions with a focus on sources having circular distributions in order to simplify their derivations. In addition to this, Douglas [14] presented an alternative version of the FastICA method for separation of complex-valued signal mixtures, without relying upon circularity assumptions of the source signals.

Essentially, the FastICA method is intended for a batch processing approach where signal statistics are estimated for a certain period of time, described as the data batch duration, whereafter the ICA problem is solved in an iterative manner until a pre-specified stopping criterion is met. Mukai et al. [11] show that a batch-based ICA method, based on a natural gradient approach, achieves better performance for fixed sources than an online method where the continuously received data is used to update statistical estimates. However, in an applied BSE application the risk is that the sources are spatiotemporally non-stationary, e.g., the spatial activity pattern of speech sources is typically non-stationary as the speakers could move around, and pauses and bursts in the speech make a speech signal temporally non-stationary. Because of this, the batch processing approach is unsuitable in such non-stationary environments due to its inherent estimation delay. Quite simply, this delay limits the method's ability to track sources in a non-stationary environment. For this reason, the use of batch-based ICA methods (e.g., FastICA) cannot be recommended in such non-stationary environments. To address this problem, this paper focuses on a new ICA method which is related to the FastICA method. However, unlike FastICA, the new method is intended for performing online estimations of the source statistics used during BSE.

The use of ICA methods described in this paper are based on a fourth order cumulant [15], i.e., the Kurtosis measure. The use of cumulants (higher order statistics) in order to find approximative and simple features distinguishing desired sources from undesired sources dates back to the early pioneering work of ICA (see for instance [16, 17, 18, 19] and references therein). However, the use of higher order cumulants for performing Kurtosis maximization has met with critique because of its sensitivity to data outliers. When an outlying data sample enters the Kurtosis algorithm it may result in divergence in the algorithm. One practical solution to this problem is to make use of some sort of signal-conditioning before the data is sampled by a device, e.g., by using a compressor before the sampler. The rationale for using such measures as the Kurtosis, despite their sensitivity to outliers, is that

they will yield a polynomial structure in their adaptive weight update equations, and this is identified in this communication as an important feature that may generate a feasible real-time Digital Signal Processor (DSP)-based implementation.

In addition to the pure ICA task, many approaches also incorporate temporal (and/or spatial) post-processors in order to improve performance further [20, 21]. The blind extraction of source signals discussed in this paper was performed purely without any post-processors. It is noted, however, that additional post-processors may be added in the future to improve the performance of the proposed method.

The outline of this paper is as follows: the adopted ICA data model is presented in Section 2. The FastICA method for complex-valued data is briefly repeated, from [13], in Section 3. The proposed new ICA method is presented in Section 4. A brief discussion on batch processing for ICA is given in Section 5, and applied ICA is discussed in Section 6. The FastICA and the proposed method are evaluated and compared in Section 7. A summary with conclusions is given in Section 8.

## 2 The ICA Data Model

The model assumes an array of  $M$  microphones where each received real-valued time signal is denoted  $x_m[t]$  for  $m : m \in \mathbb{N}, m < M$  and where  $t$  denotes continuous time. Each received time signal is sampled and decomposed into a time-frequency representation using a filter bank with  $K$  subbands, and where each subband signal is denoted  $x_m^{[k]}[n]$  with  $k : k \in \mathbb{N}, k < K$  and  $n$  is a sample index in the subband domain. The subband decomposition is carried out in the evaluation part by using a Discrete Fourier Transform (DFT) modulated uniform analysis filter bank and an efficient polyphase realization (see for instance [22] for details regarding the filter bank). Since the analysis in the next section is general and identical for all subbands, the notation will intentionally omit the subband index  $k$ . For the sake of

simplicity, the focus in the presentation is only on one of the  $K$  subbands. The received subband signals are represented using a signal vector  $\mathbf{x}^{[k]}[n] = \left(x_0^{[k]}[n], x_1^{[k]}[n], \dots, x_{M-1}^{[k]}[n]\right)^T$  of size  $M \times 1$ , where  $(\cdot)^T$  denotes the vector transpose. In this paper, the commonly used noise-free ICA mixing model for complex-valued data is adopted as

$$\mathbf{x}^{[k]}[n] = \mathbf{A}^{[k]}[n]\mathbf{s}^{[k]}[n]. \quad (1)$$

Here  $\mathbf{A}^{[k]}[n]$  is a time-varying source mixing matrix of size  $M \times I$ , while the  $I$  original, independent sources  $\mathbf{s}^{[k]}[n] = \left(s_0^{[k]}[n], s_1^{[k]}[n], \dots, s_{I-1}^{[k]}[n]\right)^T$  are assumed to obey  $\mathbb{E} \left\{ \mathbf{s}^{[k]}[n]\mathbf{s}^{[k]}[n]^H \right\} = \mathbf{I}_I$ . In this formula,  $(\cdot)^H$  denotes the complex conjugate transpose while  $\mathbb{E} \{ \cdot \}$  represents the expectation operator, and  $\mathbf{I}_I$  is an identity matrix of the size  $I \times I$ . According to [13] and [15], the Kurtosis value of a complex-valued signal  $x^{[k]}[n]$  that has a circular distribution can be defined as

$$\kappa \left\{ x^{[k]}[n] \right\} = \mathbb{E} \left\{ \left| x^{[k]}[n] \right|^4 \right\} - 2\mathbb{E} \left\{ \left| x^{[k]}[n] \right|^2 \right\}^2. \quad (2)$$

The Kurtosis value of each original source signal is, due to the assumption  $\mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\} = 1$ , equal to  $\kappa \left\{ s_i^{[k]}[n] \right\} = \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^4 \right\} - 2$ , for  $i : i \in \mathbb{N}, i < I$ , and it is assumed that the sources are ordered so that the dominant source with the highest absolute Kurtosis value, is  $s_0^{[k]}[n]$ , i.e.,  $\left| \kappa \left\{ s_0^{[k]}[n] \right\} \right| > \left| \kappa \left\{ s_i^{[k]}[n] \right\} \right|$  for  $i \in \{1, 2, \dots, I-1\}$ . This assumption implies that a blind extraction method would extract the dominant source  $s_0^{[k]}[n]$ . The subband output signal  $y^{[k]}[n]$  is a linear combination of the observed input signals  $\mathbf{x}^{[k]}[n]$  weighted by the filter vector  $\mathbf{w}^{[k]}[n]$  of size  $M \times 1$  according to

$$y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{A}^{[k]}[n]\mathbf{s}^{[k]}[n]. \quad (3)$$

To continue, the adopted signal model uses one filter vector tap per subband. This model captures signal dynamics up to the frame length

used in the filter bank. While it is possible to capture longer time scales by using several filter-taps per subband, i.e., FIR filtering, this method is not considered in this paper since the theoretical analysis is greatly simplified if using only one filter-tap per subband. It is desirable that the BSE method finds a  $\mathbf{w}^{[k]}[n]$  so that  $y^{[k]}[n] = s_0^{[k]}[n]$ , i.e.,  $\mathbf{w}^{[k]}[n]^H \mathbf{A}^{[k]}[n] = (1, 0, \dots, 0)$ . The time-domain output signal  $y[t]$  is then computed from the subband output signals  $y^{[k]}[n]$  by a DFT modulated synthesis filter bank matched to the analysis filter bank [22].

### 3 The FastICA Method Revisited

The FastICA [13] contrast function used here is the Kurtosis contrast function  $G(u) = \frac{1}{2}u^2$ , i.e.,

$$J_G(\mathbf{w}^{[k]}[n]) = \mathbb{E} \left\{ G \left( \left| \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n] \right|^2 \right) \right\} = \frac{1}{2} \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\}. \quad (4)$$

The rationale of focusing on the Kurtosis contrast function is that it will yield a polynomial structure in its update equation. A polynomial structure can be realized in hardware using a predetermined series of multiplication operations. This is often beneficial in a real-time implementation where a low complexity is preferable. According to [13], the optima of  $J_G(\mathbf{w}^{[k]}[n])$  conditioned on  $\mathbb{E} \left\{ \left| \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n] \right|^2 \right\} = 1$  follows the Kuhn-Tucker conditions, where the cost function  $J_C(\mathbf{w}^{[k]}[n]) = J_G(\mathbf{w}^{[k]}[n]) - \beta \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}$  is being used, and  $\beta$  is a real-valued parameter.

The filter vector update equation at a point  $\mathbf{w}^{[k]}[n-1]$  according to an approximative Newton's method [13, 23], with an additional unity

norm constraint applied, is

$$\mathbf{w}_+^{[k]} = -\mathbf{w}^{[k]}[n-1] + \frac{\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}^{-1} \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^* |\tilde{y}^{[k]}[n]|^2 \mathbf{x}^{[k]}[n] \right\}}{2\mathbb{E} \left\{ |\tilde{y}^{[k]}[n]|^2 \right\}}, \quad (5)$$

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\left\| \mathbf{w}_+^{[k]} \right\|_2}, \quad (6)$$

where  $\mathbf{w}_+^{[k]}$  is a temporary variable, and  $\tilde{y}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}^{[k]}[n]$ . The normalization approach in (6) is used to avoid the trivial solution  $\mathbf{w}^{[k]}[n] = \mathbf{0}_{M \times 1}$ , where  $\mathbf{0}_{M \times 1}$  denotes a null-vector of size  $M \times 1$ . This normalization approach will preserve the power of the source signal [2]. The normalization is henceforth assumed to be performed after each update of the temporary variable  $\mathbf{w}_+^{[k]}$ .

### 3.1 Fixed-Point Behavior of the FastICA Method

It is customary to perform a pre-processing whitening of the input data using, for instance, Principal Component Analysis (PCA) (e.g., [1]) in order to speed up the convergence of the FastICA method. The PCA decorrelation has the same impact as if the mixing matrix  $\mathbf{A}^{[k]}[n]$  would possess a Unitary property, i.e.,  $\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} = \mathbf{A}^{[k]}[n] \mathbb{E} \left\{ \mathbf{s}^{[k]}[n] \mathbf{s}^{[k]}[n]^H \right\} \mathbf{A}^{[k]}[n]^H = \mathbf{A}^{[k]}[n] \mathbf{A}^{[k]}[n]^H = \mathbf{I}_M$ .

The output signal of the FastICA method is denoted  $y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{A}^{[k]}[n] \mathbf{s}^{[k]}[n]$ , according to (3). It is convenient to define a vector  $\mathbf{q}^{[k]}[n]$  as  $\mathbf{q}^{[k]}[n] = \mathbf{A}^{[k]}[n]^H \mathbf{w}^{[k]}[n]$  (or  $\mathbf{w}^{[k]}[n] = \mathbf{A}^{[k]}[n] \mathbf{q}^{[k]}[n]$ ), which yields that  $y^{[k]}[n] = \mathbf{q}^{[k]}[n]^H \mathbf{s}^{[k]}[n]$ . The unity norm constraint yields that  $\left\| \mathbf{w}^{[k]}[n] \right\|_2 = 1$  and  $\left\| \mathbf{q}^{[k]}[n] \right\|_2 = 1$ .

To prove the fixed-point behavior of FastICA, let the filter vector at the previous iteration equal the optimal solution that extracts the

dominant source  $s_0^{[k]}[n]$ :

$$\mathbf{q}^{[k]}[n-1] = \mathbf{q}_{\text{opt}}^{[k]} = (1, 0, \dots, 0)^T. \quad (7)$$

In other words, the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]H} = \mathbf{q}_{\text{opt}}^{[k]H} \mathbf{A}^{[k]}[n]^H$  corresponds to the first row of the inverse (or if  $I \neq M$ , the Moore-Penrose pseudo inverse)  $\mathbf{A}^{[k]}[n]^{-1} = \mathbf{A}^{[k]}[n]^H$  of the matrix  $\mathbf{A}^{[k]}[n]$ . The definition of  $\mathbf{q}_{\text{opt}}^{[k]}$  in (7) yields that

$$\tilde{y}^{[k]}[n] = \mathbf{q}^{[k]}[n-1]^H \mathbf{s}^{[k]}[n] = \mathbf{q}_{\text{opt}}^{[k]H} \mathbf{s}^{[k]}[n] = s_0^{[k]}[n]. \quad (8)$$

In this way, the behavior of the FastICA method at an optimal solution is

$$\mathbf{q}_+^{[k]} = \frac{1}{2} \kappa \left\{ s_0^{[k]}[n] \right\} \mathbf{q}_{\text{opt}}^{[k]}. \quad (9)$$

If the dominant source,  $s_0^{[k]}[n]$ , possesses a non-zero Kurtosis value, i.e.,  $\kappa \left\{ s_0^{[k]}[n] \right\} \neq 0$ , then the updated filter vector provided by FastICA,  $\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\|\mathbf{w}_+^{[k]}\|_2} = \pm \mathbf{w}_{\text{opt}}^{[k]}$ , is a stable fixed-point optimal solution, and the sign  $\pm$  is determined by the sign of the Kurtosis of the dominant source. The FastICA method is, by virtue of this property, denoted a fixed-point method. However, the FastICA is undefined for Gaussian-only mixtures, since  $\kappa \left\{ s_0^{[k]}[n] \right\} = 0$  yields a division-by-zero in the filter vector normalization stage in (6). These conclusions are already established properties of the FastICA method (see for instance [2, 13]).

### 3.2 Local Consistency of the FastICA Method

This paper follows the analysis of the local consistency of FastICA conducted in [13]. This analysis was conducted around the optimal point  $\mathbf{w}_{\text{opt}}^{[k]}$  which extracts the dominant source. The analysis is conducted by evaluating a second-order Taylor expansion (any term with

an order higher than two is omitted) of  $J_G(\mathbf{w}^{[k]}[n-1])$  at the point  $\mathbf{w}^{[k]}[n-1] = \mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}$ , where  $\mathbf{p}^{[k]}$  is a small perturbation vector to the optimal solution (here, the term “small” implies that  $\|\mathbf{p}^{[k]}\|_2 \ll 1$ ). The unity norm constraint (6) yields that the perturbation vector must satisfy  $\|\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}\|_2^2 = \|\mathbf{w}_{\text{opt}}^{[k]}\|_2^2 = 1$ , and thus the perturbed optimal solution is evaluated at a hyper-sphere. The second-order Taylor series expansion of  $J_G(\mathbf{w}^{[k]}[n-1])$  around the point  $\mathbf{w}_{\text{opt}}^{[k]}$  is

$$J_G(\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}) = J_G(\mathbf{w}_{\text{opt}}^{[k]}) + 2\text{E} \left\{ |s_0^{[k]}[n]|^4 \right\} \text{Re} \left\{ \mathbf{p}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]} \right\} + 2 \|\mathbf{p}^{[k]}\|_2^2. \quad (10)$$

As long as the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  as well as the perturbed optimal solution  $\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}$  obey the unity norm constraint, the following relationship holds (from [13]):

$$2\text{Re} \left\{ \mathbf{p}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]} \right\} = - \|\mathbf{p}^{[k]}\|_2^2. \quad (11)$$

This relationship yields

$$J_G(\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}) = J_G(\mathbf{w}_{\text{opt}}^{[k]}) - \kappa \left\{ s_0^{[k]}[n] \right\} \|\mathbf{p}^{[k]}\|_2^2. \quad (12)$$

The term  $\|\mathbf{p}^{[k]}\|_2^2$  is always greater than, or equal to, 0, and the type (local maximum or minimum) of the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  is therefore dependent on the sign of  $\kappa \left\{ s_0^{[k]}[n] \right\}$ , i.e., the sign of the Kurtosis of the dominant source. If  $\kappa \left\{ s_0^{[k]}[n] \right\} > 0$ , the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  is a local maximum. The optimum is a local minimum if  $\kappa \left\{ s_0^{[k]}[n] \right\} < 0$ , and it is a saddle point if  $\kappa \left\{ s_0^{[k]}[n] \right\} = 0$ , i.e., if  $s_0^{[k]}[n]$  is Gaussian distributed. This result is identical to that in [13].

### 3.3 Implications of Performing Applied BSE using Fast-ICA

When the FastICA is used for Blind Signal Extraction, it is required that at least the dominant source is non-Gaussian, i.e., has a non-zero Kurtosis value. This has implications in a real-time realization of the FastICA method, where the true source statistics are estimated using sample-based estimators. Such a sample-based estimator should have a finite memory, or at least a rather short integration time, in order to be able to track changes in the environment and to restrain the memory requirement. Hence, the FastICA method risks divergence in a scenario where a spatiotemporally non-stationary source is mixed with one or several stationary Gaussian sources, and the non-Gaussian source becomes inactive for the duration of the sample-based estimator's memory length. This behavior is emphasized in the evaluation, Section 7, where a recursive sample-based estimator is used to estimate source signal statistics. The remedy to this behavior is increasing the memory length of the sample-based estimator. However, in that case the method fails in tracking sources in a rapidly changing environment. In addition to this, the memory requirement in the realization can, in some cases, be increased. These aforementioned issues regarding applied BSE, and more specifically those related to the shortcomings of the FastICA method, are the main reasons for proposing the new ICA method.

## 4 The Proposed ICA Method

The problem that needs to be resolved is that the FastICA method is divergent for Gaussian-only source mixtures. The idea of this paper is to propose an alternative Kurtosis measure that weighs the fourth-order term  $E \left\{ |y^{[k]}[n]|^4 \right\}$  towards the square second-order term  $E \left\{ |y^{[k]}[n]|^2 \right\}^2$  in the Kurtosis measure so as to avoid the divergent behavior. An alternative method for BSE is therefore proposed that

shares important properties with the FastICA method, such as the fixed-point behavior, but which circumvents the divergent behavior of FastICA for a mixture of Gaussian-only sources. A general cost function for a weighted Kurtosis measure that encapsulates both the FastICA method and the proposed method is described by the cost function  $C(\mathbf{w}^{[k]}[n], \alpha)$

$$\begin{aligned} C(\mathbf{w}^{[k]}[n], \alpha) &= \frac{1}{2} \kappa \{y^{[k]}[n]\} + \alpha \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2 = \\ &= \frac{1}{2} \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\} - (1 - \alpha) \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2, \end{aligned} \quad (13)$$

where  $\alpha$  is a real-valued parameter. If, for instance,  $\alpha = 1$ , then  $C(\mathbf{w}^{[k]}[n], 1) = \frac{1}{2} \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\}$  equals the Kurtosis contrast function used in the FastICA method, i.e., (4). Various  $\alpha$ -values yield BSE algorithms with different properties. This paper focuses on a certain  $\alpha$ -value:  $\alpha = -1$ . This value yields a non-divergent BSE algorithm that can be used to construct a new method, henceforth denoted Kurtosis maximization in the Subband domain ICA (KSICA). With the help of this method, the corresponding contrast function can be defined as  $J(\mathbf{w}^{[k]}[n]) = C(\mathbf{w}^{[k]}[n], -1)$

$$J(\mathbf{w}^{[k]}[n]) = \frac{1}{2} \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\} - 2 \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2. \quad (14)$$

A Newton-based method is used here to optimize the contrast function  $J(\mathbf{w}^{[k]}[n])$  of the KSICA method, where the gradient and the Hessian

matrix of  $J(\mathbf{w}^{[k]}[n])$  are evaluated as

$$\begin{aligned} \nabla J(\mathbf{w}^{[k]}[n]) &= \mathbb{E} \left\{ y^{[k]}[n]^* |y^{[k]}[n]|^2 \mathbf{x}^{[k]}[n] \right\} \\ &\quad - 4\mathbb{E} \left\{ |y^{[k]}[n]|^2 \right\} \mathbb{E} \left\{ y^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\}, \end{aligned} \quad (15)$$

$$\begin{aligned} \nabla^2 J(\mathbf{w}^{[k]}[n]) &= 2\mathbb{E} \left\{ |y^{[k]}[n]|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \\ &\quad - 4 \left( \mathbb{E} \left\{ y^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\} \mathbb{E} \left\{ y^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \right. \\ &\quad \left. + \mathbb{E} \left\{ |y^{[k]}[n]|^2 \right\} \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \right). \end{aligned} \quad (16)$$

Following the approximations outlined for the FastICA method, it is assumed that  $\mathbb{E} \left\{ |y^{[k]}[n]|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \approx \mathbb{E} \left\{ |y^{[k]}[n]|^2 \right\}$ .

$\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}$  and that  $\mathbb{E} \left\{ y^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\} \mathbb{E} \left\{ y^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \approx \mathbb{E} \left\{ |y^{[k]}[n]|^2 \right\} \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}$ , which yields

$$\nabla^2 J(\mathbf{w}^{[k]}[n]) \approx -6\mathbb{E} \left\{ |y^{[k]}[n]|^2 \right\} \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}. \quad (17)$$

The filter vector update for KSICA according to Newton's method at a point  $\mathbf{w}^{[k]}[n-1]$ , and under a unity norm constraint, is

$$\begin{aligned} \mathbf{w}_+^{[k]} &= \mathbf{w}^{[k]}[n-1] \\ &\quad + \frac{\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}^{-1} \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^* |\tilde{y}^{[k]}[n]|^2 \mathbf{x}^{[k]}[n] \right\}}{2\mathbb{E} \left\{ |\tilde{y}^{[k]}[n]|^2 \right\}}, \end{aligned} \quad (18)$$

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\left\| \mathbf{w}_+^{[k]} \right\|_2}. \quad (19)$$

The update equation of KSICA in (18) is identical to the update equation of FastICA in (5), except for the point that FastICA has a negative  $-\mathbf{w}^{[k]}[n-1]$  term in its update.

#### 4.1 Fixed-Point Behavior of the KSICA Method

Following the assumptions in Section 3.1, where a pre-processing whitening of the input data is performed so that  $E\{\mathbf{x}^{[k]}[n]\mathbf{x}^{[k]}[n]^H\} = \mathbf{I}_M$ , and the analysis is conducted around the optimal point  $\mathbf{q}_{\text{opt}}^{[k]} = \mathbf{A}^{[k]}[n]\mathbf{w}_{\text{opt}}^{[k]}$ , it may be concluded that the behavior of the KSICA method at an optimal solution follows

$$\mathbf{q}_+^{[k]} = \frac{1}{2} \left( \kappa \left\{ s_0^{[k]}[n] \right\} + 4 \right) \mathbf{q}_{\text{opt}}^{[k]}, \quad (20)$$

$$\mathbf{q}^{[k]}[n] = \frac{\mathbf{q}_+^{[k]}}{\|\mathbf{q}_+^{[k]}\|_2} = \mathbf{q}_{\text{opt}}^{[k]}. \quad (21)$$

Consequently, the KSICA is a fixed-point method like the FastICA method, i.e., once the KSICA has found an optimal solution, it stays at that optimal solution as the iterations proceed. Furthermore, since the term  $\frac{1}{2} \left( \kappa \left\{ s_0^{[k]}[n] \right\} + 4 \right)$  in (20) is always positive and non-zero, the KSICA is always stable and it is not inhibited by any probability distribution assumptions regarding the source signals. This is contrary to the FastICA method, in which at least the dominant source must have a non-zero Kurtosis value in order to avoid divergence. This property of the KSICA makes it tractable in a real-time application where the non-Gaussianity assumption cannot always be guaranteed. It must be stressed that while the KSICA and FastICA share the fundamental assumptions imposed by the theory of ICA, which disallows separation of Gaussian-only sources, the KSICA does not diverge in the case of Gaussian-only sources. It is this difference from FastICA which makes the KSICA a superior candidate for performing BSE in an online setting where constant activity of non-Gaussian sources may not be guaranteed.

## 4.2 Local Consistency of KSICA

An analysis of the local consistency of the KSICA method follows the local consistency-analysis of the FastICA method in Section 3.2. Therefore, a pre-processing whitening stage of the data is also included here. The second-order Taylor expansion of  $J(\mathbf{w}^{[k]}[n])$  around the point  $\mathbf{w}_{\text{opt}}^{[k]}$  is

$$J(\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}) = J(\mathbf{w}_{\text{opt}}^{[k]}) + 2\left(\kappa \left\{s_0^{[k]}[n]\right\} - 2\right) \text{Re} \left\{ \mathbf{p}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]} \right\} - 6 \left\| \mathbf{p}^{[k]} \right\|_2^2. \quad (22)$$

Using the relationship in (11) yields

$$J(\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}) = J(\mathbf{w}_{\text{opt}}^{[k]}) - \left(\kappa \left\{s_0^{[k]}[n]\right\} + 4\right) \left\| \mathbf{p}^{[k]} \right\|_2^2. \quad (23)$$

Consequently, the term  $-\left(\kappa \left\{s_0^{[k]}[n]\right\} + 4\right) \left\| \mathbf{p}^{[k]} \right\|_2^2$  is always less than, or equal to, 0, and the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  is thereby always a local maxima to  $J(\mathbf{w}^{[k]}[n])$ , independent of the distribution of the sources. This result further implies that the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  related to the dominant source  $s_0^{[k]}[n]$  is in fact always a global maximum solution.

## 5 Batch Processing for ICA

When performing BSE using a batch processing approach, input data is collected for a certain amount of time corresponding to the batch duration. The set of recorded data in a batch indexed  $b$  is denoted  $\mathbf{X}_b^{[k]}$  of the size  $N \times M$ , where  $N$  is the number of data samples in the batch and  $M$  is the number of microphones, as before. The statistical measures are then estimated for all data available in the data batch. The BSE method then follows a pre-specified iterative schedule for each data batch  $b$  in order to numerically find an optimal solution:

1. Compute an *a priori* output signal vector  $\tilde{\mathbf{y}}_b^{[k]}[i] = \mathbf{X}_b^{[k]} \mathbf{w}_b^{[k]}[i-1]^*$  at each iteration  $i = 1, 2, \dots$
2. Estimate source signal statistics in the temporary variables  $\hat{p}_b^{[k]}[i]$ ,  $\hat{\mathbf{R}}_b^{[k]}$ , and  $\hat{\mathbf{r}}_b^{[k]}[i]$

$$\begin{aligned}\hat{p}_b^{[k]}[i] &= \frac{1}{N} \tilde{\mathbf{y}}_b^{[k]}[i]^H \tilde{\mathbf{y}}_b^{[k]}[i], \\ \hat{\mathbf{R}}_b^{[k]} &= \frac{1}{N} \mathbf{X}_b^{[k]T} \mathbf{X}_b^{[k]*}, \\ \hat{\mathbf{r}}_b^{[k]}[i] &= \frac{1}{N} \mathbf{X}_b^{[k]T} \text{diag}\{\tilde{\mathbf{y}}_b^{[k]}[i]^*\}^2 \tilde{\mathbf{y}}_b^{[k]}[i],\end{aligned}$$

where  $\text{diag}\{ \}$  produces a diagonal matrix.

