# Subband Beamforming for Speech Enhancement in Hands-Free Communication

Zohra Yermeche

Ronneby, December 2004
Department of Signal Processing
School of Engineering
Blekinge Institute of Technology
372 25 Ronneby, Sweden

Licentiate Dissertation Series No. 2004:13
ISSN 1650-2140
ISBN 91-7295-053-6

Published 2004
Printed by Kaserntryckeriet AB
Karlskrona 2004
Sweden

# Preface

This licentiate thesis summarizes my work within the field of array speech enhancement for hands-free communication applications. The work has been performed at the Department of Signal Processing at Blekinge Institute of Technology. The thesis consists of three parts, which are:

**Part**

**I** A Constrained Subband Beamforming Algorithm for Speech Enhancement.

**II** Spatial Filter Bank Design for Speech Enhancement Beamforming Applications.

**III** Beamforming for Moving Source Speech Enhancement.

# Acknowledgements

I would like to start by expressing my deep gratitude towards Prof. Ingvar Claesson for giving me the opportunity to pursue my interest in signal processing in the form of a PhD student position at the Blekinge Institute of Technology (BTH). The road leading to my licentiate dissertation would have been hard to follow without the constant guidance and advice from my advisor, Dr. Nedelko Grbić.

Prof. Sven Nordholm recommendation to start doctoral studies at BTH was the best career advice I have received so far. For that I will always be thankful.

During my time in Sweden, Dr. Jan Mark de Haan has been always there for me offering precious help and encouragement. I thank him for being a wonderful friend. Many thanks also to all my colleagues at the Department of Signal Processing for their help, for making me feel at home in their company and for the friendly yet competitive floor-ball games.

My thoughts go also to my family and friends, in Sweden and abroad, who deserve my gratitude for their moral support.

Last but not least, I am grateful to Dragos for his love, support and patience throughout the last year of my thesis work.

*Zohra Yermeche*
*Karlskrona, November 2004*

# Publication list

**Part I is published as:**

Z. Yermeche, N. Grbić and I. Claesson, “A Constrained Subband Beamforming Algorithm for Speech Enhancement,” Research Report, ISSN: 1103-1581, December 2004.

Parts of this research report have been published as:

Z. Yermeche, P. Marquez, N. Grbić and I. Claesson, “A Calibrated Subband Beamforming Algorithm for Speech Enhancement,” published in Second IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, Washington DC, USA, August 2002.

**Part II is published as:**

Z. Yermeche, P. Cornelius, N. Grbić and I. Claesson, “Spatial Filter Bank Design for Speech Enhancement Beamforming Applications,” published in Third IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, Sitges, Spain, July 2004.

**Part III has been submitted for publication in its original form as:**

Z. Yermeche, N. Grbić and I. Claesson, “Beamforming for Moving Source Speech Enhancement,” submitted to IEEE Transactions on Speech and Audio Processing, December 2004.

**Other publications:**

P. Cornelius, Z. Yermeche, N. Grbić, and I. Claesson, “A Spatially Constrained Subband Beamforming Algorithm for Speech Enhancement,” published in Third IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, Sitges, Spain, July 2004.

# Contents

# Introduction

With the maturity in speech processing technologies and the prevalence of telecommunications, a new generation of speech acquisition applications is emerging. This is motivated by the modern society's continuous crave for improving and extending interactivity between individuals, while providing the user with better comfort, flexibility, quality and ease of use.

The emerging broadband wireless communication technology has given rise to the extension of voice connectivity to personal computers, allowing for the development of tele- and video-communication devices, with the objective of enabling natural and accurate communication in both desktop and mobile environments. In today's technology, conference calling stands out as one of the predominant alternatives for conducting high level communications in both small and large companies. This is essentially due to the fact that audio conferencing is convenient and cost effective, considering the reduction of travel expenses it involves.

As a result of the convergence taking place between personal computers and communication devices, telephones and other interactive devices are increasingly being powered by voice. More generally, future ambitions are to replace hand-controlled functions with voice controls, necessitating the development of efficient and robust voice recognition techniques.

The detection, characterization and processing of a various range of signals by technological means is experiencing a growing influence in the biomedical field. More specifically, speech processing techniques have proven to be effective in improving speech intelligibility in noise for hearing-impaired listeners. In addition to the task of helping hearing damages, through the development of the hearing aid industry, speech processing can further be exploited for the task of preventing hearing damages in high noise environment such as air crafts, factories and other industrial working sites.

All these applications have as common denominator, the hands-free acquisition of speech. In other words, the receiver is at a remote distance from the speech transmitting body. This context causes problems of environment noise and interfering sound corrupting the received speech, as well as reverberations of the voice from walls or ceilings, which additionally impairs the received speech signal [1]. In the case of a duplex hands-free communication, the acoustic feedback constitutes another disturbance for the talker who hears his or her voice echoed. Successful speech enhancement solutions should achieve speech dereverberation, efficient noise and interference reduction and, for mobile environments, they should also provide an adaptation capacity to speaker movement.