3. Compute a temporary filter weight vector  $\mathbf{w}_+^{[k]}$

$$\begin{aligned}\text{KSICA} &: \mathbf{w}_+^{[k]} = \mathbf{w}_b^{[k]}[i-1] + \frac{\hat{\mathbf{R}}_b^{[k]-1} \hat{\mathbf{r}}_b^{[k]}[i]}{2\hat{p}_b^{[k]}[i]}, \\ \text{FastICA} &: \mathbf{w}_+^{[k]} = -\mathbf{w}_b^{[k]}[i-1] + \frac{\hat{\mathbf{R}}_b^{[k]-1} \hat{\mathbf{r}}_b^{[k]}[i]}{2\hat{p}_b^{[k]}[i]}.\end{aligned}$$

4. Update the normalized filter weight vector

$$\mathbf{w}_b^{[k]}[i] = \frac{\mathbf{w}_+^{[k]}}{\|\mathbf{w}_+^{[k]}\|_2}.$$

5. When a stopping criterion is met, or if a specific number of iterations has passed, let  $\mathbf{w}_{b,\text{opt}}^{[k]} = \mathbf{w}_b^{[k]}[i]$  and stop the iterations for this batch. Otherwise, go to 1).

One may initialize the starting point  $\mathbf{w}_b^{[k]}[0]$  as a random vector with unit norm.

## 6 On Applied BSE using ICA

In an applied real-time application of ICA, it is necessary to use sample-based estimators to estimate the statistics of the unknown sources since their true statistics are generally not directly available. Expectations are replaced by their sample-based averages, where Auto Regressive (AR) averages are convenient while their “memory length” or integration time can easily be changed by adjusting a small set of parameters [1]. In this paper, we have made use of a first-order AR averaging technique, where an approximation  $\hat{m}_x^{[k]}[n]$  of the expectation operator for the signal  $x^{[k]}[n]$  is defined as

$$\hat{m}_x^{[k]}[n] = \lambda^{[k]}\hat{m}_x^{[k]}[n-1] + (1 - \lambda^{[k]})x^{[k]}[n]. \quad (24)$$

The parameter  $\lambda^{[k]} \in [0, 1]$  is a constant associated with the integration time of the AR-average.

### 6.1 Online Parameter Estimation

The KSICA and the FastICA share the expectations  $E\left\{\left|\tilde{y}^{[k]}[n]\right|^2\right\}$ ,  $E\left\{\mathbf{x}^{[k]}[n]\mathbf{x}^{[k]}[n]^H\right\}$  and  $E\left\{\tilde{y}^{[k]}[n]^* \left|\tilde{y}^{[k]}[n]\right|^2 \mathbf{x}^{[k]}[n]\right\}$ . These expectations are herein approximated using the following first-order AR averages:

$$\hat{p}^{[k]}[n] = \lambda^{[k]}\hat{p}^{[k]}[n-1] + (1 - \lambda^{[k]})\left|\tilde{y}^{[k]}[n]\right|^2, \quad (25)$$

$$\hat{\mathbf{R}}^{[k]}[n] = \lambda^{[k]}\hat{\mathbf{R}}^{[k]}[n-1] + (1 - \lambda^{[k]})\mathbf{x}^{[k]}[n]\mathbf{x}^{[k]}[n]^H, \quad (26)$$

$$\hat{\mathbf{r}}^{[k]}[n] = \lambda^{[k]}\hat{\mathbf{r}}^{[k]}[n-1] + (1 - \lambda^{[k]})\tilde{y}^{[k]}[n]^* \left|\tilde{y}^{[k]}[n]\right|^2 \mathbf{x}^{[k]}[n]. \quad (27)$$

The inverse of the estimator in (26) is found through the matrix inversion lemma [23], as

$$\begin{aligned} \mathbf{t}^{[k]} &= \hat{\mathbf{R}}^{[k]}[n-1]^{-1} \mathbf{x}^{[k]}[n], \\ \hat{\mathbf{R}}^{[k]}[n]^{-1} &= \lambda^{[k]-1} \hat{\mathbf{R}}^{[k]}[n-1]^{-1} \\ &\quad - \frac{\mathbf{t}^{[k]} \mathbf{x}^{[k]}[n]^H \hat{\mathbf{R}}^{[k]}[n-1]^{-1}}{\lambda^{[k]2} (1 - \lambda^{[k]})^{-1} + \lambda^{[k]} \mathbf{x}^{[k]}[n]^H \mathbf{t}^{[k]}}, \end{aligned} \quad (28)$$

where the temporary vector  $\mathbf{t}^{[k]}$  has been incorporated for the sake of clarity in the presentation. The online KSICA update equation is

$$\mathbf{w}_+^{[k]} = \mathbf{w}^{[k]}[n-1] + \frac{\hat{\mathbf{R}}^{[k]}[n]^{-1} \hat{\mathbf{r}}^{[k]}[n]}{2\hat{\rho}^{[k]}[n]}, \quad (29)$$

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\left\| \mathbf{w}_+^{[k]} \right\|_2}. \quad (30)$$

The online FastICA update equation looks similar to this:

$$\mathbf{w}_+^{[k]} = -\mathbf{w}^{[k]}[n-1] + \frac{\hat{\mathbf{R}}^{[k]}[n]^{-1} \hat{\mathbf{r}}^{[k]}[n]}{2\hat{\rho}^{[k]}[n]}, \quad (31)$$

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\left\| \mathbf{w}_+^{[k]} \right\|_2}. \quad (32)$$

## 7 Evaluation of KSICA and FastICA

Two different evaluations are performed to get an overall picture of the proposed method. First, the KSICA and FastICA methods are evaluated in a batch processing approach according to Section 5. The second part of the evaluation deals with an analysis of the methods' capabilities to extract human speech from an observed mixture of speech and interfering noise. In the second part, online estimates are used to measure the source statistics, and the methods are operating in an online mode according to Section 6.

## 7.1 Evaluation in a Batch Processing Mode

Two sources and two sensors are used in this part of the evaluation, and the elements of the mixing matrix  $\mathbf{A}_b^{[k]}$  from (1) are randomized and computed so that the Unitary property of  $\mathbf{A}_b^{[k]}$  is preserved.

Two cases are evaluated; first, the sources are assumed to have circular probability distribution functions where the source signals are  $s_{b,0}^{[k]}[n] = a_{b,0}^{[k]}[n]e^{j\phi_{b,0}^{[k]}[n]}$  and  $s_{b,1}^{[k]}[n] = a_{b,1}^{[k]}[n]e^{j\phi_{b,1}^{[k]}[n]}$ , where  $a_{b,0}^{[k]}[n]$ ,  $a_{b,1}^{[k]}[n]$ ,  $\phi_{b,0}^{[k]}[n]$ , and  $\phi_{b,1}^{[k]}[n]$  are independent real-valued random processes. Here,  $a_{b,0}^{[k]}[n]$  follows a Laplacian distribution which is common in speech models [24],  $a_{b,1}^{[k]}[n]$  follows a Gaussian distribution, while the random phase signals  $\phi_{b,0}^{[k]}[n]$  and  $\phi_{b,1}^{[k]}[n]$  follow a uniform distribution in the interval  $[-\pi, +\pi]$ . In the second case, the source signals are not circular and the real and imaginary parts of  $s_{b,0}^{[k]}[n]$  also follow a Laplacian distribution, whereas the real and imaginary parts of  $s_{b,1}^{[k]}[n]$  follow a Gaussian distribution. The variances of the two sources are identical and set to unity. Furthermore, the two sources are independent. To evaluate the performance of the batch approach, one can analyze the behavior of the vector  $\mathbf{q}_b^{[k]}[i] = \mathbf{w}_b^{[k]}[i]^H \mathbf{A}_b^{[k]}$  at iteration index  $i$ . When the extraction method converges, the absolute value of the elements in  $\mathbf{q}_b^{[k]}[i]$  should have the values one and zero, i.e.,  $|q_{b,0}^{[k]}[i]| = 1$  and  $|q_{b,1}^{[k]}[i]| = 0$  since that implies that  $\mathbf{y}_b^{[k]}[i] = \mathbf{s}_{b,0}^{[k]}$ . Our performance index  $p_b^{[k]}[i]$  measures the deviation of the solution from this desired behavior according to

$$p_b^{[k]}[i] = \frac{|q_{b,1}^{[k]}[i]|^2}{|q_{b,0}^{[k]}[i]|^2}. \quad (33)$$

In other words, if  $p_b^{[k]}[i] = 0$  which corresponds to  $-\infty$  dB, this implies that the source signal is fully extracted. The outcome of 1000 realizations is averaged, and the resulting mean performance index, to-

gether with the standard deviation for various batch sizes, is provided in Figure 1. The analysis shows that the performance of the KSICA method exceeds that of the FastICA when the batch size is lowered. This is true both in the case with circular sources and in the case with non-circular sources. Furthermore, the standard deviation of the performance index is significantly lower in the proposed KSICA method when compared to the FastICA method. This result is further validation that the KSICA method is more appropriate to use in an online setting. It can be added in this discussion that the FastICA approach generally converges within 3 to 5 iterations and that the KSICA converges to the same steady state value within about 7 more iterations.

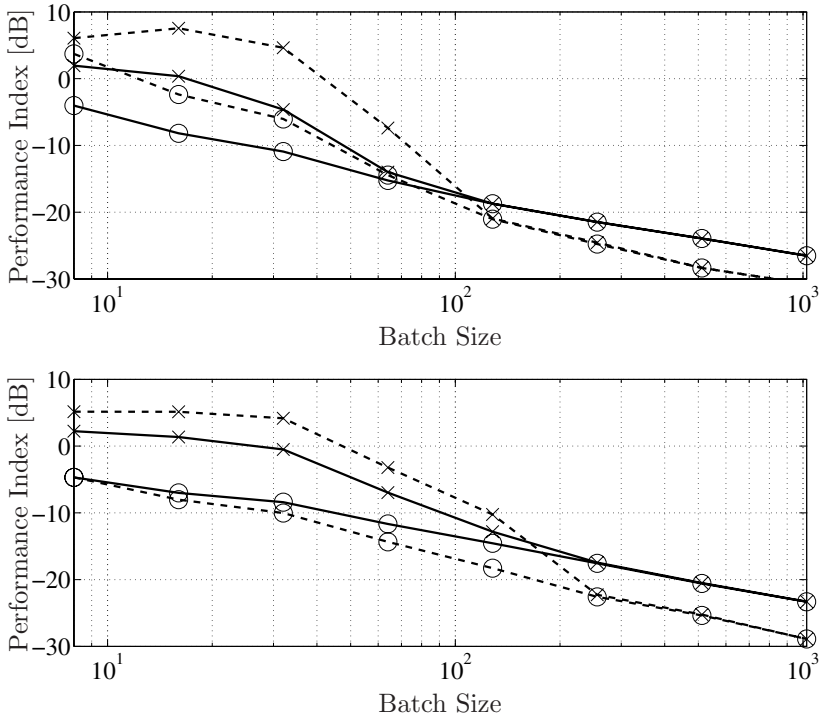


Figure 1: Mean (solid) and standard deviation (dashed) of performance index (33) in dB in a batch processing setup where KSICA “ $\circ$ ” and FastICA “ $\times$ ”. The source signals have circular distributions (upper), and the source signals do not have circular distributions (lower). The mean and standard deviation of the performance index should ideally be  $-\infty$  dB.

## 7.2 Evaluation in an Online Mode

This evaluation concerns the performance of the KSICA and FastICA methods in an online setting in which two microphones are used. Two different source configurations are evaluated: one configuration uses spatially stationary sources and the second configuration uses a spatially non-stationary moving speech source. The configurations are outlined according to Figure 2 where two microphones are situated in  $p_1$  and  $p_2$ , and the noise source is situated in  $p_3$ . For the first configuration the speech source is spatially stationary and situated in  $p_4$ . In the second configuration the speech source is moving along the path  $p_7$  from  $p_5$  to  $p_6$  and is in a constant motion.

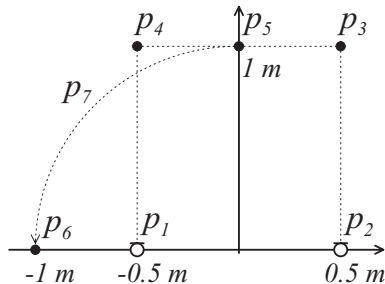


Figure 2: *Source setup in an online mode: The two microphones are situated in  $p_1$  and  $p_2$ , the interfering noise source is situated in  $p_3$ . For a stationary evaluation, the speech source is situated in  $p_4$ , and for a non-stationary evaluation, the speech source is moving along  $p_7$  starting in  $p_5$  and ending in  $p_6$ .*

### 7.2.1 Evaluation with a Spatially Stationary Speech Source

The evaluation assesses improvements in Signal to Interference Ratio (SIR) and degradation of perceptual speech quality according to the ITU-T standard p.862, Perceptual Evaluation of Speech Quality (PESQ) [25]. A key factor in an online BSE method are the learning

rates  $\lambda^{[k]}$  (see Section 6). The performance measures will be assessed for a variety of  $\lambda^{[k]}$ -values in order to provide a complete picture of the methods' performances.

The interfering noise source is spatially and temporally stationary, and it is situated in  $p_3$  (see Figure 2) while the speech source is spatially stationary and temporally non-stationary, and situated in  $p_4$  (see Figure 2). The source signals are pre-recorded with a sampling frequency of 8 kHz, 25 seconds long and subject to free-field propagation [26]. The speech source is active 50 % of the time. The filter bank uses 128 subbands and a two times oversampling. The prototype filter is a 192 tap long Hamming window.

A measure of the SIR improvement and the PESQ measure is used to evaluate the BSE methods in a spatially stationary online configuration. The filter weights at each iteration are stored and used for filtering the original convolved, but unmixed, source signals. This enables direct access to the evaluation measures. The SIR improvement performance measure, denoted  $P_{\text{SIR}}$ , is defined as

$$P_{\text{SIR}} = \frac{\text{Var} \{y_s[t]\} \text{Var} \{x_{0,v}[t]\}}{\text{Var} \{x_{0,s}[t]\} \text{Var} \{y_v[t]\}}, \quad (34)$$

where  $\text{Var} \{ \cdot \}$  denotes an estimator of variance,  $y_s[t]$  and  $y_v[t]$  represent the speech and the interfering noise components of the enhanced output signal, and, similarly, the signals  $x_{0,s}[t]$  and  $x_{0,v}[t]$  represent the speech and interfering noise components of the first simulated microphone signal. The first microphone is used as a reference in the analysis. The PESQ is an automated method for the objective assessment of perceptual sound quality, and it uses a perceptual model of how sound quality is perceived by humans. The PESQ computes a perceptual model for a clean reference speech signal,  $x_{0,s}[t]$ , and a perceptual model for the processed output speech component,  $y_s[t]$ . The perceptual difference between the clean speech signal and the processed speech signal is mapped on the Mean Opinion Score (MOS) [27], yielding a value between one and five where one indicates a poor perceptual speech quality and five indicates an excellent perceptual speech quality.

The evaluated performances of the two BSE methods are presented in Figure 3. This figure shows that the KSICA method provides a stable and good performance for a time constant around 0.15 s, which corresponds to a fast converging method. FastICA provides a performance that is only similar to the KSICA for a time constant above 1 s, which corresponds to a comparatively slower converging method. It may also be noted that the FastICA continuously provides a lower speech quality as opposed to the proposed KSICA method, where the MOS-distance is exceeding one MOS-unit for a time constant around 0.15 s in the 10 dB SIR case and exceeding 0.5 MOS-unit in the -10 dB SIR case. The difference in speech quality decreases between the two methods as the time constant increases.

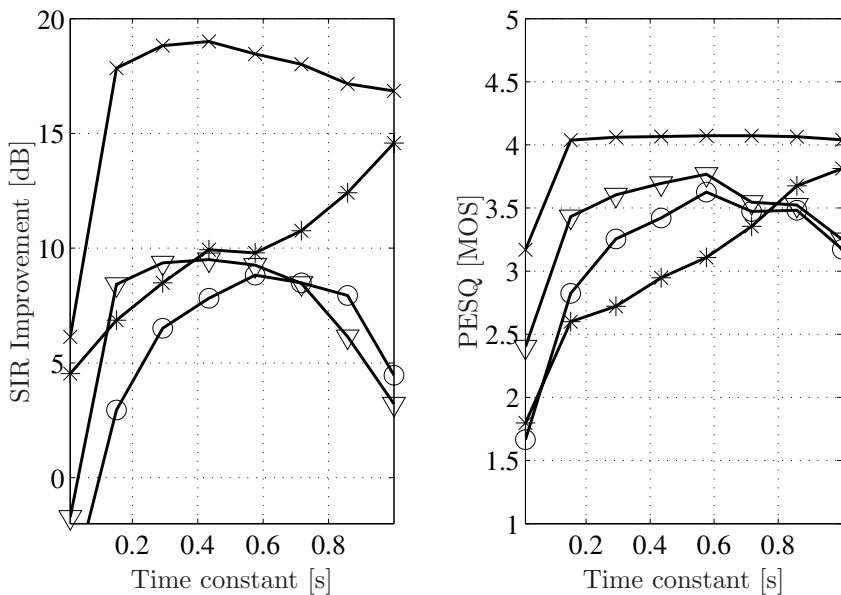


Figure 3: Performance measures SIR improvement (left) and PESQ (right). Input SIR is +10 dB “ $\times$ ” KSICA and “ $*$ ” FastICA. Input SIR is -10 dB “ $\nabla$ ” KSICA and “ $\circ$ ” FastICA.

### 7.2.2 Evaluation with a Moving Speech Source

The interfering noise source is spatially and temporally stationary, and it is situated in  $p_3$  (see Figure 2) while the speech source is spatially and temporally non-stationary, moving along the path  $p_7$  from  $p_5$  to  $p_6$  (see Figure 2). The source signals, the propagation model, and the system parameters are the same as in the previous online evaluation (see Section 7.2.1).

In order to capture the performance in this dynamic configuration, the SIR improvement of the adaptive filter weights is assessed for each frame of the input data (the frame length is 64 samples), and in each subband. It is assumed that the source is stationary during the length of a frame (the radial velocity of the speech source is  $0.0288^\circ$  / frame). The SIR improvement is computed as the mean array gain in the speech source direction over the mean array gain in the direction of the interfering source.

The evaluated performances of the two BSE methods are presented in Figure 4. This figure shows that the KSICA method provides 10 dB higher SIR improvement in relation to the FastICA method when the time constant of the  $\lambda^{[k]}$ -parameters corresponds to 0.15 s. The FastICA does perform as well as the KSICA if the time constant is set to 1 s. However, long time constants are obviously undesirable in a non-stationary application, as can be seen in Figure 4 where the 1 s. setting approaches the 0.15 s. setting first after 20 s of adaptation.

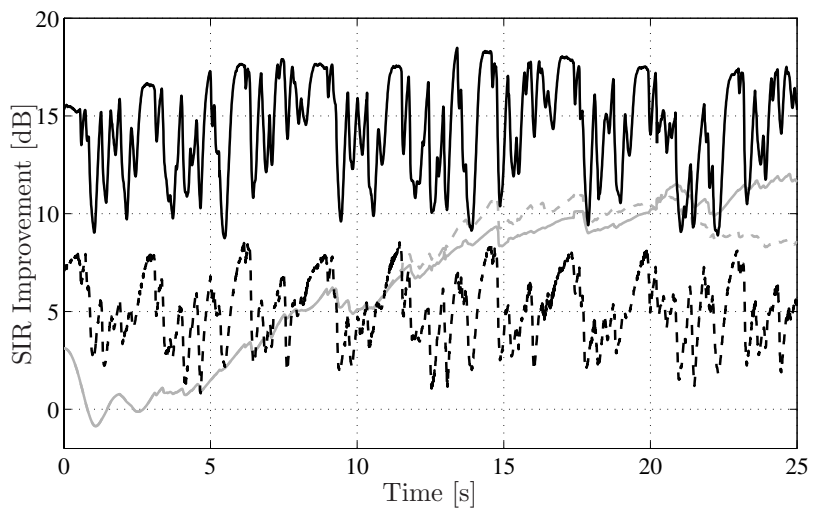


Figure 4: *SIR improvement for KSICA (black, solid) and FastICA (black, dashed) where the time constant of the  $\lambda^{[k]}$ -parameters is 0.15 s. KSICA (grey, solid) and FastICA (grey, dashed) where the time constant of the  $\lambda^{[k]}$ -parameters is 1 s. The input SIR is 0 dB.*

## 8 Summary and Conclusions

This paper presents a new method, denoted KSICA, for online BSE of complex-valued signal mixtures. The proposed method uses a Kurtosis contrast function that is a modification of the FastICA Kurtosis contrast function, and this modification is introduced in order to improve certain aspects of the FastICA method in an applied setting. The KSICA method and the FastICA method are encapsulated by a uniform cost function (13) in which a scalar constant weighs the fourth-order term to the square second-order term in the Kurtosis measure. The weighting introduced by FastICA (based on the Kurtosis contrast function) yields a sensitivity to Gaussian-only signal mixtures, while the FastICA method diverges for such signal mixtures. The proposed method uses a specific choice of weighting parameter in order to circumvent this divergent behavior of the FastICA.

An analysis of the proposed method is derived through comparison with the FastICA method. Evaluation of the method in a batch processing configuration shows that FastICA is sensitive to low batch sizes, whereas the proposed KSICA method is considerably less sensitive. For instance, the performance of the KSICA method is improved by 10 dB (also with 10 dB lower standard deviation) with respect to the FastICA method if the data batch size is less than 100 samples. The two methods both derive from the assumption that the sources have circular distributions. In the batch approach, it is furthermore shown that the KSICA is insensitive to a violation of this circularity assumption. In other words, the KSICA yields good results also when the sources are non-circular. The performance of the two methods equates as the batch size is considerably increased, above 100 data samples. The insensitivity of the proposed KSICA method compared to the FastICA method in a batch setup is seen as an indication that the KSICA method is preferable in an online setup where the requirement of non-Gaussianity for at least the dominant source cannot be constantly guaranteed.

The two methods are further analyzed in an online approach, with

two active sources in a free-field propagation model. The interfering noise source, a Gaussian source, is both spatially and temporally stationary, while the target speech source, with a higher Kurtosis value, is spatially stationary and temporally non-stationary. The statistical measures are estimated in the online setup using received real data only and the KSICA and FastICA update schemes are performed on a sample-by-sample basis. Also, the KSICA method shows significant performance improvements over the FastICA method in the online setup, both in terms of SIR improvement and in preservation of perceptual speech quality. For example, when performed at a time constant of 0.15 s in the algorithm update equation, the KSICA provides a 10 dB higher SIR improvement and more than one MOS-unit better perceptual speech quality in a 10 dB input SIR scenario. In order for the FastICA to reach the same performance, the time constant needs to be increased above 1 s. Furthermore, in an online approach, where the speech source is moving, the KSICA provides about 10 dB higher interference suppression compared to the FastICA at a time constant of 0.15 s. If the time constant is 1 s. the FastICA approaches the performance of KSICA after 20 s.

Further research may include an analysis of various values for the parameter  $\alpha$  in the general cost function (13). This paper only evaluates two values for  $\alpha$ , where  $\alpha = -1$  yields the proposed KSICA method while  $\alpha = 1$  yields the FastICA method (with a Kurtosis contrast function). Future research should also investigate the algorithm's performance under the influence of observation noise. Future analysis should extend the presented evaluation by evaluating real measured data, under various operating conditions, e.g., the number of microphones, the number of subbands, the use of more taps in the filter-and-sum subband beamformer's FIR filters, and changing the oversampling ratio. Also, evaluation of the algorithm performance under double-talker situations is also important to undertake.

## References

- [1] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [3] H. Sawada, S. Araki, R. Mukai, and S. Makino. Blind extraction of a dominant source signal from mixtures of many sources. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 3:61–64, March 2005.
- [4] B. Sällberg, M. Swartling, N. Grbić, and I. Claesson. Real-time implementation of a blind beamformer for subband speech enhancement using kurtosis maximization. *International Workshop on Acoustics, Echo and Noise Control*, pages 485–489, September 2006.
- [5] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of subband kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE 15th International Conference on Digital Signal Processing*, pages 603–606, July 2007.
- [6] B. Sällberg, N. Grbić, and I. Claesson. Online blind speech extraction based on a locally quadratic kurtosis criteria and a preprocessing automatic gain controller. *IEEE 49th International Symposium ELMAR*, pages 139–142, September 2007.
- [7] B. Sällberg, N. Grbić, and I. Claesson. An adaptive blind beamformer with an integrated single-channel noise reduction method for robust realtime blind speech extraction. *IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 309–312, March 2008.

- [8] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [9] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.
- [10] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio. Speech dereverberation via maximum-kurtosis subband adaptive filtering. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 6:3701–3704, May 2001.
- [11] R. Mukai, H. Sawada, S. Araki, and S. Makino. Robust real-time blind source separation for moving speakers in a room. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:469–472, May 2003.
- [12] R. Mukai, H. Sawada, S. Araki, and S. Makino. Blind source separation of many signals in the frequency domain. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:969–972, May 2006.
- [13] E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, February 2000.
- [14] S. C. Douglas. Fixed-point fastica algorithms for the blind separation of complex-valued signal mixtures. *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1320–1325, October 2005.
- [15] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.

- [16] J. F. Cardoso. Source separation using higher order moments. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 4:2109–2112, May 1989.
- [17] A. Cichocki, R. Thawonmas, and S. Amari. Sequential blind signal extraction in order specified by stochastic properties. *IEEE Electronic Letters*, 33(1):64–65, January 1997.
- [18] J. P. LeBlanc and P. L. de Léon. Speech separation by kurtosis maximization. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 2:1029–1032, May 1998.
- [19] F. J. Theis and Y. Inouye. On the use of joint diagonalization in blind signal processing. *IEEE International Symposium on Circuits and Systems*, pages 3586–3589, May 2006.
- [20] S. Y. Low, S. Nordholm, and R. Togneri. Convolutional blind signal separation with post-processing. *IEEE Transactions on Speech and Audio Processing*, 12(5):539–548, September 2004.
- [21] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann. Post-processing for convolutional blind source separation. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:37–40, May 2006.
- [22] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [23] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [24] W. Zhang and S. Gazor. Statistical modelling of speech signals. *IEEE International Conference on Signal Processing*, 1:480–483, August 2002.
- [25] ITU-T p.862. *Perceptual evaluation of speech quality (PESQ)*.
- [26] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.

- [27] ITU-T p.800, Annex B. *Methods for subjective determination of transmission quality*, 1996.



## PART II

# Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson, “Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction”, *accepted for publication to IEEE Signal Processing Letters*, November 2008.

© 2008 IEEE. Reprinted, with permission, from IEEE Signal Processing Letters.

**Modification to the original paper:**

The notations have been standardized so as to fit the other parts of this thesis.

# Statistical Analysis of a Local Quadratic Criterion for Blind Speech Extraction

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

## Abstract

This paper aims at complementing previous empirical work regarding a certain beamforming technique for blind speech extraction that uses a local quadratic approximation of a Kurtosis expression. It is shown here that the proposed method possesses a fixed-point property which means that it remains at an optimal solution once this solution has been reached. The proposed method's fixed-point property is valid for a range of source signals including Gaussian sources. This is an improvement over the FastICA method which diverges at the optimal points that correspond to a Gaussian source. In a real application, it can not be assured that non-Gaussian mixtures are constantly observed, hence, the proposed method is a viable alternative in that case. The fixed-point property further implies that the approximative Kurtosis expression is identical to the true Kurtosis value at an optimal point which, in turn, means that the approximation error is zero. In addition, the convergence towards an optimal solution is always in the direction of a local minimum point even though the optimal solution that correspond to a super-Gaussian source is always a maximum solution which harmonizes with the concept of Kurtosis maximization.

**Keywords:** Array signal processing, Speech enhancement, Higher order statistics.

## 1 Introduction

Speech extraction refers to the process of filtering a received mixture of acoustic signals of which at least one is a speech signal while the other signals are regarded as undesired noise, so that the speech must be extracted or, in fact, enhanced. Beamforming is a versatile speech extraction method as it can filter signals both in the temporal domain and in the spatial domain [1]. A blind adaptive beamformer is able to obtain the speech extraction effect gradually without any explicit references such as knowledge about the microphones' positions and the number and locations of acoustic sources [2]. Instead, some other assumption is made about the source signals [3, 4]. One such assumption is that source signals may carry different Kurtosis values, e.g., noise is often assumed to be Gaussian and having a zero Kurtosis value, while a speech signal generally carries a much higher Kurtosis value [5]. A blind beamformer can therefore be constructed by maximizing the Kurtosis value at the beamformer's output signal [6]. Approximative Newton methods, including a class of FastICA methods based on the Kurtosis contrast function (see e.g. [7, 8]) and a closely related method [9], have been proposed in relation to this. However, the FastICA-like methods require a number of simplifications inside the optimization routine which make their approximations inexact at an optimum solution. A recently proposed technique approximates the Kurtosis value using a local quadratic function which is then solved at each iteration [10, 11, 12]. Robust methods for solving a quadratic problem exist in a rich variety today (see e.g. [2]) which makes the proposed method practically tractable. Furthermore, the quadratic approximation of Kurtosis has been analyzed empirically in the past [10, 11, 12] and this method has shown itself capable of delivering a high level of speech extraction in various adverse environments. However, to date, no statistical analysis has been carried out for this particular approximation technique. The objective of this paper is to provide a statistical analysis of the quadratic Kurtosis criterion in order to explain some of its expected statistical behaviors and benefits.

The outline is as follows: The adopted system and the data model is described in Section 2. The recently introduced approximation technique is provided in Section 3 and a blind beamformer based on this approximation technique is derived in Section 4. A statistical analysis is performed in Section 5, and Section 6 concludes this paper.

## 2 System and Data model

In this section, a convolution model is adopted where a set of source signals are emitted in a room and received by an array of microphones. The adopted filtering (beamformer) structure is commonly denoted as a filter-and-sum beamformer in the frequency domain [1], and the system model of the structure is illustrated in Figure 1.

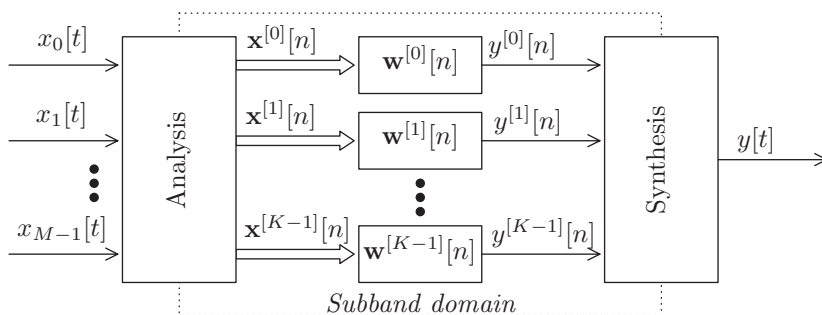


Figure 1: Subband domain filter-and-sum beamformer for  $M$  sensor signals and  $K$  subbands.

### 2.1 System Model

The system model assumes an array of  $M$  microphones where each received real-valued time signal is denoted as  $x_m[t]$  for  $m : m \in \mathbb{N}, m < M$ , where  $t$  represents continuous time. Each received time signal is sampled and decomposed into a time-frequency representation using a filter bank [13] with  $K$  subbands, and where each microphone subband

signal is denoted  $x_m^{[k]}[n]$  with subband index  $k : k \in \mathbb{N}, k < K$  and where  $n$  is a sample index in the subband domain. The reason for using a subband representation of the data is that a convolution model in the time domain corresponds to a multiplicative model in the frequency domain [14, 15]. The received signal vector is composed as

$$\mathbf{x}^{[k]}[n] = \mathbf{A}^{[k]} \mathbf{s}^{[k]}[n]. \quad (1)$$

The matrix  $\mathbf{A}^{[k]}$  of size  $M \times I$  is a source mixing matrix. The signal vector  $\mathbf{s}^{[k]}[n] = (s_0^{[k]}[n], s_1^{[k]}[n], \dots, s_{I-1}^{[k]}[n])^T$  contains the  $I$  original, independent source signals. The notation  $(\cdot)^T$  denotes the vector transpose. The output signal  $y^{[k]}[n]$  of the beamformer is computed by filtering the received signal vector by the beamformer filter vector  $\mathbf{w}^{[k]}[n]$ , i.e.,

$$y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n]. \quad (2)$$

The notation  $(\cdot)^H$  denotes the complex conjugate vector transpose. This particular filtering uses one filter-tap per subband. While it is possible to extend the beamformer filter to include multiple filter-taps, it is not done here in order to simplify the statistical analysis. It is noted that, if  $I > M$  an inverse to  $\mathbf{A}^{[k]}$  does not exist in general and the best solution is found through a pseudo-inverse of  $\mathbf{A}^{[k]}$ . However, in order to conserve the readability in this paper it is assumed that an inverse to  $\mathbf{A}^{[k]}$  exists. The time-domain output signal  $y[t]$  is then computed from the subband output signal  $y^{[k]}[n]$  by a synthesis filter bank matched to the analysis filter bank.

## 2.2 Data Model

The original sources are assumed to have unity power, i.e.,

$$\mathbb{E} \left\{ \mathbf{s}^{[k]}[n] \mathbf{s}^{[k]}[n]^H \right\} = \mathbf{I}_I, \quad (3)$$

where  $E\{ \}$  represents the expectation operation, and  $\mathbf{I}_I$  denotes an identity matrix of the size  $I \times I$ . Following [7, 9], it is furthermore assumed that the sources have circular distributions, which implies that

$$E \left\{ \mathbf{s}^{[k]}[n] \mathbf{s}^{[k]}[n]^T \right\} = \mathbf{0}_{I \times I}. \quad (4)$$

The matrix  $\mathbf{0}_{I \times I}$  is a null-matrix of the size  $I \times I$ . Furthermore, the Kurtosis value of each source is

$$\kappa \left\{ s_i^{[k]}[n] \right\} = E \left\{ \left| s_i^{[k]}[n] \right|^4 \right\} - 2. \quad (5)$$

### 3 A Local Quadratic Criterion

In [10, 11, 12], a local quadratic criterion was derived for the purpose of Kurtosis maximization in a speech extraction application. The criterion was derived based on the inspiring paper by Yang regarding an iterative procedure for subspace tracking [16]. The key idea when constructing a quadratic approximation of a high order expression is to make use of the current filter weights  $\mathbf{w}^{[k]}[n-1]$  to form an *a priori* output signal  $\hat{y}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}^{[k]}[n]$ . Some output signal components  $y^{[k]}[n]$  are then replaced by this *a priori* output signal in the high order expression, e.g., the Kurtosis expression. In the end, this procedure renders an expression that is quadratic with regard to the new beamformer filter weights  $\mathbf{w}^{[k]}[n]$ .

To clarify further derivations, the Kurtosis value of the beamformer's output signal is notated as  $\kappa \left\{ \mathbf{w}^{[k]}[n] \right\} \equiv \kappa \left\{ y^{[k]}[n] \right\}$ , which is equal to

$$\kappa \left\{ \mathbf{w}^{[k]}[n] \right\} = E \left\{ \left| y^{[k]}[n] \right|^4 \right\} - 2E \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2. \quad (6)$$

This expression is obviously of order four with regard to  $\mathbf{w}^{[k]}[n]$ . Following [10, 11, 12], this Kurtosis expression can be approximated and

rewritten into a quadratic form:

$$\begin{aligned} \hat{\kappa} \left\{ \mathbf{w}^{[k]}[n] \right\} &= \mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \left| y^{[k]}[n] \right|^2 \right\} \\ &\quad - 2\mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \operatorname{Re} \left\{ \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^* y^{[k]}[n] \right\} \right\}. \end{aligned} \quad (7)$$

This expression is clearly quadratic with regard to  $\mathbf{w}^{[k]}[n]$ . In this expression, the notation  $(\ )^*$  refers to the complex conjugate, and  $\operatorname{Re} \{ \}$  takes the real part of its argument and it is employed in order to ensure that the approximative Kurtosis expression is real valued. The Brandwood gradient and Hessian [17] of (7) with regard to  $\mathbf{w}^{[k]}[n]^*$  are

$$\begin{aligned} \nabla \hat{\kappa} \left\{ \mathbf{w}^{[k]}[n] \right\} &= \mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 y^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\} \\ &\quad - 2\mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\}, \end{aligned} \quad (8)$$

and

$$\nabla^2 \hat{\kappa} \left\{ \mathbf{w}^{[k]}[n] \right\} = \mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}. \quad (9)$$

There may be other ways to approximate the fourth-order Kurtosis expression that are not outlined here regarding which terms  $y^{[k]}[n]$  are exchanged with  $\tilde{y}^{[k]}[n]$  in the approximation. Such reformulations can be the focus of future research.

## 4 Blind Adaptive Beamforming

The proposed blind adaptive beamformer continuously updates its beamformer weights  $\mathbf{w}^{[k]}[n]$  in order to obtain a maximization of the output signal's Kurtosis value. While the approximative Kurtosis expression (7) is quadratic with regard to  $\mathbf{w}^{[k]}[n]$  its optimal solution is found at a point where the gradient with regard to  $\mathbf{w}^{[k]}[n]$  evaluates to zero, i.e.,

$\nabla \hat{\kappa} \{ \mathbf{w}^{[k]}[n] \} = \mathbf{0}_{M \times 1}$ . However, experiments on real data have shown that this direct approach leads to a highly fluctuating solution with a degraded speech quality as a result. A viable approach to reduce the fluctuations is by using a smoothed filter update equation:

$$\begin{aligned} \mathbf{w}_+^{[k]} &= \mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \cdot \\ &\quad \mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}^{-1} \cdot \\ &\quad \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\}, \end{aligned} \quad (10)$$

where  $0 < \gamma^{[k]} \leq 1$  is a smoothing parameter. A temporary vector  $\mathbf{w}_+^{[k]}$  is used in order to compute a unity-norm constraint for the updated filter coefficients as  $\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\|\mathbf{w}_+^{[k]}\|_2}$ .

When the algorithm has converged to an optimal solution it is expected that  $\mathbf{w}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]$ . If this is the case,  $y^{[k]}[n] = \tilde{y}^{[k]}[n]$  which in turn yields that  $\hat{\kappa} \{ \mathbf{w}^{[k]}[n] \} = \kappa \{ \mathbf{w}^{[k]}[n] \}$ , i.e., the introduced approximative Kurtosis value is equal to the true Kurtosis value. The next section will show that the quadratic method has a fixed-point property which means that the method stays at an optimal solution when it finds one, which, in turn, yields that the approximation error is zero. In addition, an analysis of the local consistency will show that the Hessian matrix is negative definite around an optimal solution. The update equation (10) has to compensate for the negative definiteness by a negative update direction.

## 5 Statistical Analysis

In [7, 9] a statistical analysis was used for a FastICA method and a closely related method. A similar analysis is used here for the local quadratic Kurtosis criterion.

It is assumed that the received signal vector is made uncorrelated prior to the analysis using a Principal Component Analysis (PCA) preprocessing stage. The PCA decorrelation has the same impact as if the mixing matrix  $\mathbf{A}^{[k]}$  would possess a Unitary property, i.e.,  $\mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} = \mathbf{A}^{[k]} \mathbb{E} \left\{ \mathbf{s}^{[k]}[n] \mathbf{s}^{[k]}[n]^H \right\} \mathbf{A}^{[k]H} = \mathbf{A}^{[k]} \mathbf{A}^{[k]H} = \mathbf{I}_M$ .

Two properties of the blind beamformer are analyzed: the fixed-point property and the local consistency. If an algorithm possesses a fixed-point property it means that the algorithm stays at an optimal solution when it finds one. The local consistency is used to analyze the algorithm behavior around an optimal solution. It is thus possible to, for instance, determine what kind of local optimal solution it is (a maximal solution, a minimal solution, or a saddle point).

### 5.1 Fixed-point Behavior

In order to analyze the fixed-point behavior it is assumed that the algorithm has found an optimal solution that extracts the  $i^{\text{th}}$  source, i.e.,  $\mathbf{w}^{[k]}[n-1] = \mathbf{w}_{\text{opt}}^{[k]}$  so that  $\tilde{y}^{[k]}[n] = \mathbf{w}_{\text{opt}}^{[k]H} \mathbf{x}^{[k]}[n] = s_i^{[k]}[n]$ . The filter vector update equation is therefore equal to

$$\begin{aligned} \mathbf{w}_+^{[k]} &= \mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\} \mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}^{-1} \\ &\quad \mathbb{E} \left\{ s_i^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\}. \end{aligned} \quad (11)$$

The various statistical expectations above are evaluated as

$$\mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\} = 1, \quad (12)$$

$$\mathbb{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\}^{-1} = \mathbf{A}^{[k]} \mathbf{\Lambda}_i^{[k]-1} \mathbf{A}^{[k]H}, \quad (13)$$

$$\mathbb{E} \left\{ s_i^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\} = \mathbf{A}^{[k]} \mathbf{e}_i^{[k]}. \quad (14)$$

The matrix  $\mathbf{\Lambda}_i^{[k]}$  (size  $I \times I$ ) is a diagonal matrix, whose diagonal entries have the value one, except for the  $i^{\text{th}}$  diagonal element that has the

value  $\kappa \left\{ s_i^{[k]}[n] \right\} + 2$ . The vector  $\mathbf{e}_i^{[k]}$  is a basis vector with zero-valued elements except for at the position  $i$  where the element has the value one. The filter vector update equation is

$$\begin{aligned} \mathbf{w}_+^{[k]} &= \mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbf{A}^{[k]} \boldsymbol{\Lambda}_i^{[k]-1} \mathbf{e}_i^{[k]} = \\ &= \mathbf{w}^{[k]}[n-1] - \frac{\gamma^{[k]}}{\kappa \left\{ s_i^{[k]}[n] \right\} + 2} \mathbf{A}^{[k]} \mathbf{e}_i^{[k]}. \end{aligned} \quad (15)$$

The optimal solution is such that  $\mathbf{w}_{\text{opt}}^{[k]H} \mathbf{A}^{[k]} \mathbf{s}^{[k]}[n] = s_i^{[k]}[n]$ , hence  $\mathbf{e}_i^{[k]} = \mathbf{A}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]}$  and therefore  $\mathbf{w}_{\text{opt}}^{[k]} = \mathbf{A}^{[k]} \mathbf{e}_i^{[k]}$ , which yields that

$$\mathbf{w}_+^{[k]} = \left( 1 - \frac{\gamma^{[k]}}{\kappa \left\{ s_i^{[k]}[n] \right\} + 2} \right) \mathbf{w}_{\text{opt}}^{[k]}. \quad (16)$$

Hence, provided that  $1 - \frac{\gamma^{[k]}}{\kappa \left\{ s_i^{[k]}[n] \right\} + 2} \neq 0$ , i.e.,  $\gamma^{[k]} \neq \kappa \left\{ s_i^{[k]}[n] \right\} + 2$ , it is clear that the blind beamformer (10) possesses the fixed-point property:

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}_+^{[k]}}{\left\| \mathbf{w}_+^{[k]} \right\|_2} = \pm \mathbf{w}_{\text{opt}}^{[k]}, \quad (17)$$

While  $0 < \gamma^{[k]} \leq 1$ , there is only one extreme case where the algorithm does not possess the fixed-point property, i.e.,  $\gamma^{[k]} = \kappa \left\{ s_i^{[k]}[n] \right\} + 2$ , namely if  $-2 < \kappa \left\{ s_i^{[k]}[n] \right\} \leq -1$ , which implies that the source  $s_i^{[k]}[n]$  is sub-Gaussian. In all other cases, the algorithm possesses a fixed-point property that is also valid for a Gaussian source ( $\kappa \left\{ s_i^{[k]}[n] \right\} = 0$ ).

## 5.2 Local Consistency

A second order Taylor series expansion is computed at a point  $\mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]}$  around the optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$ , where  $\mathbf{p}^{[k]}$  is a small pertur-

bation vector (the term *small* implies that  $\|\mathbf{p}^{[k]}\|_2 \ll 1$ ), according to

$$\begin{aligned} \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]} \right\} &= \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} + 2\text{Re} \left\{ \mathbf{p}^{[k]H} \nabla \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} \right\} \\ &\quad + \mathbf{p}^{[k]H} \nabla^2 \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} \mathbf{p}^{[k]}. \end{aligned} \quad (18)$$

The gradient vector  $\nabla \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\}$  is

$$\begin{aligned} \nabla \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} &= \text{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} \mathbf{w}_{\text{opt}}^{[k]} \\ &\quad - 2\text{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \right\} \text{E} \left\{ s_i^{[k]}[n]^* \mathbf{x}^{[k]}[n] \right\} = \\ &= \mathbf{A}^{[k]} \mathbf{\Lambda}_i^{[k]} \mathbf{e}_i^{[k]} - 2\mathbf{A}^{[k]} \mathbf{e}_i^{[k]} = \\ &= \kappa \left\{ s_i^{[k]}[n] \right\} \mathbf{w}_{\text{opt}}^{[k]}. \end{aligned} \quad (19)$$

The Hessian matrix  $\nabla^2 \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\}$  is evaluated as

$$\begin{aligned} \nabla^2 \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} &= \text{E} \left\{ \left| s_i^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \right\} = \\ &= \mathbf{A}^{[k]} \mathbf{\Lambda}_i^{[k]} \mathbf{A}^{[k]H}. \end{aligned} \quad (20)$$

Inserting (19) and (20) into (18) yields

$$\begin{aligned} \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]} \right\} &= \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} + 2\kappa \left\{ s_i^{[k]}[n] \right\} \text{Re} \left\{ \mathbf{p}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]} \right\} \\ &\quad + \mathbf{p}^{[k]H} \mathbf{A}^{[k]} \mathbf{\Lambda}_i^{[k]} \mathbf{A}^{[k]H} \mathbf{p}^{[k]}. \end{aligned} \quad (21)$$

The unity norm constraint renders the relationship  $2\text{Re} \left\{ \mathbf{p}^{[k]H} \mathbf{w}_{\text{opt}}^{[k]} \right\} = -\|\mathbf{p}^{[k]}\|_2^2$  that yields

$$\begin{aligned} \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]} \right\} &= \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} - \left( \kappa \left\{ s_i^{[k]}[n] \right\} - 1 \right) \|\mathbf{p}^{[k]}\|_2^2 \\ &\quad + \left( \kappa \left\{ s_i^{[k]}[n] \right\} + 1 \right) \mathbf{p}^{[k]H} \mathbf{A}^{[k]} \mathbf{\Lambda}_i^{[k]} \mathbf{A}^{[k]H} \mathbf{p}^{[k]}. \end{aligned} \quad (22)$$

The matrix  $\Delta_i^{[k]}$  (size  $I \times I$ ) has elements that are zero-valued except for a single element at position  $(i, i)$  which has the value one. Obviously,  $0 \leq \mathbf{p}^{[k]H} \mathbf{A}^{[k]} \Delta_i^{[k]} \mathbf{A}^{[k]H} \mathbf{p}^{[k]} \leq \|\mathbf{p}^{[k]}\|_2^2$ , which equals  $\alpha^{[k]} \|\mathbf{p}^{[k]}\|_2^2 = \mathbf{p}^{[k]H} \mathbf{A}^{[k]} \Delta_i^{[k]} \mathbf{A}^{[k]H} \mathbf{p}^{[k]}$  for a factor  $\alpha^{[k]} \in [0, 1]$ :

$$\hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} + \mathbf{p}^{[k]} \right\} = \hat{\kappa} \left\{ \mathbf{w}_{\text{opt}}^{[k]} \right\} - \left( \left( 1 - \alpha^{[k]} \right) \left( \kappa \left\{ s_i^{[k]}[n] \right\} + 2 \right) - 3 + \alpha^{[k]} \right) \left\| \mathbf{p}^{[k]} \right\|_2^2. \quad (23)$$

While  $\|\mathbf{p}^{[k]}\|_2^2 \geq 0$ , the type of optimal solution  $\mathbf{w}_{\text{opt}}^{[k]}$  is dependent on the relationship between  $\alpha^{[k]}$  and the Kurtosis value  $\kappa \left\{ s_i^{[k]}[n] \right\}$  through the sign of  $(1 - \alpha^{[k]}) \left( \kappa \left\{ s_i^{[k]}[n] \right\} + 2 \right) - 3 + \alpha^{[k]}$ . This relationship is shown in Figure 2. An optimal solution that corresponds to a source with the Kurtosis value  $-2 < \kappa \left\{ s_i^{[k]}[n] \right\} < 1$  (this includes a Gaussian source with  $\kappa \left\{ s_i^{[k]}[n] \right\} = 0$ ) is always, independent of the value of  $\alpha^{[k]}$ , a local minimum solution. An optimal solution corresponding to a super-Gaussian source, with  $\kappa \left\{ s_i^{[k]}[n] \right\} > 1$ , can be any kind of local optimum (maximum, minimum, or saddle point) depending on the value of  $\alpha^{[k]}$ . However, when the algorithm starts to converge towards an optimal solution, it implies that  $\alpha^{[k]} \approx 1$ , independent of what Kurtosis value the source carries. This becomes clear if it is assumed that  $\mathbf{p}^{[k]} = \delta \mathbf{w}_{\text{opt}}^{[k]}$ , for a value  $\delta > 0$ , while  $\alpha^{[k]} = 1$ . Hence, the Hessian matrix in (9) is negative definite and the algorithm always converges towards a local minimum solution. In order to compensate for the negative definiteness of the Hessian matrix it is necessary for the algorithm to update in a direction opposite to a point where the gradient evaluates to zero in (10). However, provided that  $\gamma^{[k]} \neq \kappa \left\{ s_i^{[k]}[n] \right\} + 2$ , the fixed-point property renders  $\mathbf{p}^{[k]} = \mathbf{0}_{M \times 1}$  (i.e.  $\alpha^{[k]} = 0$ ) at an optimal solution. An optimal solution that corresponds to a super-Gaussian source ( $\kappa \left\{ s_i^{[k]}[n] \right\} > 1$ ) is always a maximum solution which harmo-

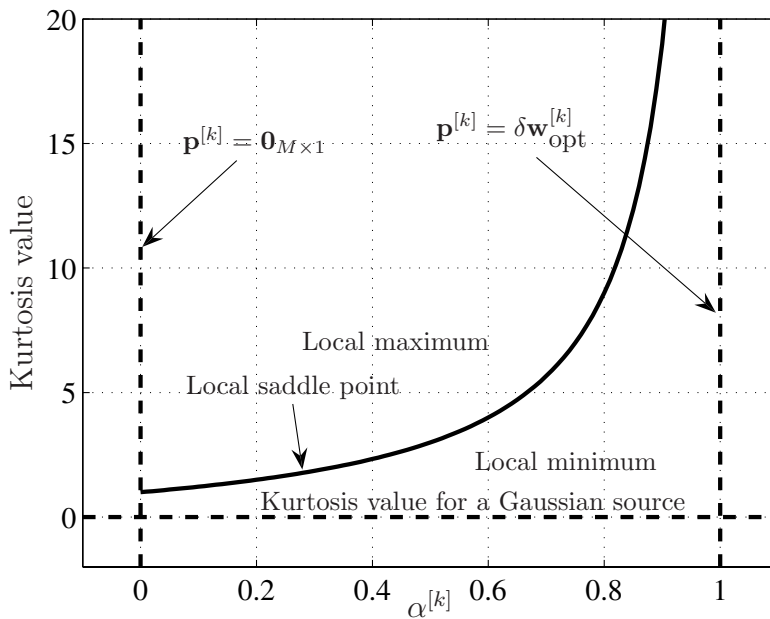


Figure 2: The solid curve shows where the solution is a saddle point, a maximum solution (points above the curve), and a minimum solution (points below the curve).

nizes with the concept of Kurtosis maximization.

## 6 Conclusions

A recent method for adaptive beamforming applied to blind speech extraction was proposed in [10, 11, 12]. The method uses an approximation of the fourth order Kurtosis value in order to construct a local quadratic expression. The proposed method is practically viable while the quadratic approximation can be solved using standard methods, e.g., the Recursive Least Squares (RLS) [2, 16]. Several empirical anal-

yses was conducted in [10, 11, 12], and they showed the method's ability to extract speech in adverse environments. A statistical analysis of this method is provided in this paper with respect to the fixed-point property and the local consistency. It is shown that the proposed method possesses a fixed-point property if the optimal solution does not correspond to a sub-Gaussian source, i.e., with a Kurtosis value in the range  $-2 < \kappa \left\{ s_i^{[k]}[n] \right\} \leq -1$ . This behavior is important in a real application where it can not always be guaranteed that non-Gaussian sources are active, and it may be contrasted with the popular FastICA method (see e.g. [7, 9]) that has been shown to diverge for a Gaussian source, i.e.,  $\kappa \left\{ s_i^{[k]}[n] \right\} = 0$ . The fixed-point property also tells us that once the algorithm has converged to an optimal solution, it stays at this optimal solution. This renders that the introduced quadratic approximation of Kurtosis is equal to the true Kurtosis value at an optimal solution, i.e., the approximation error is equal to zero at an optimal solution. In addition, it is shown that the introduced quadratic approximation renders a local minimum in the direction towards an optimal solution. This means that the Hessian matrix in (9) is negative definite and the optimization approach results in a minimization strategy instead of a maximization strategy, even though the optimal solution that correspond to a super-Gaussian source ( $\kappa \left\{ s_i^{[k]}[n] \right\} > 1$ ) is always a maximum solution which harmonizes with the concept of Kurtosis maximization.

## References

- [1] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [2] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.

- [4] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [5] W. Zhang and S. Gazor. Statistical modelling of speech signals. *IEEE International Conference on Signal Processing*, 1:480–483, August 2002.
- [6] Z. Ding. A new algorithm for automatic beamforming. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2:689–693, November 1991.
- [7] E. Bingham and A. Hyvärinen. A fast fixed-point algorithm for independent component analysis of complex valued signals. *International Journal of Neural Systems*, 10(1):1–8, February 2000.
- [8] S. C. Douglas. Fixed-point fastica algorithms for the blind separation of complex-valued signal mixtures. *IEEE Asilomar Conference on Signals, Systems and Computers*, pages 1320–1325, October 2005.
- [9] B. Sällberg, N. Grbić, and I. Claesson. Complex-valued independent component analysis for online blind speech extraction. *IEEE Transactions on Speech and Audio Processing*, 16(8):1624–1632, November 2008.
- [10] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of sub-band kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE 15th International Conference on Digital Signal Processing*, pages 603–606, July 2007.
- [11] B. Sällberg, N. Grbić, and I. Claesson. Online blind speech extraction based on a locally quadratic kurtosis criteria and a preprocessing automatic gain controller. *IEEE 49th International Symposium ELMAR*, pages 139–142, September 2007.

- [12] B. Sällberg, N. Grbić, and I. Claesson. An adaptive blind beamformer with an integrated single-channel noise reduction method for robust realtime blind speech extraction. *IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 309–312, March 2008.
- [13] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [14] P. Smaragdīs. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [15] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.
- [16] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, January 1995.
- [17] D. H. Brandwood. A complex gradient operator and its application in adaptive array theory. *Proceedings of the IEE*, 130(1):11–16, February 1983.



## PART III

# **Online Maximization of Subband Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction**

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson, “Online Maximization of Sub-band Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction”, *IEEE 15th International Conference on Digital Signal Processing*, 2007.

© 2007 IEEE. Reprinted, with permission, from IEEE 15th International Conference on Digital Signal Processing.

**Modification to the original paper:**

The notations have been standardized so as to fit the other parts of this thesis.

# Online Maximization of Subband Kurtosis for Blind Adaptive Beamforming in Realtime Speech Extraction

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

## **Abstract**

This paper presents a method for blind beamforming with application in realtime speech extraction in a non-stationary environment. The blind beamforming is carried out using an on-line Kurtosis maximization approach where the optimization is based on Newton's method. The main novelty of the paper lies in the approximated subband Kurtosis measure where a locally quadratic criterion is solved at each iteration. Further, a realtime digital signal processor (DSP) implementation of the method is conducted and results with real data are presented.

**Keywords:** Blind Speech Extraction, Adaptive Beamforming, Kurtosis Maximization.

## 1 Introduction

Blind adaptive beamforming has features that are attractive for speech enhancement in human communication. The motivation for employing a beamformer is that it uses several microphones, thus operating in the spatiotemporal domain, and it has a higher degree of freedom as opposed to corresponding single-channel techniques that utilize only the temporal domain [1]. The inherent virtue of a blind control algorithm in beamforming is that no knowledge about the spatiotemporal environment is needed, such as the position of the sources relative to the microphone array or knowledge regarding the physical dimension of the array itself. The merging of an adaptive beamformer with a blind control algorithm results in a structure that continuously tracks sources in a changing environment [2].

This paper presents a novel approach of approximating the subband Kurtosis measure using a local quadratic criterion. The intended application is Blind Speech Extraction (BSE), where a dominant speech source (dominant in the Kurtosis measure) is extracted from an observed convolutive mixture of sources [3, 4, 5]. The proposed method in this paper is an extension of [6], where the new formulation of the Kurtosis measure provides significantly faster convergence. This proposed approach stands in contrast to Blind Signal Separation (BSS) in which all dominant sources, or groups of sources, are separated, see [7, 8].

The outline of this paper is as follows. The background to the proposed method is presented in Section 2. The assumed signal model and the beamforming notation are given in Section 3. The proposed approximated Kurtosis measure is provided in Section 4, and a gradient based maximization thereof is given in Section 5. A realtime Digital Signal Processor (DSP) implementation is presented in Section 6 to illustrate the method's robustness and performance. Evaluation results are given in Section 7, and a summary with conclusions and topics for future research are provided in Section 8.

## 2 Background

One fundamental approach for ICA is to exploit the non-Gaussianity of the observed signals [3, 4]. This may be achieved by utilizing the Kurtosis measure as an estimate of Gaussianity. The estimated Kurtosis measure of a beamformer's output signal can then control the update of a beamformer so as to maximize the output Kurtosis value [2]. By doing so, it is intended that the beamformer blocks the noise signal(-s) (low Kurtosis value) and passes the speech component (high Kurtosis value).

In order to successfully apply blind algorithms in the application of speech extraction, it is necessary to construct a time-varying and robust structure to allow for source movement tracking and to compensate for a changing background noise environment. Many of the fundamental ICA methods require stationary instantaneous mixtures (at least for the duration of each data batch) to yield robust performance [3, 9]. Hence, some of these methods are not suitable in a real environment where the spatiotemporal settings may vary rapidly. The main focus in this paper is therefore on Newton's method, utilizing an estimate of the Kurtosis measure in each subband and allowing for fast source tracking in such a time-varying environment.

In many BSE/BSS algorithms, there is a need for a whitening (decorrelation) of the input signals in order to improve the convergence speed [3, 4]. This is often done in a preprocessing step, e.g., by using the Principle Component Analysis (PCA) [10, 11]. However, in a non-stationary environment such as the environment intended here, the tracking performance of the PCA method may potentially deteriorate the performance of the following online ICA algorithm. Therefore, the proposed method does not utilize prewhitening and operates directly upon the observed signals using an online update of the beamformer filter.

The permutation ambiguity of subband ICA may pose a potential problem in many applications and may therefore require special attention [8]. However, the permutation problem is not considered an

issue in this paper because we focus on the extraction of one dominant speech source (dominant in the Kurtosis measure) that is convolved and mixed with one or many noise sources with low Kurtosis values.

### 3 Signal Model

In this paper, we assume one dominant desired source and one or many undesired sources (with lower Kurtosis values). The sources' relative positions to the beamformer are unknown, and the beamformer's spatial configuration is also unknown. The beamformer employs  $M$  microphones that sense the acoustical wavefield, and the recorded time signals for each microphone  $m : m \in \mathbb{N}, m < M$  at time index  $t$  are denoted  $x_m[t]$ . The set of all microphone signals is represented by the vector  $\mathbf{x}[t] = (x_0[t], x_1[t], \dots, x_{M-1}[t])^T$ , where the superscript  $(\ )^T$  denotes the transpose. The received time signals are efficiently decomposed into a time-frequency representation, denoted  $\mathbf{x}^{[k]}[n]$ , where  $k : k \in \mathbb{N}, k < K$  is the subband index and  $n$  is the block time index, using a polyphase realization of Discrete Fourier Transform (DFT) modulated analysis filterbank [12]. The observed convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [5], and the observed subband signals are assumed:

$$\mathbf{x}^{[k]}[n] = \mathbf{h}^{[k]}[n]s^{[k]}[n] + \mathbf{v}^{[k]}[n], \quad (1)$$

where  $\mathbf{h}^{[k]}[n]$  represents a spatiotemporal transfer function related to the desired speech source with source signal  $s^{[k]}[n]$  and  $\mathbf{v}^{[k]}[n]$  represents the subband noise component for subband index  $k$ . A linear weighting of this subband input signal using a time-varying beamformer filter vector  $\mathbf{w}^{[k]}[n] = (w_0^{[k]}[n], w_1^{[k]}[n], \dots, w_{M-1}^{[k]}[n])^T$ , denoted a filter-and-sum beamformer [1], yields a subband output signal

$$y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n], \quad (2)$$

where the superscript  $(\ )^H$  denotes the Hermitian transpose. The time domain output signal  $y[t]$  is efficiently reconstructed from the subband

output signals  $y^{[k]}[n]$  using a polyphase DFT modulated synthesis filterbank matched to the analysis filterbank [12].

## 4 Approximation of Subband Kurtosis Measure

A listing of 16 different variants of the Kurtosis measure for complex valued data is given in [13]. One of these variants of the Kurtosis measure is applied herein for the beamformer's subband output signal

$$\kappa \{y^{[k]}[n]\} = \text{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\} - 2\text{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2 - \left| \text{E} \left\{ y^{[k]}[n]^2 \right\} \right|^2, \quad (3)$$

where  $\text{E} \{ \cdot \}$  represents the expectation operator.  $\kappa \{y^{[k]}[n]\}$  designates the Kurtosis value of the signal  $y^{[k]}[n]$ , and it is approximated in this paper by the time-varying function  $\hat{\kappa}^{[k]}[n]$ , i.e.,  $\hat{\kappa}^{[k]}[n] \approx \kappa \{y^{[k]}[n]\}$ . This approximation utilizes an *a-priori* output signal denoted  $\tilde{y}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}^{[k]}[n]$  and is formulated as

$$\begin{aligned} \hat{\kappa}^{[k]}[n] = & \text{E} \left\{ \left| y^{[k]}[n] \right|^2 \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \\ & - 2\text{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \text{Re} \left\{ \text{E} \left\{ y^{[k]}[n] \tilde{y}^{[k]}[n]^* \right\} \right\} \\ & - \text{Re} \left\{ \text{E} \left\{ \tilde{y}^{[k]}[n]^2 \right\}^* \text{E} \left\{ y^{[k]}[n] \tilde{y}^{[k]}[n] \right\} \right\}, \end{aligned} \quad (4)$$

where the operator  $\text{Re} \{ \cdot \}$  takes the real part of its argument and  $(\cdot)^*$  denotes the complex conjugate. The real-operator is introduced to ensure that this approximation of the Kurtosis measure is real valued and that it forms an analytic function of  $\mathbf{w}^{[k]}[n]$ . The objective is now to maximize the approximated Kurtosis measure  $\hat{\kappa}^{[k]}[n]$  in (4) by continuously updating the filter  $\mathbf{w}^{[k]}[n]$ , using information in the previous filter vector  $\mathbf{w}^{[k]}[n-1]$ . The introduced approximation using the *a-priori* output signal is inspired by the derivation of the Projection Approximation Subspace Tracking (PAST) technique in [10].

## 5 Gradient-based Kurtosis Maximization

We may rewrite the approximated Kurtosis measure in (4) using (2) as

$$\begin{aligned}\hat{\kappa}^{[k]}[n] &= \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \mathbf{w}^{[k]}[n] \\ &\quad - 2\mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \operatorname{Re} \left\{ \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]^* \right\} \right\} \\ &\quad - \operatorname{Re} \left\{ \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^2 \right\}^* \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n] \right\} \right\}, \quad (5)\end{aligned}$$

The approximation of the beamformer's output signal Kurtosis value in (5) is (locally) quadratic in the filter vector  $\mathbf{w}^{[k]}[n]$ , and the gradient at block time index  $n$  with respect to the filter vector  $\mathbf{w}^{[k]}[n]$  is

$$\begin{aligned}\frac{\partial \hat{\kappa}^{[k]}[n]}{\partial \mathbf{w}^{[k]}[n]^*} &= \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \mathbf{w}^{[k]}[n] \\ &\quad - 2\mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]^* \right\} \\ &\quad - \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^2 \right\}^* \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n] \right\}. \quad (6)\end{aligned}$$

Optimization of the approximative Kurtosis value, according to Newton's method [14], follows

$$\begin{aligned}\mathbf{w}^{[k]}[n] &= \mathbf{w}^{[k]}[n-1] - \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \left| \tilde{y}^{[k]}[n] \right|^2 \right\}^{-1} \\ &\quad \left( 2\mathbb{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\} \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]^* \right\} \right. \\ &\quad \left. + \mathbb{E} \left\{ \tilde{y}^{[k]}[n]^2 \right\}^* \mathbb{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n] \right\} \right). \quad (7)\end{aligned}$$

A feasible implementation of the method is achieved through the introduction of a set of suitable approximations. Auto-regressive averaging

is one type of approximation (see [4]) that are used here

$$\mathbf{P}^{[k]}[n] \approx \mathbf{E} \left\{ \mathbf{x}^{[k]}[n] \mathbf{x}^{[k]}[n]^H \left| \tilde{y}^{[k]}[n] \right|^2 \right\}^{-1}, \quad (8)$$

$$u^{[k]}[n] \approx \mathbf{E} \left\{ \left| \tilde{y}^{[k]}[n] \right|^2 \right\}, \quad (9)$$

$$z^{[k]}[n] \approx \mathbf{E} \left\{ \tilde{y}^{[k]}[n]^2 \right\}, \quad (10)$$

$$\mathbf{u}^{[k]}[n] \approx \mathbf{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]^* \right\}, \quad (11)$$

$$\mathbf{z}^{[k]}[n] \approx \mathbf{E} \left\{ \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n] \right\}. \quad (12)$$

The matrix  $\mathbf{P}^{[k]}[n]$  in (8) is updated according to the matrix inversion lemma [14] as

$$\mathbf{t}^{[k]} = \mathbf{P}^{[k]}[n-1] \mathbf{x}^{[k]}[n], \quad (13)$$

$$\begin{aligned} \mathbf{P}^{[k]}[n] &= \lambda^{[k]-1} \mathbf{P}^{[k]}[n-1] \\ &\quad - \frac{\left| \tilde{y}^{[k]}[n] \right|^2 \mathbf{t}^{[k]} \mathbf{x}^{[k]}[n]^H \mathbf{P}^{[k]}[n-1]}{\lambda^{[k]2} + \lambda^{[k]} \left| \tilde{y}^{[k]}[n] \right|^2 \mathbf{x}^{[k]}[n]^H \mathbf{t}^{[k]}}, \end{aligned} \quad (14)$$

where  $\mathbf{t}^{[k]}$  is a temporary vector, and the parameter  $\lambda^{[k]} \in [0, 1]$  controls the convergence rate (i.e., the tracking performance) of the method. First order auto-regressive averages are used in (9) to (12):

$$u^{[k]}[n] = \lambda^{[k]} u^{[k]}[n-1] + (1 - \lambda^{[k]}) \left( \tilde{y}^{[k]}[n] \right)^2, \quad (15)$$

$$z^{[k]}[n] = \lambda^{[k]} z^{[k]}[n-1] + (1 - \lambda^{[k]}) \tilde{y}^{[k]}[n]^2, \quad (16)$$

$$\mathbf{u}^{[k]}[n] = \lambda^{[k]} \mathbf{u}^{[k]}[n-1] + (1 - \lambda^{[k]}) \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]^*, \quad (17)$$

$$\mathbf{z}^{[k]}[n] = \lambda^{[k]} \mathbf{z}^{[k]}[n-1] + (1 - \lambda^{[k]}) \mathbf{x}^{[k]}[n] \tilde{y}^{[k]}[n]. \quad (18)$$

The update equation in (7) is now approximated as

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbf{P}^{[k]}[n] \mathbf{\Delta}^{[k]}[n]}{\left\| \mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbf{P}^{[k]}[n] \mathbf{\Delta}^{[k]}[n] \right\|_2}, \quad (19)$$

where

$$\Delta^{[k]}[n] = 2u^{[k]}[n]\mathbf{u}^{[k]}[n] + z^{[k]}[n]^*\mathbf{z}^{[k]}[n]. \quad (20)$$

The parameter  $\gamma^{[k]} \in [0, 1]$  is introduced in order to control the fluctuations in the filter weights due to the random input data. The normalization in (19) has been incorporated in order to avoid the trivial solution  $\mathbf{w}^{[k]}[n] = \mathbf{0}$ . However, this normalization is not optimal with respect to the spectral whitening of the enhanced speech since only  $\|\mathbf{w}^{[k]}[n]\|_2 = 1$  is ensured, and the term  $|\mathbf{w}^{[k]}[n]^H \mathbf{h}^{[k]}[n]|$  is thereby not guaranteed to equal unity. It may be noted that, except for the normalization, the filter update in (19) is similar to the well-known Recursive Least Squares (RLS) algorithm [14]. A suitable initialization of this algorithm is  $\mathbf{P}^{[k]}[0] = \mathbf{I}_M$ , where  $\mathbf{I}_M$  is the  $(M \times M)$  identity matrix,  $u^{[k]}[0] = z^{[k]}[0] = 0$ ,  $\mathbf{u}^{[k]}[0] = \mathbf{z}^{[k]}[0] = \mathbf{0}_{M \times 1}$ , and  $\mathbf{w}^{[k]}[0] = (1, 0, \dots, 0)^T$ .

## 6 Realtime Implementation

The implementation of the proposed method is made on a floating point DSP named SHARC-21262 from Analog Devices. This is a high-performance DSP that supports effective parallel computations through the Single Instruction Multiple Data (SIMD) mode, suitable for vector-based operations. The algorithm is efficiently implemented using a transformed approach to reduce the overhead. A rule-of-thumb for the transformed approach is to perform all computations so that the largest data dimension is used in the element-wise vector operations. This alternative way of managing the data was introduced in [6] for a MATLAB-based realtime implementation. The transformation approach reduces the number of computations in this DSP platform by 7.5 times for a typical vector operation [15].

## 6.1 System Configuration

The filterbank configuration used  $K = 64$  subbands, with four times oversampling. The prototype filter was designed using the window method with a Hamming window. This configuration was selected as a trade-off between the audio quality performance and the introduced signal delay, measured to 8.5 ms. The algorithmic parameters were set such that the integration time of the  $\lambda^{[k]}$  parameter was 0.2 s and the integration time of the  $\gamma^{[k]}$  parameter was 1.5 s.

## 7 Evaluation Results

The performance of the proposed method is analyzed first using an offline setting with real measured data. This setting allows comparison with the analytically tractable Wiener beamformer [1]. The realtime implementation is analyzed using short-term power estimates where the algorithm may be switched on and off.

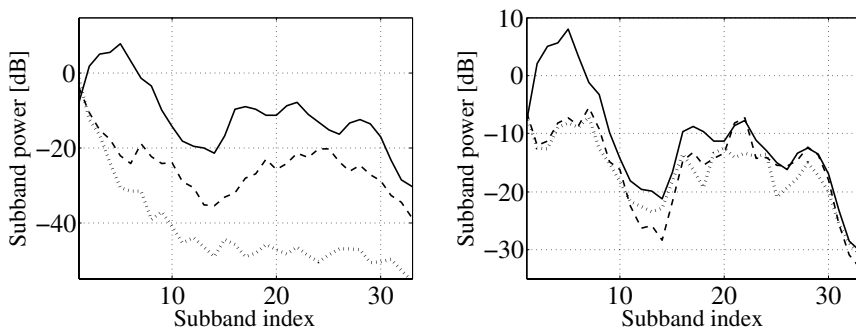


Figure 1: Estimated subband power (left) for the speech sequence (solid), ferry engine noise (dotted), and factory noise (dashed). Beamformer's spectral effect on the enhanced speech (right) with original speech (solid), enhanced speech in the ferry engine noise case (dotted), and enhanced speech in the factory noise case (dashed).

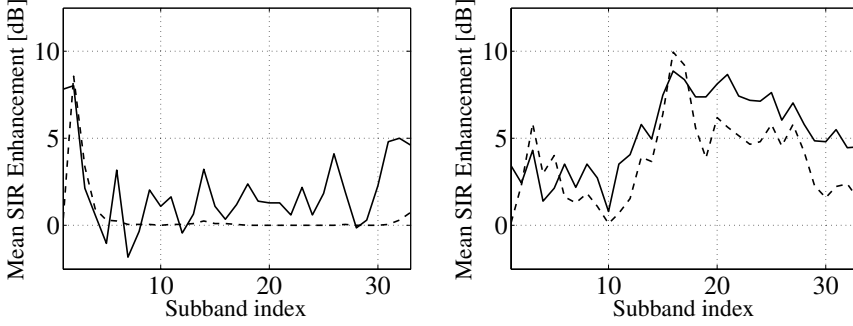


Figure 2: Subband mean SIR enhancement for speech and ferry engine noise (left) and speech and factory noise (right). The proposed method (solid) and Wiener Beamformer (dashed).

## 7.1 Evaluation Measures

The mean Signal to Interference Ratio (SIR) enhancement  $Q^{[k]}$  is analyzed in the subband domain using

$$Q^{[k]} = \frac{\text{Var} \{y_s^{[k]}[t]\} \text{Var} \{x_v^{[k]}[t]\}}{\text{Var} \{y_v^{[k]}[t]\} \text{Var} \{x_s^{[k]}[t]\}}, \quad (21)$$

where  $\text{Var} \{ \cdot \}$  designates an estimator of variance and the subband signals  $x_s^{[k]}[t]$ ,  $x_v^{[k]}[t]$ ,  $y_s^{[k]}[t]$ , and  $y_v^{[k]}[t]$  represent the components of the input and output signals that are related to the speech and interference, respectively. The mean SIR enhancement in the time domain,  $Q$ , is evaluated as

$$Q = \frac{\text{Var} \{y_s[t]\} \text{Var} \{x_v[t]\}}{\text{Var} \{y_v[t]\} \text{Var} \{x_s[t]\}}, \quad (22)$$

A short-term power estimate  $q[t]$  is used in the time domain to evaluate the method's realtime performance. The short-term power of a signal  $x[t]$  is estimated according to

$$q[t] = \alpha q[t-1] + (1-\alpha)x[t]^2, \quad (23)$$

---

where the parameter  $\alpha \in [0, 1]$  is selected so as to yield a 10 ms integration time.

## 7.2 Offline Results using Real Data

Human speech (male and female) was sent through a loudspeaker and recorded using two microphones separated by 5 cm in an office room (reverberation time  $RT_{60} = 130$  ms) with sampling frequency 8 kHz. Previously recorded ferry engine noise and factory noise were subsequently emitted and recorded using the same setup. The signals' subband power estimates are given on the left side of Figure 1. The speech signal is then mixed at 0 dB SIR with each of the two interfering noise signals. The subband mean SIR enhancement is provided in Figure 2. This figure shows that the proposed method increases the subband mean SIR at several of the higher frequencies. The beamformer's spectral effect on the enhanced speech is given on the right side of Figure 1. The time domain mean SIR enhancement in the ferry engine noise case is 10.6 dB for the proposed method and 12.4 dB for the Wiener beamformer. In the factory noise case, the mean SIR enhancement was 5.6 dB for the proposed method and 8.8 dB for the Wiener beamformer.

## 7.3 Online Results using Real Data

The online evaluation includes prerecorded signals consisting of male speech spatially added with music. It is provided in [16]. The speaker and the music were recorded in an office room using two microphones where the distance between the sources and the microphones is 60 cm in a square ordering. The two input channels are used as input to the proposed realtime DSP implementation, and the resulting output time signal is presented in Figure 3 with a corresponding short-term power estimate. The proposed method is activated after 6.8 s. The two input channels are exchanged momentarily at 13.7 s in order to illustrate the method's capability to track signals in a non-stationary environment. The SIR enhancement is 15 dB to 20 dB in this case.

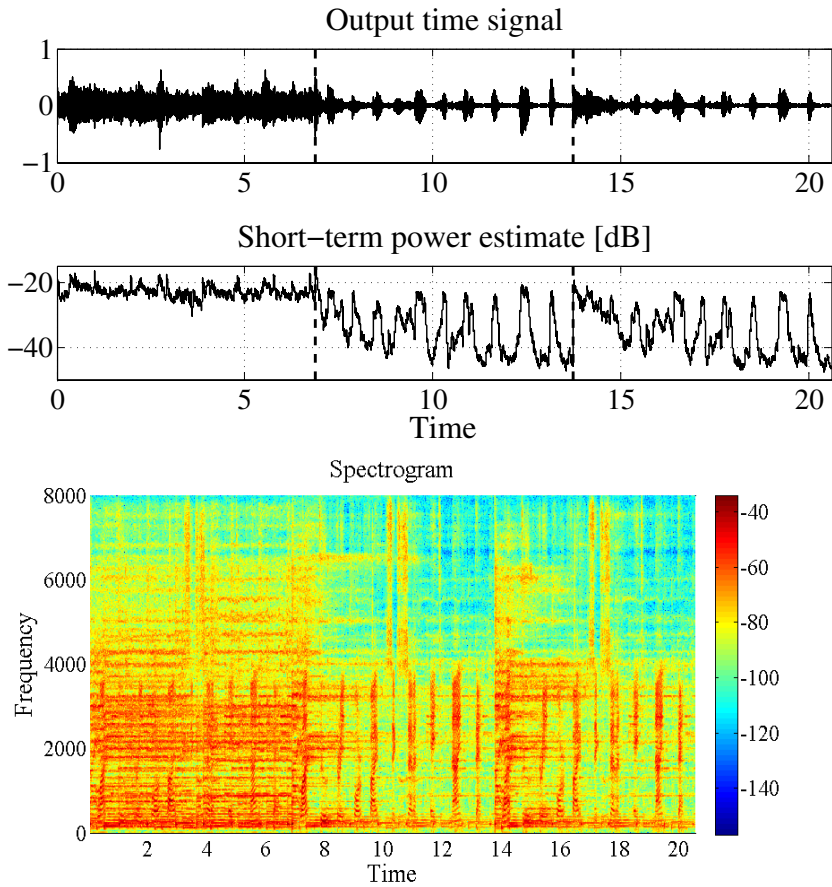


Figure 3: Output time signal for speech and music (upper), short-term power estimate (middle), and corresponding spectrogram (lower). The proposed method is activated at 6.8 s, and the input channels are exchanged at 13.7 s.

## 8 Summary and Conclusions

This paper presents a blind beamformer based on the well-known Kurtosis maximization strategy. The novelty of this contribution lies in the introduced approximations where the subband Kurtosis measure is approximated locally using a quadratic criterion. This criterion is continuously updated online according to a Newton-based search method, a variant of the RLS method. The method successfully extracts speech in convolutive mixtures of speech and ferry engine noise, speech and factory noise, and speech and music, at the cost of introduced spectral whitening of the enhanced speech. It is noticeable that, except for the spectral whitening, the speech distortion is perceptually very low. This method shows promising performance for speech enhancement in real environments. However, some questions are still open for future research:

The strategy for normalization of the adaptive filter used here, (19), is not optimal with respect to the spectral whitening of the enhanced speech. Other normalization strategies may reduce this undesired spectral whitening.

A natural extension of our proposed method is to introduce a power normalization of the approximated subband Kurtosis measure in order to yield a scale-invariant solution.

A rigorous analysis of the spatiotemporal behavior of this proposed method is required in order to relate and compare its performance to existing state-of-the-art solutions.

## References

- [1] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [2] Z. Ding. A new algorithm for automatic beamforming. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2:689–693, November 1991.

- [3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [4] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [5] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [6] B. Sällberg, M. Swartling, N. Grbić, and I. Claesson. Real-time implementation of a blind beamformer for subband speech enhancement using kurtosis maximization. *International Workshop on Acoustics, Echo and Noise Control*, pages 485–489, September 2006.
- [7] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.
- [8] R. Mukai, H. Sawada, S. Araki, and S. Makino. Blind source separation of many signals in the frequency domain. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:969–972, May 2006.
- [9] F. J. Theis and Y. Inouye. On the use of joint diagonalization in blind signal processing. *IEEE International Symposium on Circuits and Systems*, pages 3586–3589, May 2006.
- [10] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, January 1995.
- [11] R. Badeau, K. Abed-Meraim, G. Richard, and B. David. Sliding window orthonormal past algorithm. *IEEE International Conference on Acoustic, Speech and Signal Processing*, 5:261–264, April 2003.

- [12] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [13] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [14] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [15] Z. Yermeche, B. Sällberg, N. Grbić, and I. Claesson. Real-time dsp implementation of a subband beamforming algorithm for dual microphone speech enhancement. *IEEE International Symposium on Circuits and Systems*, pages 353–356, May 2007.
- [16] T. W. Lee. Blind source separation: Audio examples. [Online] [http://www.cnl.salk.edu/~tewon/Blind/blind\\_audio.html](http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html), Feb. 17 2007.



## PART IV

# **An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Ro- bust Realtime Blind Speech Ex- traction**

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson, “An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Robust Realtime Blind Speech Extraction”, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2008.

© 2008 IEEE. Reprinted, with permission, from IEEE International Conference on Acoustics, Speech, and Signal Processing.

**Modification of the original paper:**

Measurement data related to a method that first uses a blind beamformer and then a single-channel speech enhancer on the beamformer output data has been added to Figure 1 and Figure 2 using a red plot color. The corresponding figure labels have been added with the text “*The SIR and PESQ improvements of the adaptive blind beamformer when the AGE method is applied to the beamformer’s output signal (red)*”. The introduced modification does not alter the results that were presented in the original paper, it merely clarifies them.

In addition, the notations have been standardized so as to fit the other parts of this thesis.

# An Adaptive Blind Beamformer with an Integrated Single-channel Noise Reduction Method for Robust Realtime Blind Speech Extraction

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

## Abstract

The performance of single-channel temporal noise reduction methods generally deteriorates in high noise environments, whereas spatial beamformers can maintain some level of speech enhancement. This paper presents a solution where a low complexity single-channel noise reduction method is integrated into the feedback control loop of an adaptive blind beamformer with the purpose of robust blind speech extraction in high noise environments. The proposed combined system outperforms each of the individual methods with respect to signal-to-interference ratio improvement for a wide range of operating conditions and where the loss in estimated perceptual speech quality due to the combined system is tolerably low. Furthermore, the excess processing load in a hardware solution is comparatively insignificant for the proposed extended approach.

**Keywords:** Speech enhancement, Array signal processing.

## 1 Introduction

Classical approaches for speech enhancement in human communication are typically based on single-channel noise reduction methods, where the data from a single microphone are used to perform the noise reduction [1, 2, 3]. Inherent in single-channel methods is the necessity to trade-off the opposing design aspects of speech distortion and noise reduction. Due to the fact that single-channel techniques are limited to the temporal domain, they generally provide a high degree of speech distortion when the noise reduction level is increased, which is often needed in a high noise environment.

Blind adaptive beamforming has features that are attractive for speech enhancement in human communication. The motivation for employing a beamformer is that it uses several microphones, thus operating in the spatiotemporal domain [4], and it has a higher degree of freedom as opposed to single-channel methods that utilize only the temporal domain. The inherent virtue of a blind control method in beamforming is that no knowledge about the spatiotemporal environment is needed, such as the position of the sources relative to the microphone array or knowledge regarding the physical dimension of the array itself [5, 6, 7]. The merging of an adaptive beamformer with a blind control method results in a structure that continuously tracks sources in a changing environment [8].

This paper investigates an approach where a single-channel noise reduction method [2, 3] is integrated into the feedback control loop of a recently proposed adaptive blind beamforming technique [9, 10]. The intended application is Blind Speech Extraction (BSE), where a dominant speech source (dominant in the Kurtosis measure) is extracted from an observed convolutive mixture of sources [5, 6, 7, 11]. The idea of integrating a noise reduction method into the feedback control loop of a blind beamformer is, to the best knowledge of the authors, novel. The approach provides a successful symbiosis where the spatial processing of the blind beamformer aids the temporal processing of the noise reduction method, and vice-versa. This is emphasized in

the evaluation where the performance of the proposed approach is increasingly better than any of the individual systems and even better than the linear addition of the individual systems' performances. The speech quality deterioration (according to the ITU-T standard P.862, Perceptual Evaluation of Speech Quality (PESQ) [12]) of the proposed system is tolerably low, and the increased processing load due to the combined system is comparatively insignificant.

The outline of this paper is as follows. The assumed signal model and the beamforming notation are given in Section 2. The proposed structure with a single-channel noise reduction method and an adaptive blind beamformer is presented in Section 3. Evaluation results are given in Section 4, and a summary with conclusions is provided in Section 5.

## 2 Signal Model

In this paper, we assume one dominant desired source (with highest Kurtosis value) and one or many undesired sources. It is further assumed that the speech has a stationarity time that is much shorter than the interfering noise. The sources' relative positions to the beamformer are unknown, and the beamformer's spatial configuration is also unknown. The beamformer employs  $M$  microphones that sense the acoustical wavefield, and the recorded time signal for each microphone is denoted  $x_m[t]$  for  $m : m \in \mathbb{N}, m < M$ , with time index  $t$ . The sampled received time signals are efficiently decomposed into a time-frequency representation, denoted  $\mathbf{x}^{[k]}[n]$ , where  $k : k \in \mathbb{N}, k < K$  is the subband index and  $n$  is the subband time index, using a poly-phase realization of a Discrete Fourier Transform (DFT) modulated analysis filterbank [13]. The observed convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [7], and the observed subband signals are assumed to be

$$\mathbf{x}^{[k]}[n] = \mathbf{h}^{[k]}[n]s^{[k]}[n] + \mathbf{v}^{[k]}[n], \quad (1)$$

where  $\mathbf{h}^{[k]}[n]$  represents a spatiotemporal transfer function related to the desired speech source with source signal  $s^{[k]}[n]$  and  $\mathbf{v}^{[k]}[n]$  represents the subband noise component for subband index  $k$ . A linear weighting of this subband input signal using a time-varying beamformer filter vector  $\mathbf{w}^{[k]}[n] = (w_0^{[k]}[n], w_1^{[k]}[n], \dots, w_{M-1}^{[k]}[n])^T$ , where  $(\cdot)^T$  denotes the transpose, yields a subband output signal

$$y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n], \quad (2)$$

where  $(\cdot)^H$  denotes the Hermitian transpose. The time output signal  $y[t]$  is efficiently reconstructed from the subband output signals  $y^{[k]}[n]$  using a polyphase DFT modulated synthesis filterbank matched to the analysis filterbank [13].

### 3 The Proposed Structure

The original formulation of the adaptive blind beamformer in [9, 10] used two signals in its feedback control loop: the input signal vector  $\mathbf{x}^{[k]}[n]$  and an *a-priori* beamformer output signal

$$\tilde{y}^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}^{[k]}[n]. \quad (3)$$

The idea of this paper is to integrate a low-complexity single-channel noise reduction method in the feedback control loop of this adaptive blind beamformer. The *a-priori* output signal of the adaptive blind beamformer,  $\tilde{y}^{[k]}[n]$ , is used as input to the noise reduction method. The output signal of the noise reduction method is denoted  $\bar{y}^{[k]}[n] = g^{[k]}[n]\tilde{y}^{[k]}[n]$ , where a real valued gain function  $g^{[k]}[n] \in [0, 1]$  is applied in each subband to facilitate the noise reduction effect. Any phase mismatch between the input signal vector  $\mathbf{x}^{[k]}[n]$  and the noise reduced signal  $\bar{y}^{[k]}[n]$  will deteriorate the beamformer's performance. The same gain function is therefore applied to the input signal vector prior to the beamformer's control loop in order to nullify this performance limitation. The input signal vector is  $\bar{\mathbf{x}}^{[k]}[n] = g^{[k]}[n]\mathbf{x}^{[k]}[n]$ . This section will

first present the single-channel noise reduction method and thereafter the proposed adaptive blind beamforming method.

### 3.1 Single-channel Noise Reduction Method

The single-channel noise reduction method used in this paper is the Adaptive Gain Equalizer (AGE) [2, 3]. The AGE is selected due to its inherent simplicity and because it does not require a supplementary structure, like a Voice Activity Detector (VAD), which is required by many other noise reduction techniques such as the spectral subtraction type of methods [1]. The AGE operates in a subband domain and utilizes a real valued gain function  $g^{[k]}[n]$  per each subband  $k$  in order to impose the noise reduction to its input signal. The input signal to the AGE method is in our case an *a-priori* output signal of the adaptive blind beamformer,  $\tilde{y}^{[k]}[n]$ , and the output signal of the AGE method is denoted  $\bar{y}^{[k]}[n] = g^{[k]}[n]\tilde{y}^{[k]}[n]$ . Two averages,  $a_{\text{fast}}^{[k]}[n]$  and  $a_{\text{slow}}^{[k]}[n]$ , are the key elements of the AGE. These averages are intended to track the speech bursts and the background noise floor level, respectively. The averages are realized using first order auto-regressive filters

$$a_{\text{fast}}^{[k]}[n] = \alpha_{\text{fast}}^{[k]} a_{\text{fast}}^{[k]}[n-1] + \left(1 - \alpha_{\text{fast}}^{[k]}\right) \left|\tilde{y}^{[k]}[n]\right|, \quad (4)$$

$$T^{[k]} = \alpha_{\text{slow}}^{[k]} a_{\text{slow}}^{[k]}[n-1] + \left(1 - \alpha_{\text{slow}}^{[k]}\right) \left|\tilde{y}^{[k]}[n]\right|, \quad (5)$$

$$a_{\text{slow}}^{[k]}[n] = \min\left(T^{[k]}, a_{\text{fast}}^{[k]}[n]\right), \quad (6)$$

where  $T^{[k]}$  is a temporary variable and  $\alpha_{\text{fast}}^{[k]}$  and  $\alpha_{\text{slow}}^{[k]}$  are constants associated to the integration time of the two averages  $a_{\text{fast}}^{[k]}[n]$  and  $a_{\text{slow}}^{[k]}[n]$ , respectively. The function  $\min(a, b)$  selects the minimal value of its two parameters  $a$  and  $b$ , and it is used to ensure that  $a_{\text{fast}}^{[k]}[n] \geq a_{\text{slow}}^{[k]}[n]$ , i.e., that  $a_{\text{fast}}^{[k]}[n]/a_{\text{slow}}^{[k]}[n] \geq 1$  for all  $k$  and  $n$ . If the parameters  $\alpha_{\text{fast}}^{[k]}$  and  $\alpha_{\text{slow}}^{[k]}$  are chosen so that the integration time of  $a_{\text{fast}}^{[k]}[n]$  is close to speech pseudo-stationarity time (20-50 ms) and the

integration time of  $a_{\text{slow}}^{[k]}[n]$  has a time frame matched to the slowly varying background noise (in the order of seconds) then the quotient  $a_{\text{fast}}^{[k]}[n]/a_{\text{slow}}^{[k]}[n]$  will be close to unity when speech is not present and  $a_{\text{fast}}^{[k]}[n]/a_{\text{slow}}^{[k]}[n] \gg 1$  when a speech burst is present. The different temporal properties of the speech and the background noise are used to form the real valued noise reducing gain function  $g^{[k]}[n]$  that continuously tracks the speech level, i.e., speech bursts, without the need of a supplementary VAD structure. The AGE utilizes the quotient of the two averages in order to construct the gain function

$$g^{[k]}[n] = f^{[k]} \left( \frac{a_{\text{fast}}^{[k]}[n]}{a_{\text{slow}}^{[k]}[n]} \right), \quad (7)$$

where the function  $f^{[k]}(\cdot)$  inhibits the quotient from ever exceeding unity. The inhibiting function  $f^{[k]}(\cdot)$  can typically be selected as a hard clipping function [2, 3]

$$f^{[k]}(x) = \begin{cases} \frac{x}{G^{[k]}}, & \text{if } \frac{x}{G^{[k]}} < 1 \\ 1, & \text{if } \frac{x}{G^{[k]}} \geq 1 \end{cases}, \quad (8)$$

where  $G^{[k]} > 1$  is a real valued subband specific maximal allowed noise reduction level. The resulting effect is that the gain function is bounded to  $\frac{1}{G^{[k]}} \leq g^{[k]}[n] \leq 1$  for all  $k$  and  $n$ . This means that if no speech is present and  $a_{\text{fast}}^{[k]}[n] \approx a_{\text{slow}}^{[k]}[n]$ , then  $g^{[k]}[n] \approx \frac{1}{G^{[k]}}$  and the noise reduction is maximal, whereas if a speech burst is present and  $a_{\text{fast}}^{[k]}[n] \gg a_{\text{slow}}^{[k]}[n]$ , then  $g^{[k]}[n] \approx 1$  and it becomes an all-pass filter.

### 3.2 Proposed Adaptive Blind Beamforming Method

A listing of 16 different variants of the Kurtosis measure for complex valued data is given in [14]. One of these variants of the Kurtosis measure was applied in [9, 10] for the beamformer's subband output

signal

$$\kappa \{y^{[k]}[n]\} = \mathbb{E} \left\{ \left| y^{[k]}[n] \right|^4 \right\} - 2\mathbb{E} \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2 - \left| \mathbb{E} \left\{ y^{[k]}[n]^2 \right\} \right|^2, \quad (9)$$

where  $\mathbb{E} \{ \cdot \}$  represents the expectation operator.  $\kappa \{y^{[k]}[n]\}$  designates the Kurtosis value of the signal  $y^{[k]}[n]$ , and it was approximated in the previous works by the time-varying function  $\hat{\kappa}^{[k]}[n]$ , i.e.,  $\hat{\kappa}^{[k]}[n] \approx \kappa \{y^{[k]}[n]\}$ . The approximation in [9, 10] utilized the *a-priori* output signal  $\tilde{y}^{[k]}[n]$  in its control loop, whereas, in this paper, a set of noise-reduced signals  $\bar{y}^{[k]}[n]$  and  $\bar{\mathbf{x}}^{[k]}[n]$  is used instead, according to

$$\begin{aligned} \hat{\kappa}^{[k]}[n] &= \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \bar{\mathbf{x}}^{[k]}[n] \bar{\mathbf{x}}^{[k]}[n]^H \left| \bar{y}^{[k]}[n] \right|^2 \right\} \mathbf{w}^{[k]}[n] \\ &\quad - 2\mathbb{E} \left\{ \left| \bar{y}^{[k]}[n] \right|^2 \right\} \operatorname{Re} \left\{ \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \bar{\mathbf{x}}^{[k]}[n] \bar{y}^{[k]}[n]^* \right\} \right\} \\ &\quad - \operatorname{Re} \left\{ \mathbb{E} \left\{ \bar{y}^{[k]}[n]^2 \right\}^* \mathbf{w}^{[k]}[n]^H \mathbb{E} \left\{ \bar{\mathbf{x}}^{[k]}[n] \bar{y}^{[k]}[n] \right\} \right\}, \quad (10) \end{aligned}$$

where the operator  $\operatorname{Re} \{ \cdot \}$  takes the real part of its argument. The real-operator is introduced to ensure that the approximated Kurtosis measure is a real valued function of  $\mathbf{w}^{[k]}[n]$ . The objective is now to maximize the approximated Kurtosis measure  $\hat{\kappa}^{[k]}[n]$  in (10) by continuously updating the filter  $\mathbf{w}^{[k]}[n]$ , using information in the previous filter vector  $\mathbf{w}^{[k]}[n-1]$ . The introduced approximation, using the *a-priori* output signal, was inspired by the derivation of the Projection Approximation Subspace Tracking (PAST) technique in [15].

### 3.2.1 Newton-based Kurtosis Maximization

The approximation of the beamformer's output signal Kurtosis value in (10) is (locally) quadratic in the filter vector  $\mathbf{w}^{[k]}[n]$ , and the optimization of this approximative Kurtosis value, according to a modified Recursive Least Squares (RLS) method [16], follows

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbf{P}^{[k]}[n] \Delta^{[k]}[n]}{\left\| \mathbf{w}^{[k]}[n-1] - \gamma^{[k]} \mathbf{P}^{[k]}[n] \Delta^{[k]}[n] \right\|_2}, \quad (11)$$

where

$$\Delta^{[k]}[n] = 2u^{[k]}[n]\mathbf{u}^{[k]}[n] + z^{[k]}[n]^*\mathbf{z}^{[k]}[n]. \quad (12)$$

The parameter  $\gamma^{[k]} \in [0, 1]$  is introduced in order to control the fluctuations in the filter weights due to the random input data. The normalization in (11) has been incorporated in order to avoid the trivial solution  $\mathbf{w}^{[k]}[n] = \mathbf{0}$ . The variables  $u^{[k]}[n]$ ,  $z^{[k]}[n]$ ,  $\mathbf{u}^{[k]}[n]$ , and  $\mathbf{z}^{[k]}[n]$  are herein implemented using first order auto-regressive averages to approximate the various statistical measures in (10) as

$$u^{[k]}[n] = \lambda^{[k]}u^{[k]}[n-1] + (1 - \lambda^{[k]}) \left| \bar{y}^{[k]}[n] \right|^2, \quad (13)$$

$$z^{[k]}[n] = \lambda^{[k]}z^{[k]}[n-1] + (1 - \lambda^{[k]}) \bar{y}^{[k]}[n]^2, \quad (14)$$

$$\mathbf{u}^{[k]}[n] = \lambda^{[k]}\mathbf{u}^{[k]}[n-1] + (1 - \lambda^{[k]}) \bar{\mathbf{x}}^{[k]}[n]\bar{y}^{[k]}[n]^*, \quad (15)$$

$$\mathbf{z}^{[k]}[n] = \lambda^{[k]}\mathbf{z}^{[k]}[n-1] + (1 - \lambda^{[k]}) \bar{\mathbf{x}}^{[k]}[n]\bar{y}^{[k]}[n], \quad (16)$$

where the parameter  $\lambda^{[k]} \in [0, 1]$  controls the convergence rate (and the source tracking performance) of the method. The matrix  $\mathbf{P}^{[k]}[n]$  is computed according to the matrix inversion lemma [16] as

$$\begin{aligned} \mathbf{t}^{[k]} &= \mathbf{P}^{[k]}[n-1]\bar{\mathbf{x}}^{[k]}[n], \\ \mathbf{P}^{[k]}[n] &= \lambda^{[k]-1}\mathbf{P}^{[k]}[n-1] \\ &\quad - \frac{|\bar{y}^{[k]}[n]|^2 \mathbf{t}^{[k]}\bar{\mathbf{x}}^{[k]}[n]^H \mathbf{P}^{[k]}[n-1]}{\lambda^{[k]2} + \lambda^{[k]} |\bar{y}^{[k]}[n]|^2 \bar{\mathbf{x}}^{[k]}[n]^H \mathbf{t}^{[k]}}, \end{aligned} \quad (17)$$

where  $\mathbf{t}^{[k]}$  is a temporary vector. A suitable initialization of this method is  $\mathbf{P}^{[k]}[0] = \mathbf{I}_M$ , where  $\mathbf{I}_M$  is the  $(M \times M)$  identity matrix,  $u^{[k]}[0] = z^{[k]}[0] = 0$ ,  $\mathbf{u}^{[k]}[0] = \mathbf{z}^{[k]}[0] = \mathbf{0}_{M \times 1}$ , and  $\mathbf{w}^{[k]}[0] = (1, 0, \dots, 0)^T$ .

## 4 Evaluation

The performance of the proposed method is analyzed using an off-line setting with real measured data and two microphones. This setting al-

lows comparison between the single-channel method, the blind adaptive beamformer, and the combined proposed structure.

#### 4.1 Evaluation Measures

A measure of the Signal to Interference Ratio (SIR) improvement and an objective measure that reflects the perceptual speech quality through the Perceptual Evaluation of Speech Quality (PESQ) [12] measure are used to evaluate the proposed approach. The filter weights at each iteration are stored and used for filtering the original convolved but unmixed source signals. This enables direct access to the evaluation measures. The SIR improvement performance measure, denoted  $P_{\text{SIR}}$ , is defined as

$$P_{\text{SIR}} = \frac{\text{Var} \{y_s[t]\} \text{Var} \{x_{v;0}[t]\}}{\text{Var} \{x_{s;0}[t]\} \text{Var} \{y_v[t]\}}, \quad (18)$$

where  $\text{Var} \{ \cdot \}$  denotes an estimator of variance,  $y_s[t]$  and  $y_v[t]$  represent the speech and noise components of the enhanced output signal, and, similarly, the signals  $x_{s;0}[t]$  and  $x_{v;0}[t]$  represent the speech and noise components of the first microphone signal. The first microphone is acting as a reference in the analysis.

The PESQ standard is an automated method for objective assessment of perceptual sound quality, and it uses a perceptual model of how sound quality is perceived by humans. The PESQ computes a perceptual model for a clean received reference speech signal  $x_{s;0}[t]$  and a perceptual model for the processed output speech component  $y_s[t]$ . The perceptual difference between the clean received speech signal and the processed speech signal is mapped on the Mean Opinion Score (MOS), which yields a value between one and five, where the score one indicates bad speech quality and the score five is used to indicate excellent speech quality.

## 4.2 System Configuration

The filterbank configuration used  $K = 64$  subbands, with two times oversampling. The prototype filter was designed using the window method with a Hamming window. The method parameters  $\lambda^{[k]}$ ,  $\gamma^{[k]}$ ,  $\alpha_{\text{fast}}^{[k]}$ , and  $\alpha_{\text{slow}}^{[k]}$  were set such that their integration times were 60 ms, 30 ms, 60 ms, and 2 s, respectively. The maximal allowed attenuation in the AGE method was  $G^{[k]} = 15$  dB (i.e.,  $10^{15/20}$ ). It should be noted that these parameter values were selected empirically, and the same values were used for all subbands. A future analysis should encompass the influence of various parameter values on the method's performance in order to find their optimal values.

## 4.3 Signal Configuration

Human speech (male and female) was sent through a loudspeaker and recorded using two microphones separated by 5 cm in an office room (reverberation time  $RT_{60} = 130$  ms) with sampling frequency 8 kHz. Previously recorded ferry engine noise and factory noise were subsequently emitted and recorded using the same setup. The speech signal is then mixed at various levels of SIR with each of the two interfering noise signals.

## 4.4 Estimated Processing Load

The estimated processing load<sup>1</sup> for a realtime Digital Signal Processor (DSP) implementation of the proposed method on an ADSP-21262 type DSP is provided in table 1. As can be seen from this analysis, the filterbanks (analysis and synthesis) together with the adaptive blind beamformer comprise the lion's share of the required processing load, and the additional noise reduction method requires merely 2.9 % of the

---

<sup>1</sup>The estimation of the processing load for the proposed method is performed in a simulation environment provided by the DSP manufacturer. The test-software is written in C language, where the compiler is set to operate at the highest optimization level.

<b>Program block</b>	<b>Est. proc. load</b>
Dual-channel analysis filter bank	25.4 %
Dual-channel adaptive blind beamformer	58.0 %
Single-channel noise reduction method	2.9 %
Single-channel synthesis filter bank	13.7 %

Table 1: Estimated processing load of a dual-channel implementation of the proposed method on an ADSP-21262 DSP. The total program package requires 13.5 % of the DSP's processing resources.

overall processing load. The total program package requires 13.5 % of the DSP's available processing resources.

#### 4.5 Evaluation Results

The evaluated performance of the single-channel AGE technique, the original adaptive blind beamformer, and the proposed combined structure are presented in Figure 1 and Figure 2 for the cases when speech is mixed with ferry engine noise and speech is mixed with factory noise. The results indicate that the combined system outperforms each of the individual systems with respect to SIR improvement. In some cases, the performance of the combined system also outperforms the linear addition of performance of each of the two subsystems. This indicates a successful symbiosis, where the spatial processor aids the temporal processor, and vice-versa. In addition, the perceptual speech degradation of the combined system never falls below 0.3 MOS-units in relation to the blind beamformer's MOS, and this further motivates the proposed solution.

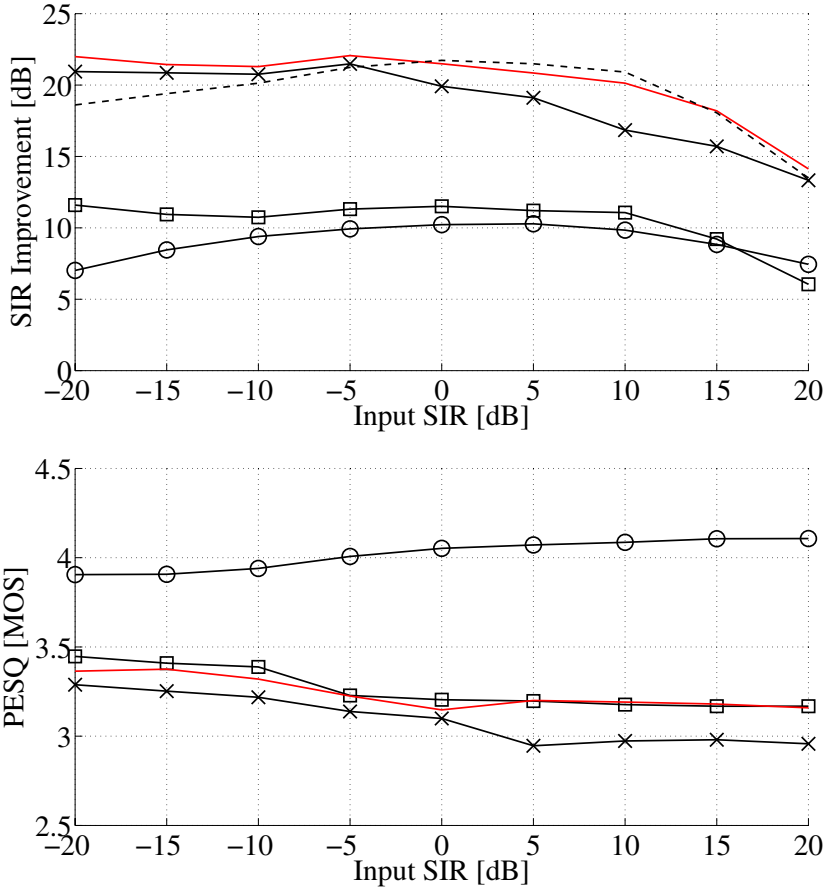


Figure 1: Evaluated performance when speech is mixed with ferry engine noise for the single-channel AGE method (circles), the adaptive blind beamformer (squares), and the proposed combined structure (crosses). The linear addition of SIR improvement of the AGE method and the adaptive blind beamformer (dashed). The SIR and PESQ improvements of the adaptive blind beamformer when the AGE method is applied to the beamformer's output signal (red).

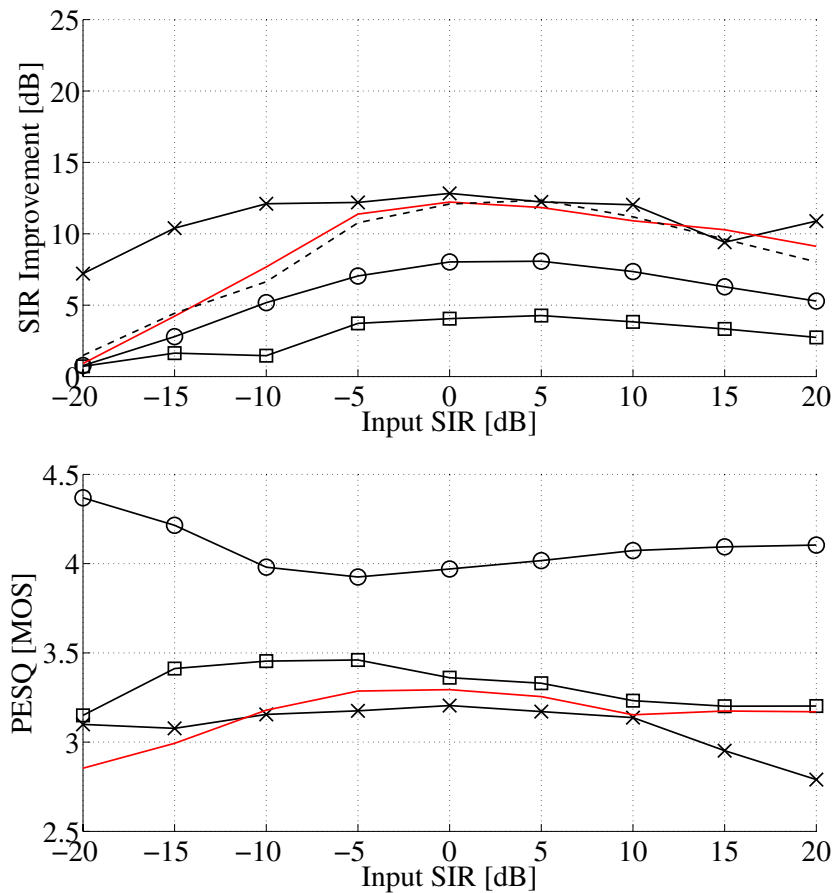


Figure 2: Evaluated performance when speech is mixed with factory noise for the single-channel AGE method (circles), the adaptive blind beamformer (squares), and the proposed combined structure (crosses). The linear addition of SIR improvement of the AGE method and the adaptive blind beamformer (dashed). The SIR and PESQ improvements of the adaptive blind beamformer when the AGE method is applied to the beamformer's output signal (red).

## 5 Summary and Conclusions

This paper presents the integration of a single-channel noise reduction technique in the feedback control loop of a recently proposed adaptive blind beamformer. The proposed combined system provides a SIR improvement that outperforms the individual systems. In some cases, the performance of the combined system also outperforms the linear addition of performance of each of the two subsystems. The introduced degradation in perceptual speech quality is tolerably low, and the extra processing load due to the extended structure is small; this further motivates the proposed combined structure. The current method parameters were selected empirically, and an important part for future research is the design of optimal parameter values that will further improve the method's performance. The proposed approach has been successfully validated in realtime using a DSP implementation with the purpose of blind speech extraction in high-noise human communication applications.

## References

- [1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(2):113–120, April 1979.
- [2] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement for personal communication using an adaptive gain equalizer. *Elsevier Signal Processing*, 85(6):1089–1101, 2005.
- [3] B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson. A mixed analog - digital hybrid for speech enhancement purposes. *IEEE International Symposium on Circuits and Systems*, 2:852–855, May 2005.
- [4] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [5] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [6] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [7] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [8] Z. Ding. A new algorithm for automatic beamforming. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2:689–693, November 1991.
- [9] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of sub-band kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE 15th International Conference on Digital Signal Processing*, pages 603–606, July 2007.

- [10] B. Sällberg, N. Grbić, and I. Claesson. Online blind speech extraction based on a locally quadratic kurtosis criteria and a preprocessing automatic gain controller. *IEEE 49th International Symposium ELMAR*, pages 139–142, September 2007.
- [11] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.
- [12] ITU-T p.862. *Perceptual evaluation of speech quality (PESQ)*.
- [13] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [14] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [15] B. Yang. Projection approximation subspace tracking. *IEEE Transactions on Signal Processing*, 43(1):95–107, January 1995.
- [16] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.





## PART V

# Online Blind Speech Extraction Based on a Local Quadratic Kur- tosis Criterion and a Preprocess- ing Automatic Gain Controller

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson, “Online Blind Speech Extraction Based on a Locally Quadratic Kurtosis Criteria and a Preprocessing Automatic Gain Controller”, *49th International Symposium ELMAR-2007 focused on Multimedia Signal Processing and Communications*, 2007.

**Modification to the original paper:**

The notations have been standardized so as to fit the other parts of this thesis.

# Online Blind Speech Extraction Based on a Local Quadratic Kurtosis Criterion and a Preprocessing Automatic Gain Controller

Benny Sällberg, Nedelko Grbić, and Ingvar Claesson

## **Abstract**

This paper focuses on realtime speech extraction using blind adaptive beamforming. The speech extraction is carried out using an approximation of the Kurtosis measure in a subband domain. The introduced approximated Kurtosis measure is an improvement over a recently proposed approximation technique where a local quadratic criterion was solved at each iteration. The improvement introduced in this paper regards an approach to normalize this same criterion using a preprocessing Automatic Gain Control (AGC) unit and thereby making the algorithm invariant to input signal scales. The proposed method outperforms the recent technique in terms of Signal to Interference Ratio (SIR) improvement. In addition, the increased memory consumption and processing load due to the proposed improvement are comparably low, and this is often desirable in a realtime Digital Signal Processor (DSP) implementation. Further, a realtime implementation of the method is conducted and results with real data are presented.

## 1 Introduction

Blind adaptive beamforming has features that are attractive for speech enhancement in human communication. The motivation for employing a beamformer is that it uses several microphones, thus operating in the spatiotemporal domain, and it has a higher degree of freedom as opposed to corresponding single-channel techniques that only utilize the temporal domain [1]. The inherent virtue of a blind control algorithm in beamforming is that no knowledge about the spatiotemporal environment is needed, such as the position of the sources relative to the microphone array or knowledge regarding the physical dimension of the array itself. The merging of an adaptive beamformer with a blind control algorithm results in a structure that continuously tracks sources in a changing environment [2, 3].

This paper presents an improvement to a recently proposed approach of approximating the subband Kurtosis measure using a local quadratic criterion [3]. The improvement proposed in this paper regards a preprocessing Automatic Gain Control (AGC) stage that normalizes the input signals prior to the adaptive beamformer's filter update equation, thus yielding a scale-invariant solution. The intended application is Blind Speech Extraction (BSE), where a dominant speech source (dominant in the Kurtosis measure) is extracted from an observed convolutive mixture of sources [3, 4, 5, 6]. The proposed approach stands in contrast to Blind Signal Separation (BSS) in which all dominant sources, or groups of sources, are separated, see [7].

The outline of this paper is as follows. The assumed signal model and the beamforming notation are given in Section 2. The improved approximated Kurtosis measure is provided in Section 3, and a Newton-based maximization thereof is given in Section 4. A realtime Digital Signal Processor (DSP) implementation is presented in Section 5, to illustrate the method's robustness and performance. Evaluation results are given in Section 6. Conclusions and topics for future work are provided in Section 7.

## 2 Signal Model

In this paper, we assume one dominant desired speech source and  $I - 1$  undesired sources (with lower Kurtosis values). The sources' relative positions to the beamformer are unknown, and the beamformer's spatial configuration is also unknown. The beamformer employs  $M$  microphones that sense the acoustical wavefield, and the received time signals for each microphone  $m : m \in \mathbb{N}, m < M$  at time index  $t$  are denoted  $x_m[t]$ . The set of all microphone signals is represented by the vector  $\mathbf{x}[t] = (x_0[t], x_1[t], \dots, x_{M-1}[t])^T$ , where the superscript  $(\ )^T$  denotes the transpose. The received time signals are efficiently decomposed into a time-frequency representation, denoted  $\mathbf{x}^{[k]}[n] = (x_0^{[k]}[n], x_1^{[k]}[n], \dots, x_{M-1}^{[k]}[n])^T$ , where  $k : k \in \mathbb{N}, k < K$  is the subband index and  $n$  is the block time index, using a polyphase realization of a Discrete Fourier Transform (DFT) modulated analysis filterbank [8]. The convolutive mixture in the time domain corresponds to instantaneous mixtures in the frequency domain [6], and the observed subband signal for subband index  $k$  is assumed to be

$$\mathbf{x}^{[k]}[n] = \mathbf{H}^{[k]}[n]\mathbf{s}^{[k]}[n] + \mathbf{v}^{[k]}[n], \quad (1)$$

where  $\mathbf{H}^{[k]}[n]$  represents a matrix of channels, the vector  $\mathbf{s}^{[k]}[n]$  contains the subband source signals, and  $\mathbf{v}^{[k]}[n]$  represents the subband noise component. The desired speech source's signal is the first signal  $s_0^{[k]}[n]$  in  $\mathbf{s}^{[k]}[n]$ , and the other signals  $s_i^{[k]}[n]$  ( $i \in \{1, 2, \dots, I - 1\}$ ) in  $\mathbf{s}^{[k]}[n]$  are considered as interferences. A linear weighting of this subband input signal using a time-varying beamformer filter vector  $\mathbf{w}^{[k]}[n] = (w_0^{[k]}[n], w_1^{[k]}[n], \dots, w_{M-1}^{[k]}[n])^T$ , denoted a filter-and-sum beamformer [1], yields a subband output signal

$$y^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}^{[k]}[n] \quad (2)$$

where the superscript  $(\ )^H$  denotes the Hermitian transpose. The time domain output signal  $y[t]$  is reconstructed from the subband output

signals  $y^{[k]}[n]$  using a polyphase DFT modulated synthesis filterbank matched to the analysis filterbank [8].

### 3 Approximation of Subband Kurtosis Measure

A listing of 16 different variants of the Kurtosis measure for complex valued data is given in [9]. One of these variants of the Kurtosis measure was applied in [3] for the beamformer's subband output signal

$$\kappa \{y^{[k]}[n]\} = E \left\{ \left| y^{[k]}[n] \right|^4 \right\} - 2E \left\{ \left| y^{[k]}[n] \right|^2 \right\}^2 - \left| E \left\{ y^{[k]}[n]^2 \right\} \right|^2, \quad (3)$$

where  $E \{ \}$  represents the expectation operator and  $\kappa \{y^{[k]}[n]\}$  designates the Kurtosis value of the signal  $y^{[k]}[n]$ .

#### 3.1 Improvement - Scale Invariance

While a linear scale  $c$  applied to the beamformer's input signal  $c\mathbf{x}^{[k]}[n]$  yields a non-linear change in the beamformer output signal's Kurtosis value  $c^4\kappa \{y^{[k]}[n]\}$ , it may be concluded that the Kurtosis measure in (3) is not invariant to input signal scale. This scale variant behavior is in many cases undesirable, because it influences the predictability of the algorithm's behavior in a realtime implementation. An Automatic Gain Control (AGC) unit is proposed in this paper as a remedy to this shortfall. The AGC is applied to the input signal before the adaptive beamformer's filter update equation. It is stressed that the AGC is not used in the filtering part of the algorithm and does thereby not influence the output signal directly, as indicated in Figure 1. The AGC normalizes the input data, denoted  $\mathbf{x}'^{[k]}[n]$ , by the square root of its mean power

$$\mathbf{x}'^{[k]}[n] = \frac{\mathbf{x}^{[k]}[n]}{\sqrt{E \left\{ \mathbf{x}^{[k]}[n]^H \mathbf{x}^{[k]}[n] \right\}}}. \quad (4)$$

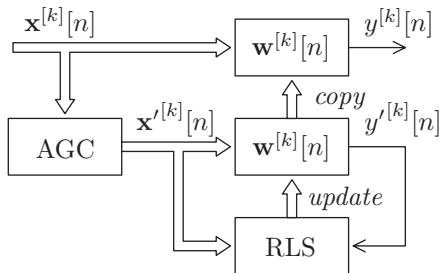


Figure 1: An Automatic Gain Control (AGC) in the control loop of an adaptive beamformer.

According to [3] a normalized output signal  $y'^{[k]}[n] = \mathbf{w}^{[k]}[n]^H \mathbf{x}'^{[k]}[n]$  and a normalized *a-priori* output signal  $\tilde{y}'^{[k]}[n] = \mathbf{w}^{[k]}[n-1]^H \mathbf{x}'^{[k]}[n]$  are used in a scale invariant Kurtosis measure,  $\hat{\kappa}'^{[k]}[n]$ , according to

$$\begin{aligned} \hat{\kappa}'^{[k]}[n] &= \text{E} \left\{ \left| y'^{[k]}[n] \right|^2 \left| \tilde{y}'^{[k]}[n] \right|^2 \right\} \\ &\quad - 2\text{E} \left\{ \left| \tilde{y}'^{[k]}[n] \right|^2 \right\} \text{Re} \left\{ \text{E} \left\{ y'^{[k]}[n] \tilde{y}'^{[k]}[n]^* \right\} \right\} \\ &\quad - \text{Re} \left\{ \text{E} \left\{ \tilde{y}'^{[k]}[n]^2 \right\}^* \text{E} \left\{ y'^{[k]}[n] \tilde{y}'^{[k]}[n] \right\} \right\}, \end{aligned} \quad (5)$$

where the operator  $\text{Re} \{ \}$  takes the real part of its argument and  $( \ )^*$  designates the complex conjugate. The real-operator is introduced to ensure that the approximated Kurtosis measure is real valued and that it forms an analytic function of  $\mathbf{w}^{[k]}[n]$ . The objective is now to maximize the approximated Kurtosis measure  $\hat{\kappa}'^{[k]}[n]$  in (5) by continuously updating the filter  $\mathbf{w}^{[k]}[n]$ , using information in the previous filter vector  $\mathbf{w}^{[k]}[n-1]$ .

## 4 Newton-based Kurtosis Maximization

The approximation of the beamformer output signal's Kurtosis value in (5) is (locally) quadratic in the filter vector  $\mathbf{w}^{[k]}[n]$ . Optimization of the approximative Kurtosis value follows Newton's method [10]. A feasible implementation of the method is achieved through the introduction of a set of suitable approximations. Auto-regressive averaging is one type of approximation (see [4, 5]) that are used here

$$\mathbf{P}^{[k]}[n] \approx \mathbb{E} \left\{ \mathbf{x}'^{[k]}[n] \mathbf{x}'^{[k]}[n]^H \left| \tilde{y}'^{[k]}[n] \right|^2 \right\}^{-1}, \quad (6)$$

$$u^{[k]}[n] \approx \mathbb{E} \left\{ \left| \tilde{y}'^{[k]}[n] \right|^2 \right\}, \quad (7)$$

$$z^{[k]}[n] \approx \mathbb{E} \left\{ \tilde{y}'^{[k]}[n]^2 \right\}, \quad (8)$$

$$\mathbf{u}^{[k]}[n] \approx \mathbb{E} \left\{ \mathbf{x}'^{[k]}[n] \tilde{y}'^{[k]}[n]^* \right\}, \quad (9)$$

$$\mathbf{z}^{[k]}[n] \approx \mathbb{E} \left\{ \mathbf{x}'^{[k]}[n] \tilde{y}'^{[k]}[n] \right\}, \quad (10)$$

$$\eta^{[k]}[n] \approx \mathbb{E} \left\{ \mathbf{x}'^{[k]}[n]^H \mathbf{x}'^{[k]}[n] \right\}. \quad (11)$$

The matrix  $\mathbf{P}^{[k]}[n]$  in (6) is recursively updated according to the matrix inversion lemma as [10]

$$\begin{aligned} \mathbf{t}^{[k]} &= \mathbf{P}^{[k]}[n-1] \mathbf{x}'^{[k]}[n], \\ \mathbf{P}^{[k]}[n] &= \lambda^{[k]-1} \mathbf{P}^{[k]}[n-1] \\ &\quad - \frac{\left| \tilde{y}'^{[k]}[n] \right|^2 \mathbf{t}^{[k]} \mathbf{x}'^{[k]}[n]^H \mathbf{P}^{[k]}[n-1]}{\lambda^{[k]2} + \lambda^{[k]} \left| \tilde{y}'^{[k]}[n] \right|^2 \mathbf{x}'^{[k]}[n]^H \mathbf{t}^{[k]}}, \end{aligned} \quad (12)$$

where  $\mathbf{t}^{[k]}$  is a temporary vector, and the parameter  $\lambda^{[k]} \in [0, 1]$  controls the convergence rate (i.e., the tracking performance) of the method.

First order auto-regressive averages are used in (7) to (11):

$$u^{[k]}[n] = \lambda^{[k]}u^{[k]}[n-1] + \left(1 - \lambda^{[k]}\right) \left| \tilde{y}'^{[k]}[n] \right|^2, \quad (13)$$

$$z^{[k]}[n] = \lambda^{[k]}z^{[k]}[n-1] + \left(1 - \lambda^{[k]}\right) \tilde{y}'^{[k]}[n]^2, \quad (14)$$

$$\mathbf{u}^{[k]}[n] = \lambda^{[k]}\mathbf{u}^{[k]}[n-1] + \left(1 - \lambda^{[k]}\right) \mathbf{x}'^{[k]}[n] \tilde{y}'^{[k]}[n]^*, \quad (15)$$

$$\mathbf{z}^{[k]}[n] = \lambda^{[k]}\mathbf{z}^{[k]}[n-1] + \left(1 - \lambda^{[k]}\right) \mathbf{x}'^{[k]}[n] \tilde{y}'^{[k]}[n], \quad (16)$$

$$\eta^{[k]}[n] = \lambda^{[k]}\eta^{[k]}[n-1] + \left(1 - \lambda^{[k]}\right) \mathbf{x}^{[k]}[n]^H \mathbf{x}^{[k]}[n]. \quad (17)$$

The input power estimate in (17) is used to normalize the input signal  $\mathbf{x}'^{[k]}[n] = \frac{\mathbf{x}^{[k]}[n]}{\sqrt{\eta^{[k]}[n]}}$ . The filter update equation, according to Newton's method (with an additional normalization step), becomes

$$\mathbf{w}^{[k]}[n] = \frac{\mathbf{w}^{[k]}[n-1] - \gamma^{[k]}\mathbf{P}^{[k]}[n]\mathbf{\Delta}^{[k]}[n]}{\left\| \mathbf{w}^{[k]}[n-1] - \gamma^{[k]}\mathbf{P}^{[k]}[n]\mathbf{\Delta}^{[k]}[n] \right\|_2}, \quad (18)$$

where

$$\mathbf{\Delta}_k(n) = 2u^{[k]}[n]\mathbf{u}^{[k]}[n] + z^{[k]}[n]^* \mathbf{z}^{[k]}[n]. \quad (19)$$

The parameter  $\gamma^{[k]} \in [0, 1]$  is introduced in order to control the fluctuations in the filter weights due to the random input data. The normalization in (18) has been incorporated in order to avoid the trivial solution  $\mathbf{w}^{[k]}[n] = \mathbf{0}$ . However, this normalization is not optimal with respect to the spectral whitening of the enhanced speech since only  $\left\| \mathbf{w}^{[k]}[n] \right\|_2 = 1$  is ensured and the term  $\left\| \mathbf{w}^{[k]}[n]^H \mathbf{h}_0^{[k]}[n] \right\|_2$  is thereby not guaranteed to equal unity ( $\mathbf{h}_0^{[k]}[n]$  is the transfer function vector related to the desired speech source). It may be noted that, except for the normalization, the filter update in (18) is similar to the well-known Recursive Least Squares (RLS) algorithm [10]. A suitable initialization of this algorithm is  $\mathbf{P}^{[k]}[0] = \mathbf{I}_M$ , where  $\mathbf{I}_M$  is the  $(M \times M)$  identity matrix,  $u^{[k]}[0] = z^{[k]}[0] = 0$ ,  $\mathbf{u}^{[k]}[0] = \mathbf{z}^{[k]}[0] = \mathbf{0}_{M \times 1}$ , and  $\mathbf{w}^{[k]}[0] = (1, 0, \dots, 0)^T$ .

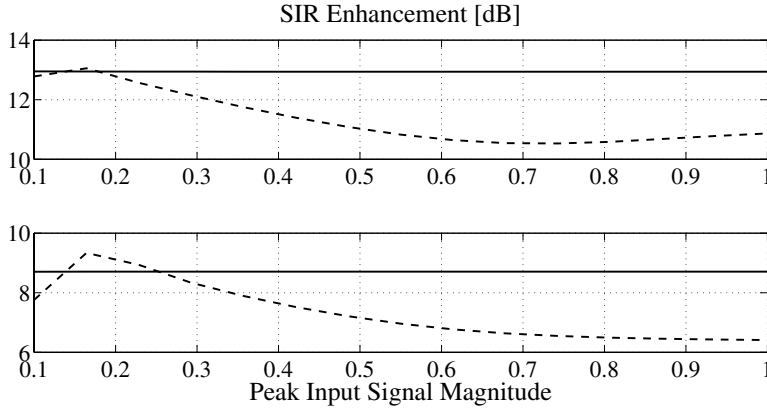


Figure 2: Mean SIR enhancement at different peak input signal magnitudes for speech and ferry engine noise signal (upper) and speech and factory noise signal (lower). The improved scale-invariant algorithm (solid) and the original algorithm [3] (dashed).

## 5 Realtime Implementation

The implementation of the proposed method is made on a floating point DSP named ADSP-21262 from Analog Devices. This is a high-performance DSP that supports effective parallel computations through the Single Instruction Multiple Data (SIMD) mode, suitable for vector-based operations. The algorithm is efficiently implemented using a transformed approach to reduce the overhead. A rule-of-thumb for the transformed approach is to perform all computations so that the largest data dimension is used in the element-wise vector operations. The transformation approach reduces the number of computations in this DSP platform by 7.5 times for a typical vector operation [11]. The same hardware platform was used in [3] to realize the predecessor to the proposed blind beamformer.

## 5.1 System Configuration

The filterbank configuration used  $K = 64$  subbands, with four times oversampling. The prototype filter was designed using the window method with a Hamming window. This configuration was selected as a trade-off between the audio quality performance and the introduced signal delay, measured to 8.5 ms. The algorithmic parameters were selected such that the integration time of the  $\lambda^{[k]}$  parameter was 0.46 s and the integration time of the  $\gamma^{[k]}$  parameter was 5 ms.

## 6 Evaluation Results

The performance of the improved proposed method is first analyzed using an offline setting with real measured data. This setting allows for comparison with the recently proposed method [3]. The realtime implementation is analyzed using short-term power estimates where the algorithm may be switched on and off.

### 6.1 Evaluation Measures

The mean Signal to Interference Ratio (SIR) enhancement  $Q$  is analyzed in the time domain using

$$Q = \frac{\text{Var}\{y_s[t]\} \text{Var}\{x_v[t]\}}{\text{Var}\{y_v[t]\} \text{Var}\{x_s[t]\}}, \quad (20)$$

where  $\text{Var}\{\cdot\}$  designates an estimator of variance and the signals  $x_s[t]$ ,  $x_v[t]$ ,  $y_s[t]$ , and  $y_v[t]$  represent the components of the input and output signals that are related to the speech and interference, respectively.

### 6.2 Offline Results using Real Data

Human speech (male and female) was sent through a loudspeaker and recorded using two microphones separated by 5 cm in an office room (reverberation time  $RT_{60} = 130$  ms) with sampling frequency 8 kHz.

A previously recorded ferry engine noise signal and a factory noise signal were subsequently emitted and recorded using the same setup, although from a different direction. The speech signal was then mixed at 0 dB SIR level with each of the two interfering noise signals. Each mixed signal was thereafter scaled such that the peak input signal magnitude corresponded to a certain level in order to test the algorithm's ability to extract speech in different operating conditions. The mean SIR enhancement for various input signal magnitudes is presented in Figure 2. It can be concluded from this figure that the proposed algorithm is invariant to different input signal magnitudes as opposed to the original algorithm [3]. Furthermore, the performance of the proposed algorithm exceeds that of the original method in most cases. Some spectral whitening can be noticed in the extracted speech signal.

### 6.3 Online Results using Real Data

The online evaluation includes prerecorded signals consisting of male speech spatially added with music. It is provided in [12]. The speaker and the music were recorded in an office room using two microphones where the distance between the sources and the microphones is 60 cm in a square ordering. The two input channels are used as input to the proposed realtime DSP implementation, and the resulting output time signal is presented in Figure 3 with a corresponding short-term power estimate. The proposed method is activated after 6.8 s. The two input channels are exchanged momentarily at 13.7 s in order to illustrate the method's capability to track signals in a non-stationary environment. The SIR enhancement is 15 dB to 20 dB in this case.

## 7 Conclusions and Future Work

This paper presents an improvement to a blind beamformer based on the well-known Kurtosis maximization strategy. The novelty of this contribution lies in a normalization approach of the recently proposed

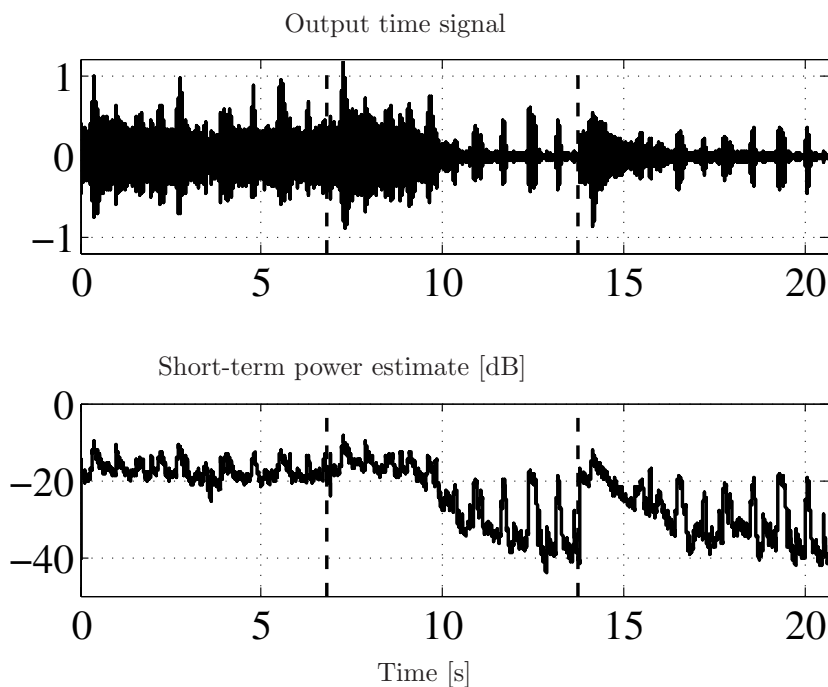


Figure 3: Output time signal for speech and music (upper) and short-term power estimate (lower). The proposed method is activated at 6.8 s, and the input channels are exchanged at 13.7 s.

(locally) quadratic Kurtosis criterion [3]. The improved criterion is continuously updated online according to a Newton-based search method. The method successfully extracts speech in convolutive mixtures of speech and ferry engine noise, speech and factory noise, and speech and music. It is noticeable that, except for some spectral whitening, the speech distortion is perceptually very low. Furthermore, the performance of the proposed method is superior to the recently proposed method for a variety of input signal scales. This method shows promising performance for speech enhancement in real environments. How-

ever, some questions are still open for future research:

The strategy for normalization of the adaptive filter used here, (18), is not optimal with respect to the spectral whitening of the enhanced speech. Other normalization strategies may reduce this undesired spectral whitening.

A rigorous analysis of the spatiotemporal behavior of this proposed method is required in order to relate and compare its performance to existing state-of-the art solutions.

## References

- [1] D. Johnson and D. Dudgeon. *Array Signal Processing – Concepts and Techniques*. Prentice Hall, 1993.
- [2] Z. Ding. A new algorithm for automatic beamforming. *IEEE Asilomar Conference on Signals, Systems and Computers*, 2:689–693, November 1991.
- [3] B. Sällberg, N. Grbić, and I. Claesson. Online maximization of sub-band kurtosis for blind adaptive beamforming in realtime speech extraction. *IEEE 15th International Conference on Digital Signal Processing*, pages 603–606, July 2007.
- [4] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. John Wiley and Sons, 2001.
- [5] A. Cichocki and S. Amari. *Adaptive Blind Signal and Image Processing - Learning Algorithms and Applications*. John Wiley and Sons, 2003.
- [6] P. Smaragdis. Blind separation of convolved mixtures in the frequency domain. *Elsevier Neurocomputing*, 22(1–3):21–34, 1998.
- [7] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson. Blind signal separation using overcomplete subband representation. *IEEE*

*Transactions on Speech and Audio Processing*, 9(5):524–533, July 2001.

- [8] P. P. Vaidyanathan. *Multirate Systems and Filter Banks*. Prentice Hall, 1993.
- [9] C. Nikias and A. Petropulu. *Higher-Order Spectral Analysis - A Nonlinear Signal Processing Framework*. Prentice Hall, 1993.
- [10] S. Haykin. *Adaptive Filter Theory*. John Wiley and Sons, 2002.
- [11] Z. Yermeche, B. Sällberg, N. Grbić, and I. Claesson. Real-time dsp implementation of a subband beamforming algorithm for dual microphone speech enhancement. *IEEE International Symposium on Circuits and Systems*, pages 353–356, May 2007.
- [12] T. W. Lee. Blind source separation: Audio examples. [Online] [http://www.cnl.salk.edu/~tewon/Blind/blind\\_audio.html](http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html), Feb. 17 2007.



## PART VI

# Implementation Aspects of the Adaptive Gain Equalizer

**This part is published as:**

B. Sällberg, N. Grbić, and I. Claesson, *Implementation Aspects of the Adaptive Gain Equalizer*, Research Report, Blekinge Institute of Technology, ISSN: 1103-1581, May, 2006.

Parts of this research report have been published as:

B. Sällberg and M. Dahl, *Speech Enhancement Implementations in the Digital, Analog and Hybrid Domain*, SSoCC - Swedish System on Chip Conference, Stockholm, Sweden, 18-19 April, 2005.

B. Sällberg, H. Åkesson, M. Dahl and I. Claesson, *A Mixed Analog-Digital Hybrid for Speech Enhancement Purposes*, IEEE ISCAS - International Symposium for Circuits and Systems, Kobe, Japan, 23-26 May, 2005.

B. Sällberg, H. Åkesson, M. Dahl and I. Claesson, *Analog Circuit Implementation for Speech Enhancement Purposes*, IEEE Asilomar Conference on Signals, Systems, and Computers, Pacific Grove/CA, USA, 7-10 November, 2004.

# Implementation Aspects of the Adaptive Gain Equalizer

Benny Sällberg and Nedelko Grbić and Ingvar Claesson

## **Abstract**

The quality of speech, or important speech parameters such as the intelligibility, clearness or naturalness of speech, can be emphasized by signal processing. Such processing for improving speech quality can be found in telecommunication applications, e.g. mobile telephony, internet telephony or personal intercom. Blind methods are preferable over conventional because they do not require calibration schemes and are independent of environmental variations. By careful selection of hardware domain for realization, i.e. digital, analog, or hybrid, implementation-specific benefits can be utilized to increase the speech quality or performance. This report stresses some implementation aspects when implementing a blind method for speech enhancement in digital, analog, and hybrid digital-analog hardware.

## **Acknowledgement**

The authors wish to thank Mattias Dahl and Nils Westerlund for rigorous groundwork testing the algorithm capabilities and Henrik Åkesson for constructive feedback and assistance during the implementations outlined in this report.

## 1 Introduction

The objective of speech signal processing is to improve the overall quality, or selected qualitative measures, of speech. A typical application is in telecommunication, where the perception of a human speech communication can be improved by speech signal processing, see Figure 1. The spectral subtraction method is a classic example of an algorithm

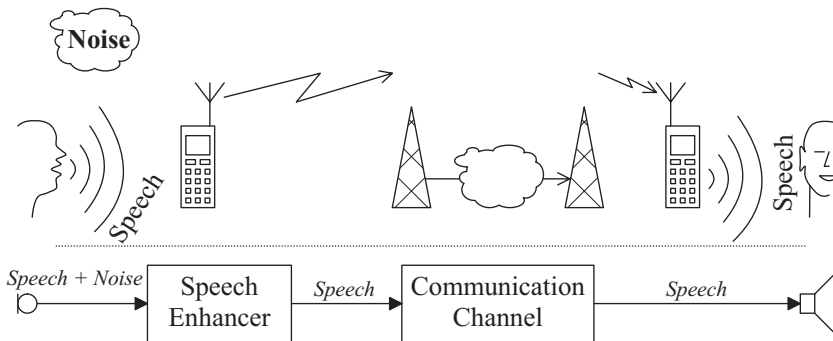


Figure 1: Human speech communication over a radio link.

for increasing the speech Signal to Noise Ratio (SNR) by reducing the level of interfering noise [1, 2]. Digital or analog hardware can be used for realizing speech enhancement algorithms [3]. Some specific algorithms are also suitable for a hybrid (mix between analog and digital) domain implementation [4]. The choice of implementation domain and the specific characteristics thereof may be intentionally utilized to increase the overall performance and efficiency. In this context, performance implies not only implementation specific performance, e.g. power consumption, but also qualitative speech performance, e.g. naturalness or intelligibility [5, 6]. However, the choice of implementation domain may lead to restrictions on the signal processing algorithm.

This report discusses selected advantages (and disadvantages) when implementing a robust, low-complexity speech enhancement algorithm (see [7, 8, 9, 10]) in various hardware domains. The report reflects

experience gained by the authors during the implementations and is a collection and extension of material provided in [11, 12, 13].

The outline of this report is as follows:

**Section 2** A general discussion of implementation aspects is provided in this section. The discussion puts emphasis on speech enhancement related issues.

**Section 3** A speech enhancer is employed for implementation in various domains. The Adaptive Gain Equalizer (AGE) is the selected speech enhancer and is outlined in this section, in its original digital form.

**Section 4** Aspects of implementing the AGE in the digital, analog, and hybrid analog-digital domain are presented in this section.

**Section 5** This section acts as a proof of concept, where the conducted evaluation indicates that the AGE carries the robustness and flexibility required for implementation in various domains.

**Section 6** A short summary and some conclusions drawn from the work collected in this report is presented in this section.

## 2 A Discussion of Implementation Aspects

Two major design approaches exist for the implementation of an algorithm in hardware; A speech enhancement algorithm is given with the objective to implement it in a specific domain. Alternatively, the implementation domain is outlined in advance and the objective is to find a speech enhancement algorithm that fits the given constraints. Independent of approach, the solution must consider the requirements of the selected speech enhancement algorithm with respect to the hardware, e.g. constraints on signal delay, speech signal quality and real time performance.

Digital and analog hardware implementations of speech enhancement methods exhibit both advantages and disadvantages. With a hybrid solution main benefits of the two domains may be utilized, while drawbacks can be circumvented to some extent. This section provides a general discussion regarding some aspects of implementing speech enhancement structures in hardware. The underlying configuration for the various domains is illustrated in Figure 2.

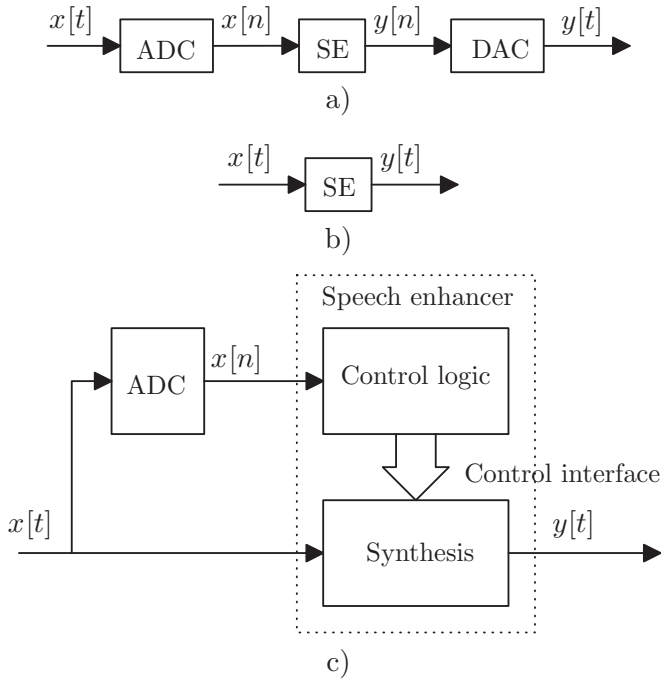


Figure 2: Implementation of a Speech Enhancer (SE) in the digital domain (a), the analog domain (b), and the hybrid domain (c). Analog-to-Digital Conversion (ADC) and Digital-to-Analog Conversion (DAC) includes anti-alias and reconstruction filtering. The signal  $x[t]$  include speech and noise, and  $y[t]$  contains enhanced speech. Continuous time is denoted by  $t$  and the sampling index is denoted by  $n$ .

## 2.1 The Digital Domain

Digital domain refers to the use of digital processors or embedded systems, such as Digital Signal Processors (DSP), Micro Controllers ( $\mu\text{C}$ ) or digital Application Specific Integrated Circuits (ASIC). The requirement to use an Analog to Digital Converter (ADC) and a Digital to Analog Converter (DAC) is common to any digital solution interfacing with the real world. This requirement is due to the sampling process, see Figure 2 a).

### 2.1.1 Advantages

The main advantage of digital solutions is their high degree of software configurability for programmable units. The implementation problem is normally well defined in the digital domain. Digital implementations may also be easily adjusted to fit the specific environment or given hardware capabilities. Mathematically advanced and complex algorithms and structures can be realized within a digital processor. A digital solution can most often also perform several consecutive tasks, sometimes even in parallel, e.g. both noise reduction and speech coding. The possibility to implement filters with a linear phase property is a vital advantage of the digital domain.

### 2.1.2 Disadvantages

When digital solutions are designed, some disadvantages of digital domain implementation need to be taken into account. For example, algorithms can be limited by processor clock rate (for synchronous systems), word length, type and number of on-chip peripherals. In the worst case scenario, the computational load is too high and introduces timing problems resulting in poor speech quality. Limitations in word length can introduce errors if not considered in the design phase. For example, there is a significant difference in short-multiplications (16 bit) and long multiplications (32 bit). The differences in fixed point and floating point arithmetic also require special attention. The sampling

process in digital solutions requires good dynamic range utilization, e.g. additional circuitry such as expanders or automatic gain control units may be employed. A poor dynamic range utilization may lead to inadequate speech quality. Digital solutions often introduce delays in the signal path due to analog anti-alias filtering in the sampling process, Analog to Digital Converter (ADC) schemes such as sigma-delta may also contribute negatively to the overall signal delay. Another factor included in digital domain disadvantages is clock system power consumption. Up to one third of the total power dissipation of digital circuitry lies in the power consumption of the clocking network [14].

## 2.2 The Analog Domain

Analog solutions utilize passive components or discrete semi-conductors, e.g. resistors, capacitors, inductors, transistors and diodes, to perform specific tasks. However, operational amplifiers and multipliers/dividers may be employed to implement more sophisticated methods. Analog solutions do not require signal sampling and operate directly on the received analog signals, see Figure 2 b).

### 2.2.1 Advantages

Similar to the digital domain, the analog domain has some key features worth mentioning. Data in an analog solution is not quantized, and often it is less restricted in bandwidth compared to a digital solution. For a speech signal processing application, the high bandwidth and lack of quantization of data may lead to very high quality of speech. The implementation is not restricted by clock rates or word length related issues, since the "operations" are performed in continuous time. Due to the continuous time signal processing, the group delay introduced by the analog domain is likely to be extremely short as opposed to corresponding digital structures. It is also likely that an analog solution is more power efficient than a corresponding digital solution while it does not in general require an inefficient clocking network.

### 2.2.2 Disadvantages

A set of bottlenecks inhibit the usability of analog solutions for signal processing implementations. Some mathematical operations can be hard to implement using pure analog hardware. Workarounds may include inexact approximations which introduce errors, e.g. bias, offset etcetera. Nonlinear phenomena are reoccurring in analog solutions which can make complex implementations hard to predict and to simulate, e.g. diodes and transistors are nonlinear by nature. The implementation problem is harder to define in comparison to digital solutions, while voltages in analog solutions are continuous and bound mainly by the supply voltages. Analog solutions may also be sensitive to variations in component values, and component ageing, leading to unpredictable results if neglected. In all, the implementation of an analog solution often requires significant engineering skills and hands-on experience.

## 2.3 The Hybrid Domain

By definition, a solution in the hybrid domain incorporates a mixture of digital and analog hardware. Key features of the two domains should be utilized in the design of hybrid solutions, whilst trying to eliminate the drawbacks of each domain. An advanced algorithm may, for example, be split into several parts in a hybrid solution, e.g. by putting computationally immense tasks in the digital domain and simpler tasks in the analog domain. For speech signal processing applications one could put the control logic in digital hardware and the actual signal processing in analog hardware, see Figure 2 c). To illustrate the outstanding performance achievements in hybrid solutions: When speech is not present, the control logic can be put in sleep mode (low power consumption) to conserve energy. In a hybrid approach, the overall solution is likely to be highly robust, while the two domains (analog and digital) may complement each other. For example, even though digital control logic is suffering from dynamic range related issues, the actual analog signal

processing is still producing high fidelity speech. However, designing a hybrid system requires special attention to ensure that analog and digital sections do not interfere with each other. Digital interference in analog audio signals may contribute negatively to the overall quality of speech. Utilization of separate ground planes and separate power supply lines for digital and analog hardware is a rule of thumb to achieve high fidelity speech quality.

### **3 A Speech Enhancer**

The Adaptive Gain Equalizer (AGE) has been shown to be a highly effective method for the enhancement of speech [7, 8, 9, 10]. Low complexity and high flexibility makes the method suitable for a wide range of implementations [11, 12, 13]. Furthermore, the AGE is scalable and does not require a Voice Activity Detector (VAD), as opposed to similar methods such as the spectral subtraction method. Here, the scalability of the AGE implies that the underlying structure is the same, independent of the number of subbands.

#### **3.1 The Adaptive Gain Equalizer**

The AGE may be viewed as an intelligent volume control, in which the volume is rapidly boosted when speech is present. Hence, the method focuses on boosting speech rather than on suppression of noise. One fundamental assumption constitutes the foundation of the AGE, namely; the stationarity time for speech is significantly lower than that of the interfering noise [15]. The method has been verified using traditional DSP technology [8], a mixed signal processor, analog hardware [11], and hybrid analog-digital hardware [12]. However, the original formulation of the AGE is in the digital domain.

### 3.1.1 Input-Output Signal Assembly

An analysis filter bank is employed for division of the sampled input signal,  $x[n]$ , into frequency selective subbands<sup>1</sup>,  $x^{[k]}[n]$ . Each input subband signal is weighted by a gain function,  $g^{[k]}[n]$ . Finally, all weighted subband signals are combined to form the total output,  $y[n]$ , according to

$$x^{[k]}[n] = h^{[k]} * x[n], \quad (1)$$

$$y[n] = \sum_{k=0}^{K-1} g^{[k]}[n] x^{[k]}[n]. \quad (2)$$

Here,  $h^{[k]}[n]$  designates the impulse response function of the subband selective filter of the analysis filter bank,  $k : k \in \mathbb{N}, k < K$  is the subband index and  $*$  denotes convolution. The input-output signal assembly of the AGE is illustrated in Figure 3, where the block for calculating a subband specific gain function is denoted a kernel (KERNEL<sub>*k*</sub> for the *k*<sup>th</sup> subband).

### 3.1.2 A Kernel for Computing a Gain Function

Each kernel employs two measures for calculating the gain function; a fast average and a slow average. The measures are derived according to

$$a_{\text{fast}}^{[k]}[n] = \alpha_{\text{fast}}^{[k]} a_{\text{fast}}^{[k]}[n-1] + \left(1 - \alpha_{\text{fast}}^{[k]}\right) \left|x^{[k]}[n]\right|, \quad (3)$$

$$T^{[k]} = \alpha_{\text{slow}}^{[k]} a_{\text{slow}}^{[k]}[n-1] + \left(1 - \alpha_{\text{slow}}^{[k]}\right) \left|x^{[k]}[n]\right|, \quad (4)$$

$$a_{\text{slow}}^{[k]}[n] = \min\left(T^{[k]}, a_{\text{fast}}^{[k]}[n]\right). \quad (5)$$

Here,  $a_{\text{fast}}^{[k]}[n]$  and  $a_{\text{slow}}^{[k]}[n]$  denotes the fast and slow averages respectively.  $T^{[k]}$  is a prototype variable for temporary use and the function

---

<sup>1</sup>The adopted filter bank topology in part VI refers to a non-decimated filter bank of band pass FIR or IIR filters. Hence, this filter bank in part VI is less versatile than the general filter bank adopted throughout this thesis.

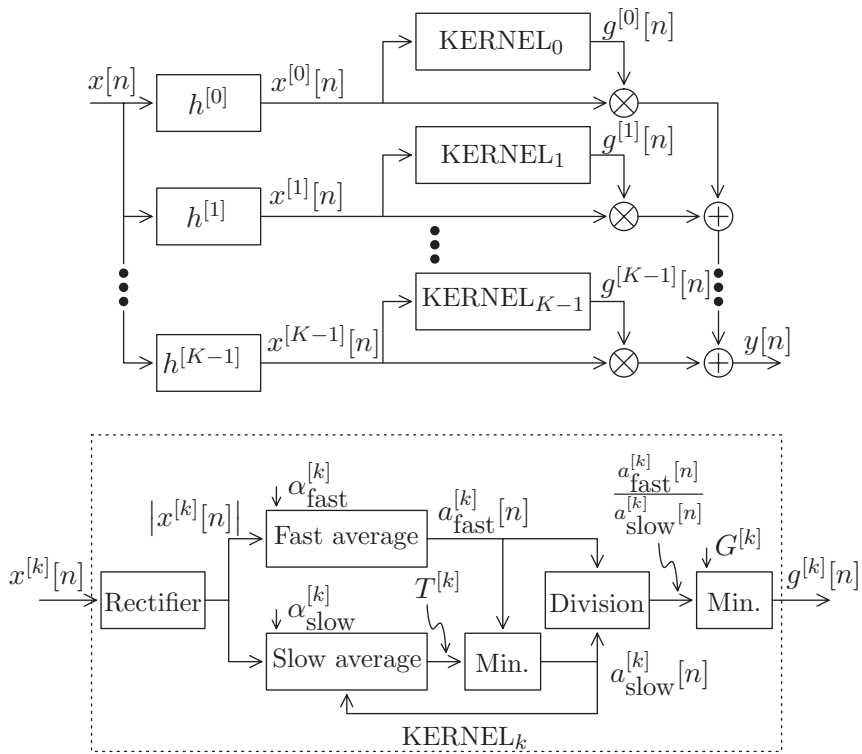


Figure 3: Digital domain Adaptive Gain Equalizer (AGE). Each  $\text{KERNEL}_k$  computes a subband-specific gain function.

$\min(a, b)$  gives the minimum of the two real valued parameters  $a$  and  $b$ . The parameters  $\alpha_{\text{fast}}^{[k]}$  and  $\alpha_{\text{slow}}^{[k]}$  control the tracking performance of the fast and slow averages, where the corresponding time constants are denoted as  $T_{\text{fast}}^{[k]}$  and  $T_{\text{slow}}^{[k]}$ , respectively. The subband-specific gain function is a constrained quotient of the two measures where an upper limit,  $G^{[k]}$ , imposed on the gain function constitutes a constraint according to

$$g^{[k]}[n] = \min \left( \frac{a_{\text{fast}}^{[k]}[n]}{a_{\text{slow}}^{[k]}[n]}, G^{[k]} \right). \quad (6)$$

Hence, the upper limit, in combination with (5), forces the gain function to be bounded to the interval  $1 \leq g^{[k]}[n] \leq G^{[k]}$ , i.e. the AGE focuses on boosting speech.

Research has been conducted using frequency dependent parameters for the AGE [9], i.e. sets of different parameters  $\alpha_{\text{fast}}^{[k]}$ ,  $\alpha_{\text{slow}}^{[k]}$ , and  $G^{[k]}$  are used for different subbands. However, the parameters can be set to the same values for simplicity. Parameters which have been shown to be suitable for many applications are:  $T_{\text{fast}}^{[k]} = 30$  ms,  $T_{\text{slow}}^{[k]} = 3$  s, and  $20 \log G^{[k]} = 10$  dB,  $\forall k$ .

## 4 AGE Implementation Aspects

The speech enhancer outlined in Section 3 will be used as a platform to illustrate implementation aspects from the different domains given in Sections 2.1, 2.2 and 2.3. The AGE will be reformulated to suit the analog domain and hybrid domain. General comments will also be given regarding selected issues thereof.

### 4.1 A Digital Domain Implementation

Although the original formulation of the AGE occurs in the digital domain, this section will comment on the implementation of the AGE

in this same domain.

#### 4.1.1 Implementation Details

There are two sensitive aspects of a digital domain AGE implementation which require special attention. According to (3) and (4) the fast and slow averages are implemented as recursive digital filters. By improper selection of coefficients a digital recursive filter has the risk of being unstable and may suffer from drawbacks such as limit cycle oscillations. Furthermore, when calculating the gain function in (6) a division operation is required. Division on digital circuitry is often tedious and time consuming. An approximation can be made by using a look-up table.

### 4.2 An Analog Domain Implementation

An analog domain implementation of the AGE is topologically identical to the digital domain implementation of the AGE (see Figure 3). The main difference is that the time is now continuous, i.e. the sampled time index,  $n$ , is replaced by the continuous time  $t$ . To facilitate an analog domain implementation, the AGE requires reformulation using analog (continuous time) operators.

#### 4.2.1 Algorithm Reformulation

As in the digital solution, an analysis filter bank is employed for division of the input signal into  $K$  subbands according to

$$x^{[k]}[t] = \int_0^{\infty} h^{[k]}[\tau]x[t - \tau]d\tau. \quad (7)$$

Here  $h^{[k]}[t]$  is the impulse response function of a continuous time band pass filter, i.e. corresponding to a subband selective filter. A vital difference between digital and analog implementation of the AGE involves the ways in which the fast and slow averages are computed. In the digital case, auto regressive averages are employed. In the analog case,

integrators are used to compute the fast and slow averages according to

$$a_{\text{fast}}^{[k]}[t] = \int_0^{\infty} i_{\text{fast}}^{[k]}[\tau] |x^{[k]}[t - \tau]| d\tau, \quad (8)$$

$$a_{\text{slow}}^{[k]}[t] = \min \left( \int_0^{\infty} i_{\text{slow}}^{[k]}[\tau] |x^{[k]}[t - \tau]| d\tau, a_{\text{fast}}^{[k]}[t] \right). \quad (9)$$

The time constants associated with the impulse response functions of the fast and slow averages, i.e.  $i_{\text{fast}}^{[k]}[t]$  and  $i_{\text{slow}}^{[k]}[t]$  respectively, should match those of the corresponding digital structure in (3) and (5). The analog domain AGE gain function is computed as

$$g^{[k]}[t] = \min \left( \frac{a_{\text{fast}}^{[k]}[t]}{a_{\text{slow}}^{[k]}[t]}, G^{[k]} \right). \quad (10)$$

The output signal,  $y[t]$ , is the sum of all weighted subband signals according to

$$y[t] = \sum_{k=0}^{K-1} g^{[k]}[t] x^{[k]}[t]. \quad (11)$$

#### 4.2.2 Implementation Details

The implementation of the analog domain AGE algorithm is made on a Printed Circuit Board (PCB). The design of the PCB is a fairly straightforward task while each individual AGE kernel is identical to one another. While the AGE supports modularized design, a multi-band structure can easily be implemented. The only structural difference between individual subbands lies in the subband-selective filters and in the subband-specific parameters, i.e. the time constants of the fast and the slow averages and the value of the upper limit,  $G^{[k]}$ .

Rudimentary electronic components are used in the PCB design such as Operational Amplifiers (OPAMP), resistors, capacitors, diodes,

transistors and analog multipliers. The classical OPAMP  $\mu A741$  is selected as it is cheap, well known and whose performance is suitable for many general analog electronic building blocks. For division and multiplication a wide bandwidth precision analog multiplier MPY634U from Texas Instruments [16] is selected. The components are powered by a positive,  $V_{DD} = +5V$ , and a negative,  $V_{EE} = -5V$ , supply voltage. The design is separated into four major building blocks: Full-wave rectification, integration, a compare and dump sub-circuit, and gain calculation. The PCB building blocks should be compared to the AGE structure presented in Figure 3. Figure 4 illustrates the full-wave rectifier sub-circuit. The implementation of the fast and the slow integrators,  $i_{\text{fast}}^{[k]}[t]$  and  $i_{\text{slow}}^{[k]}[t]$ , is illustrated in Figure 5. The lower gain limit is implemented by a compare and dump circuit illustrated in Figure 6. The compare and dump sub-circuit compares the level of the slow average to the level of the fast average. If the slow average level is greater than or equal to the fast average level the comparator signals and drives the base of an NPN-BJT transistor, which, in turn, short circuits (dumps) the slow average integrating capacitor towards the ground. Thus, the level of the slow average is inhibited as to never exceed the fast average level. Two sub-circuits are used for gain calculation as illustrated in Figure 7. The first sub-circuit uses an MPY634 circuit in divider mode for calculating the fast and slow averages quotient. Secondly, an upper gain limit is imposed by a zener diode inhibiting the voltage level of the quotient to not exceed the zener diode voltage. Finally, the output of the AGE kernel is formed by multiplication of the gain function and the original input subband signal by using an MPY634 multiplier, see Figure 8.

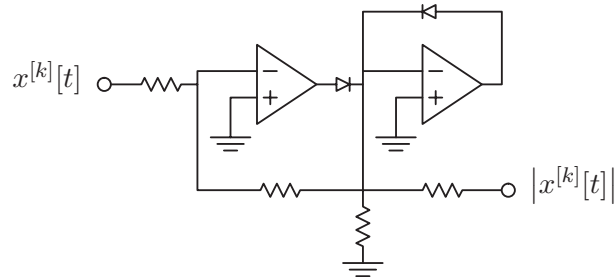


Figure 4: Full-wave rectifier where the output is the absolute value of the input.

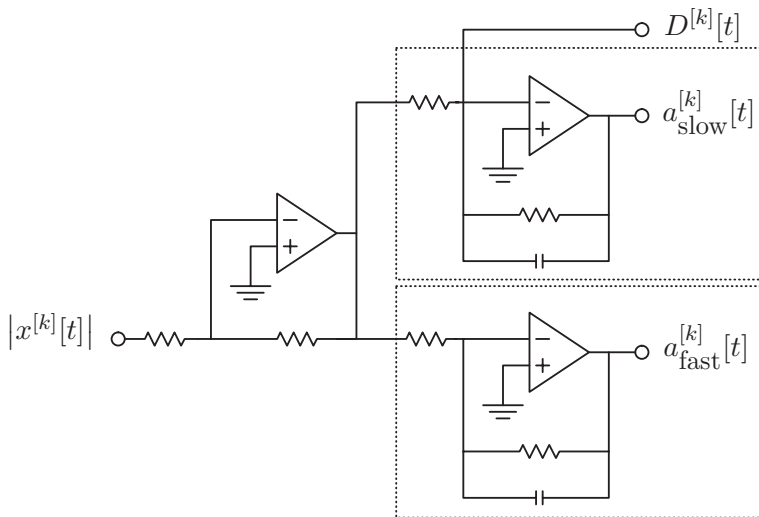


Figure 5: Fast average integrator and a slow average integrator applied to the full wave rectified input signal. The  $D^{[k]}[t]$  wire is for inhibiting the slow average to never exceed the fast average.

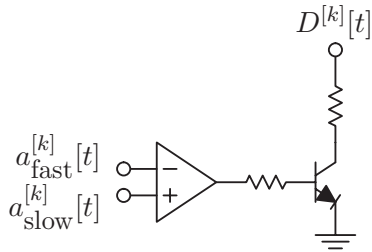


Figure 6: A compare and dump sub-circuit composed of a fast average and a short average comparator. The level of the slow average is ensured to never exceed the fast average by pulling the  $D^{[k]}[t]$  wire towards ground.

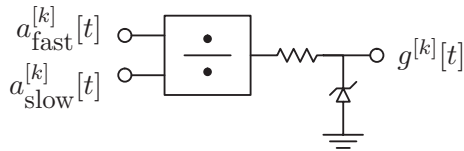


Figure 7: Gain function calculation where a Zener diode constitutes an upper gain function limit.

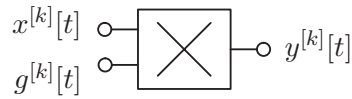


Figure 8: Applying the subband-specific gain function to the subband input signal.

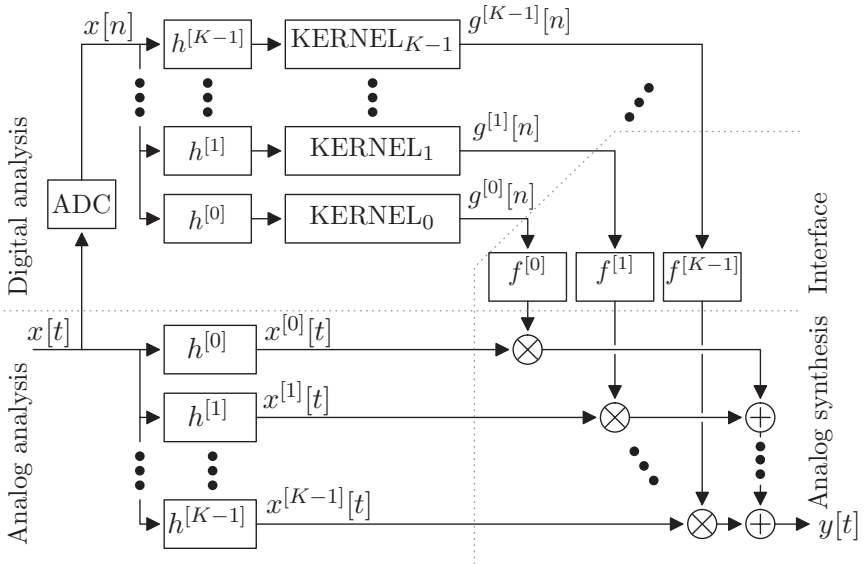


Figure 9: Hybrid domain implementation of the Adaptive Gain Equalizer (AGE) employing digital and analog domain analysis and pure analog synthesis.

### 4.3 A Hybrid Domain Implementation

A hybrid implementation of the AGE has been shown to be very effective, giving high quality speech [12]. However, the AGE needs some reformulation to fit the hybrid domain. In the current implementation, the AGE uses digital and analog analysis and pure analog synthesis, i.e. a digital and an analog signal path are used in parallel. The split analysis and synthesis scheme of the AGE algorithm is illustrated in Figure 9.

### 4.3.1 Algorithm Reformulation

The aim of the hybrid domain AGE implementation is to utilize advantages from analog and digital solutions, such as high signal bandwidth, no quantization of data, reconfigurability, etcetera. Thus, the implementation is split into two parts; digital analysis and analog synthesis. A mapping function,  $f^{[k]}(\cdot)$ , maps the digital analysis gain function,  $g^{[k]}[n]$ , to a corresponding analog synthesis gain function,  $g^{[k]}[t]$ , according to  $g^{[k]}[t] = f^{[k]}(g^{[k]}[n])$ . The structure of the mapping function depends on implementation-specific parameters, such as the maximal gain function value,  $G^{[k]}$ . The analog analysis and synthesis of the AGE algorithm constitutes an analog signal chain from input,  $x[t]$ , to output,  $y[t]$ .

### 4.3.2 Implementation Details

The main issues regarding a hybrid implementation of the AGE, are the design of an analog filter bank, and the control interface between the digital analysis and the analog synthesis. The novel solution to the filter bank issue involves the use of a custom integrated circuit, the Mitsubishi 7-band graphic equalizer M5289P, where the gain in each subband can be individually controlled. The corresponding filter bank in the digital analysis is designed by conventional digital filter design methods. The control interface between the digital analysis and analog synthesis constitutes digitally steered potentiometers controlling the gain in each subband of the analog filter bank.

**Digital Analysis** The digital analysis is performed on a Texas Instrument Mixed Signal Processor (MSP) MSP430F149. A filter bank is implemented in the MSP for approximation of the analog filter bank in the synthesis (see Section 4.3.2). Two-poles and two-zeroes infinite impulse response filter sections form the digital filter bank. The digital subband-selective filters,  $h^{[k]}[n]$  in (1), are designed so that they match the corresponding analog subband selective filters,  $h^{[k]}[t]$ . All parts (ex-

cept for the actual summation of the output signal) are implemented in the MSP, i.e. full wave rectifying, fast average integrator, slow average integrator, lower- and upper gain limiting, and gain function calculation.

While the MSP is a low speed micro controller, a two-level priority scheme is required to ensure full functionality of the processor, constituting a high priority stream and a low priority stream. The high priority stream is running at full sample rate, i.e.  $F_S$ . The subband filtering, full wave rectifying, fast average estimation, and gain steering via a control interface are operations with high priority. The low priority stream uses a Round-Robin time sharing algorithm in which one subband is managed at a time. Calculation of the subband specific slow average, upper- and lower gain limiting, and gain function calculation are time shared with low priority.

**Analog Synthesis** The Mitsubishi Electric M5289P integrated circuit is used as an analog synthesis filter bank. The M5289P is an analog electronics Hi-Fi 7-element graphic equalizer, and employs seven potentiometers for gain control in each subband. The gain of each subband of the M5289P can be controlled in a span of  $10^{-13/20}$  to  $10^{+13/20}$ , by altering the value of the subband-specific potentiometer. Additional capacitors and resistors are used in a subband filter selection network.

**Control Interface** The digital potentiometer, X9C104P, from Xicor is used for individual control of the M5289P subband specific gains. The X9C104P has a resolution of one hundred steps spanning  $100\text{ k}\Omega$  and is steered by the MSP via a three wire digital control interface:  $\overline{U/D}$  - Up/Down,  $\overline{INC}$  - Increment, and  $\overline{CS}$  - Chip Select. The analog synthesis gain function,  $g^{[k]}[t]$ , corresponds to a potentiometer value and is mapped from the digital analysis gain function,  $g^{[k]}[n]$ , using a mapping function,  $f^{[k]}(\cdot)$ .

## 5 Implementation Evaluation

This evaluation is conducted in two parts; First, the analog solution is evaluated and secondly the hybrid solution. These are then compared to a corresponding digital solution. Short term power estimates of the original and enhanced signals are used as a benchmark in the evaluation.

### 5.1 Performance Measures

The performance and quality of a speech enhancement algorithm is not easily quantified. Several objective- and subjective tests exist today as presented in [6]. Examples of objective tests are: Itakura-Saito (IS) Distortion measure, Log-Likelihood Ratio (LLR) measure, Log-Area-Ratio (LAR) measure, and Segmental Signal-to-Noise Ratio (SNR) measure. Examples of subjective tests are: Modified Rhyme Test (MRT), Diagnostic Rhyme Test (DRT), and Mean Opinion Score (MOS). An objective test is selected for evaluation in this report; a short term power estimate. The power measure,  $\Gamma_x[n]$ , of a signal,  $x[n]$ , is computed using an auto regressive average and is herein defined by

$$\Gamma_x[n] = \gamma\Gamma_x[n-1] + (1-\gamma)|x[n]|^2, \quad (12)$$

where  $\gamma = 1 - \frac{1}{F_s T_\gamma}$  (this relationship is valid for  $T_\gamma \gg 1/F_s$ ) is a constant controlling the integration time,  $T_\gamma$  (in [s]), of the power measure. The integration time is set to 25 ms in this evaluation. For comparing the signal before,  $x[n]$ , and after speech enhancement,  $y[n]$ , a power measure quotient is used, defined as  $\Gamma_y[n]/\Gamma_x[n]$ . This quotient is also referred to as the differential short term power estimate.

### 5.2 Analog Domain AGE Evaluation

While all individual subband kernels are identical in their structure, only one kernel is evaluated at a time. The circuit implementation is adjusted before evaluation such that it fulfills recommended algorithm

settings, see Section 3.1. Experiments indicate that, for many practical cases, the recommended algorithm settings ensure natural-sounding speech. The analog implementation is evaluated by real time, on-site measurements. The evaluation setup constitutes a noisy speech signal, which is band pass filtered digitally by a linear phase FIR filter, prior to being presented to the analog circuitry. Several subband filtered signals are processed by the analog AGE structure and relevant signals are recorded by a multi channel Digital Audio Tape (DAT) recorder. The recorded signals are synchronized and summed off-line to form the output signal. In Figure 10, short term power estimates for an input speech signal are presented before and after enhancement, and in Figure 11 a corresponding differential short term power estimate is presented indicating the level of speech enhancement. A speech enhancement performance comparison of the analog method and a corresponding digital implementation is illustrated in Figure 12.

### 5.3 Hybrid Domain AGE Evaluation

This part of the evaluation aims to compare the hybrid AGE implementation to a corresponding digital implementation. Transfer function comparisons show that the filter bank in the digital implementation matches the filter bank in the hybrid implementation. Furthermore, the time constants of the fast average and the slow average are in the same order of magnitude in both the hybrid implementation and the digital implementation. A maximal gain of  $10^{+13/20}$  is used in the digital implementation, i.e. corresponding to 13 dB maximal gain in the hybrid solution.

In Figure 13 the speech enhancement performance of the hybrid method is illustrated. The performance evaluation shows a speech enhancement of maximum 13 dB speech gain, confirmed by the differential short term power measure in Figure 14. A speech enhancement performance comparison of the hybrid method and a corresponding digital implementation is illustrated in Figure 15. The performance of the hybrid method and the digital implementation are remarkably

equal. Subjective listening tests confirm the good performance and high quality of the hybrid speech enhancement implementation.

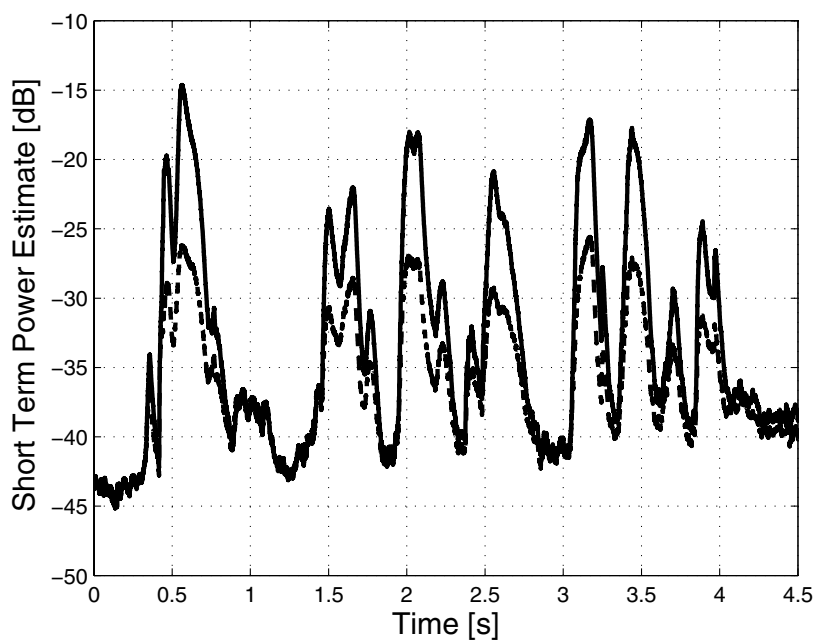


Figure 10: Short term power estimate of a speech signal disturbed by additive noise (dashed) and corresponding power estimate after enhancement by the analog AGE implementation (solid).

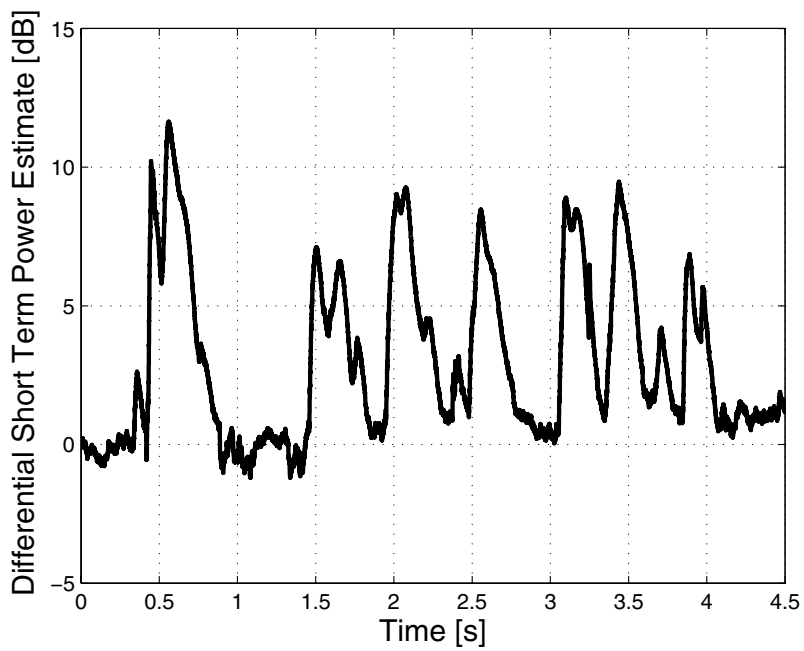


Figure 11: Differential short term power estimate indicating the level of speech enhancement of the analog implementation.

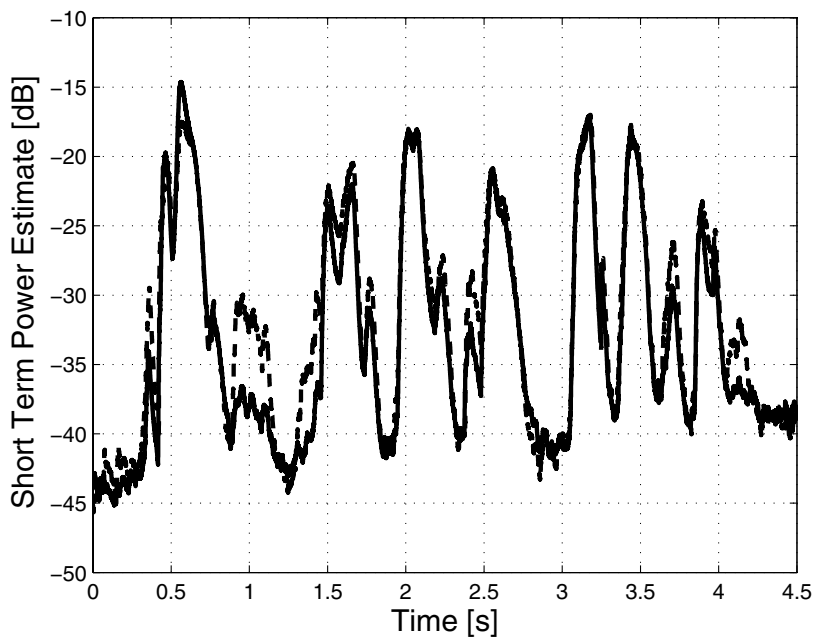


Figure 12: Short term power estimate of a speech signal enhanced by the analog AGE implementation (solid) and a corresponding digital MATLAB implementation (dashed).

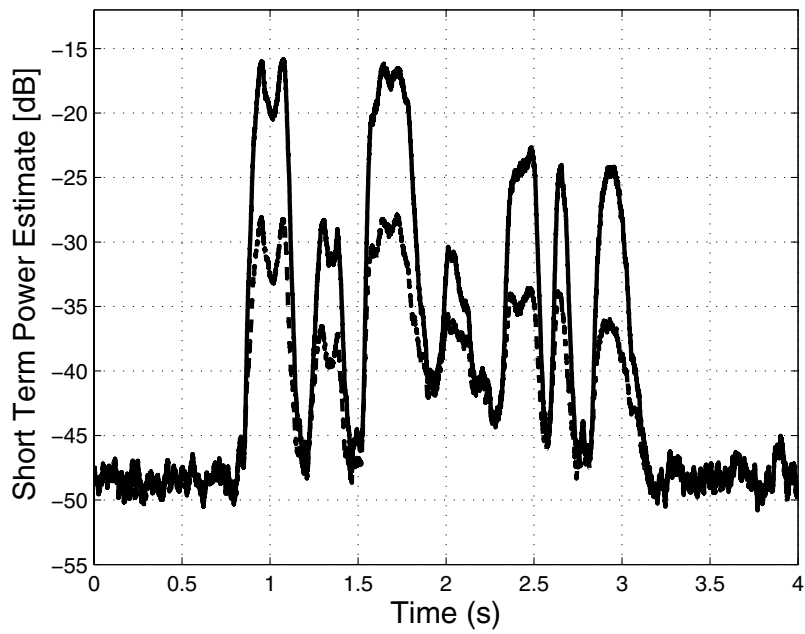


Figure 13: Short term power estimates of an unprocessed speech signal (dashed), and corresponding speech signal processed by the hybrid implementation (solid).

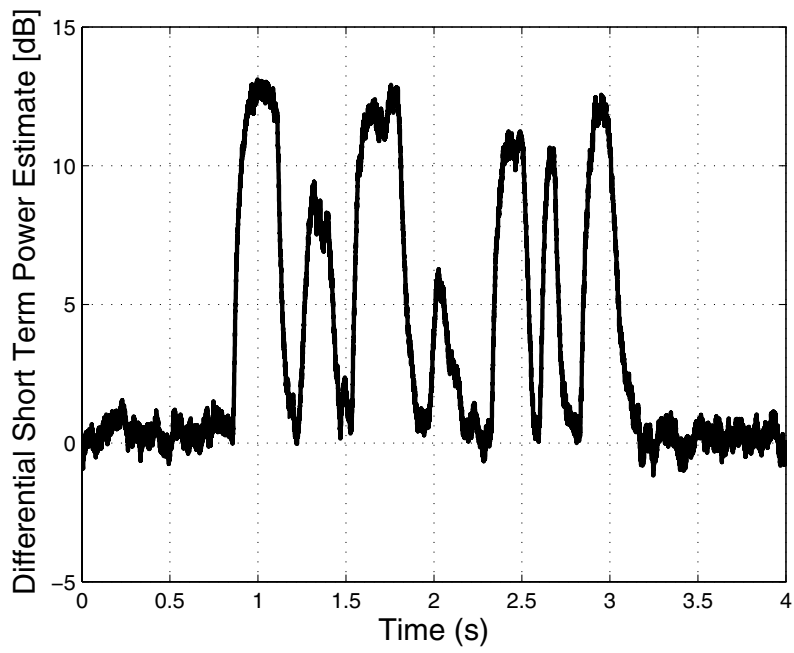


Figure 14: Differential short term power estimate for the hybrid method.

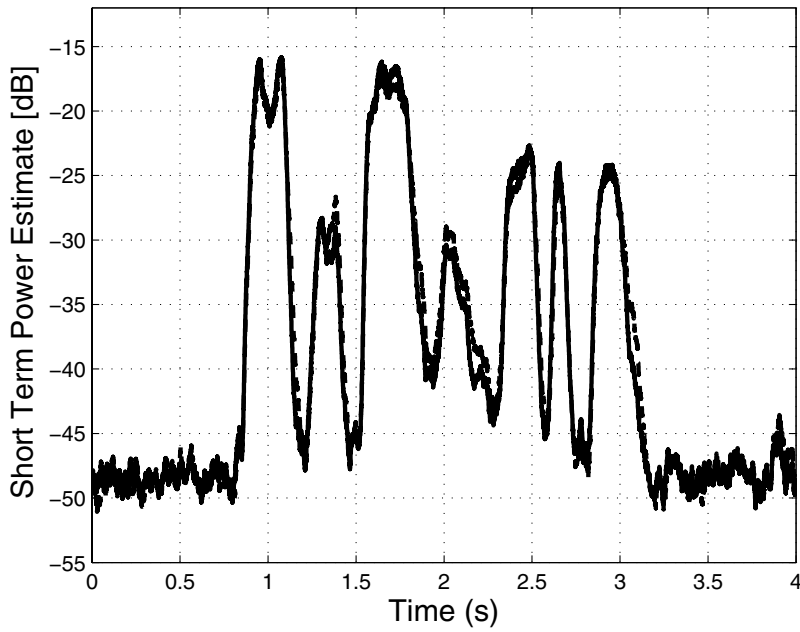


Figure 15: Short term power estimates of a speech signal processed by the hybrid implementation (solid), and processed by a digital MATLAB implementation (dashed).

## 6 Summary and Conclusions

Algorithms for speech enhancement can be realized on digital, analog, and hybrid hardware. The different domains have unique advantages (and disadvantages). When implementing a speech enhancement algorithm the advantages and disadvantages should be taken into consideration. Generally, not all algorithms for speech enhancement are suitable for implementation in a wide range of domains. Special constraints may be put by a specific domain which can not be fulfilled by the algorithm without introducing errors. The algorithm, or alternatively the implementation domain, should be selected with care.

A low complexity speech enhancement method that has been successfully implemented in all three domains is the Adaptive Gain Equalizer (AGE). The predominant advantage of implementing the AGE in the digital domain is its scalability and reconfigurability. It is often important to have the possibility to use digital circuitry for more than one purpose. A major benefit of the analog domain implementation is the continuous time processing which leads to high quality of speech. The hybrid implementation of the AGE is a combination of many advantages. While the analysis is performed in the digital domain, the main advantages of this domain are utilized; such as reconfigurability of the implementation. Due to the analog domain synthesis, the signal path from input to output is completely analog. Hence, the hybrid solution draws benefits from the analog domain such as avoiding quantization of data in the input-to-output signal path. It also introduces negligible restrictions in bandwidth.

In all, the inbound simplicity and ingenuity of the AGE algorithm makes it suitable for implementation in all three domains.

## References

- [1] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal*

*Processing*, ASSP-27:113–120, April 1979.

- [2] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32:1109–1121, December 1984.
- [3] H. Yoo, R. Ellis, D. Anderson, P. Hasler, D. Graham, and M. Hans. A continuous-time speech enhancement front-end for microphone inputs. Technical report, Mobile and Media Systems Laboratory, HP Laboratories Palo Alto, HPL-2002-311, November 7th 2002.
- [4] P. Hasler and D. Andersson. Cooperative analog-digital signal processing. *IEEE International Symposium on Circuits and Systems*, IV:3972–3975, 2002.
- [5] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice Hall, 1988.
- [6] J.H.L. Hansen and B. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. *ICSLP, Sydney, Australia*, pages 2819–2822, Dec 1998.
- [7] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer. *DSPCS*, September 2003.
- [8] N. Westerlund, M. Dahl, and I. Claesson. Real-time implementation of an adaptive gain equalizer for speech enhancement purposes. *WSEAS*, September 2003.
- [9] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement using an adaptive gain equalizer with frequency dependent parameter settings. *VTC04*, September 2004.
- [10] N. Westerlund, M. Dahl, and I. Claesson. Speech enhancement for personal communication using an adaptive gain equalizer. *Elsevier Signal Processing*, 85(6):1089–1101, 2005.

- [11] B. Sällberg, H. Åkesson, N. Westerlund, M. Dahl, and I. Claesson. Analog circuit implementation for speech enhancement purposes. *38th Asilomar Conference on Circuits, Systems and Computers*, Nov 2004.
- [12] B. Sällberg, H. Åkesson, M. Dahl, and I. Claesson. A mixed analog - digital hybrid for speech enhancement purposes. *IEEE International Symposium on Circuits and Systems*, May 2005.
- [13] B. Sällberg and M. Dahl. Speech enhancement implementations in the digital, analog, and hybrid domain. *Swedish System on Chip Conference*, April 2005.
- [14] D. Duarte, V. Narayanan, and M. J. Irwin. Impact of technology scaling in the clock system power. *IEEE ISVLSI*, pages 52–57, 2002.
- [15] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete time processing of speech signals*. Macmillan Publishing Company, 1993.
- [16] Texas Instruments. *MPY634 Wide Bandwidth Precision Analog Multiplier*. Texas Instruments, Dallas, Texas, 2000.





## ABSTRACT

Acoustic disturbances influence human speech communication by interfering with the communication process. In the worst case, it is impossible to communicate at all due to these disturbances. Methods that reduce the influence of the disturbances while preserving speech intelligibility are often desired. This thesis proposes real-world solutions for applied speech enhancement using autonomous and robust methods. Most of the work of the thesis concerns solutions to the problem of reducing acoustic disturbances within the framework of Blind Speech Enhancement (BSE). Notably, the term "blind" is assigned a positive attribute as it implies that the speech enhancement is carried out without any explicit references required. Instead, an assumption about the statistical independence between the sources coupled with an assumption regarding distinguishing statistical properties of the sources underpin the proposed methods. The unifying theory is Independent Component Analysis (ICA), which is performed by means of spatial filtering.

Two of the methods that are proposed in this thesis are shown, both in a theoretical and an empirical framework, to be robust in a real application while preserving stability even for Gaussian-only sources. Existing methods cannot guarantee stability in this scenario and Gaussian-only source mixtures may be the case in a real environment. The difference between the two methods lies in the different optimization strategies and the introduced approximations. The idea of injecting a single-channel method into the control loop of a blind beamformer is also proposed. In particular, two approaches are derived that aim at improving the blind beamformer in the case of disturbing noise and maintaining the same performance for different signal input levels.

Finally, implementation aspects of a single-channel speech enhancer are discussed. The implementation aspects deal with the implementation of a speech enhancer in several different platforms such as analogue hardware, digital hardware, as well as hybrid analogue and digital hardware.