Many signal processing techniques address these issues separately. Echo cancellation as a research area has been widely explored in the last decades [2, 3, 4]. Speech enhancement in reverberant environment has been considered in [5, 6]. Various background noise reduction methods using one microphone have been developed [7, 8, 9]. Methods using multiple microphones, also referred to as microphone array techniques, aim at addressing the problem in its totality [10, 11, 12, 13, 14, 15, 16]. A large diversity of array processing algorithms derived from classical signal processing methods can be found in the literature. For instance, blind source separation [17, 18] has open the path to speech separation algorithms [19, 20]. Other microphone array techniques using spectral subtraction have been proposed in [21, 22].

The inherent ability of microphone arrays to exploit the spatial correlation of the multiple received signals has enabled the development of combined temporal and spatial filtering algorithms known as beamforming techniques [10]. Some of the classical beamformers include the Delay-and-Sum beamformer, the Filter-and-Sum Beamformer and the Generalized Sidelobe Canceller (GSC). The GSC has been predominantly used for noise suppression. However, it has proven to be sensitive to reverberation [23, 24]. Other advanced beamforming techniques using optimal filtering or signal subspace concepts have been suggested [25, 26, 27]. Many of these algorithms rely on Voice Activity Detection (VAD). This is needed in order to avoid source signal cancellation effects [10], which may result in unacceptable levels of speech distortion. Methods based on calibration data have been developed to circumvent the need of a VAD [28].

Microphone arrays have also permitted the emergence of localization algorithms to detect the presence of speech, determine the direction of the speaker and track it when it moves [29, 30, 31]. Combined with video technology, these techniques can allow the system to steer and concentrate on the speaker, thus provide a combined video and audio capability [32].

In this thesis an adaptive subband RLS beamforming approach is investigated and evaluated in real hands-free acoustical environments. The proposed methodology is defined such to perform background noise and acoustic coupling reduction, while producing an undistorted filtered version of the signal originating from a desired location. This adaptive structure allows for a tracking of the noise characteristics, such to accomplish its attenuation in an efficient manner. A soft constraint built from calibration data in low noise conditions guarantee the integrity of the desired signal without the need of any speech detection. A subband beamforming structure is used in order to improve the performance of the time-domain filters and reduce their computational complexity. A new spatial filter bank design method, which includes the constraint of signal passage at a certain position, is suggested for speech enhancement beamforming applications. Further, a soft constrained beamforming approach with built-in speaker localization, is proposed to accommodate for source movement. Real speech signals are used in the simulations and results show accurate speaker movement tractability with good noise and interference suppression.

In this chapter, a brief description of sound propagation is first given, followed by an introduction to acoustic array theory. The next section contains a summary of the existing microphone array beamforming concept and techniques for speech enhancement. Then, approaches for localization and tracking are presented. At last, an overview of this thesis content is given.

# 1 Sound Propagation

Sound waves propagate through an air medium by producing a movement of the molecules in the direction of propagation and are referred to as compressional waves. The wave equation for acoustic waves propagating in a homogeneous and lossless medium can be expressed as [10]

$$\nabla^2 x(t,\mathbf{r}) - \frac{1}{c^2}\frac{\delta^2}{\delta t^2}\, x(t,\mathbf{r}) = 0, \tag{1}$$

where $x(t,\mathbf{r})$ is a function representing the sound pressure at a time instant $t$ for a point in space with Cartesian coordinates $\mathbf{r} = [x,y,z]^T$. Here, $\nabla^2$ stands for the Laplacian operator and $(^T)$ is the transpose. The variable $c$ is the speed of propagation, which depends upon the pressure and density of the medium, and thus is constant for a given wave type and medium. For the specific case of acoustic waves in air, the speed of propagation is approximately 340 ms$^{-1}$.

In general, waves propagate from their source as spherical waves, with the amplitude decaying at a rate proportional to the distance from the source [33]. These properties imply a rather complex mathematical analysis of propagating signals, which is a major issue in array processing of near-field signals. However, at a sufficiently long distance from the source, acoustic waves may be considered as plane waves, considerably simplifying the analysis. The solution of the wave equation for a monochromatic plane wave is given by [10]

$$x(t, \mathbf{r}) = A\, e^{j(\omega t - \mathbf{k}^{\mathbf{T}}\mathbf{r})}, \tag{2}$$

where $A$ is the wave amplitude, $\omega = 2\pi f$ is the angular frequency, and $\mathbf{k}$ is the *wave number vector*, which indicates the speed and direction of the wave propagation. The wave number vector is given by

$$\mathbf{k} = \frac{2\pi}{\lambda}[sin\phi\, cos\theta \quad sin\phi\, sin\theta \quad cos\phi]^T, \tag{3}$$

where $\theta$ and $\phi$ are the spherical coordinates, as illustrated in Fig. 1 and $\lambda$ is the wavelength.

Due to the linearity of the wave equation, the monochromatic solution can be expanded to the more general polychromatic case by considering the solution as a sum of complex exponentials. More generally, the Fourier theory can be exploited to form an integral of complex exponentials to represent an arbitrary wave shape [10].

A band limited signal can be reconstructed by temporally sampling the signal at a given location in space, or spatially sampling the signal at a given instant in time. Additionally, the superposition principle applies to propagating wave signals, allowing multiple waves to occur without interaction [10]. Based on these conclusions, the information carried by a propagating acoustic wave can be recovered by proper processing using the temporal and spatial characteristics of the wave.

## 1.1 Noise Field

The acoustic field in the absence of information transmission is commonly referred to as a *noise field* (or *background noise field*). In general, it consists of the summation of a large diversity of unwanted or disturbing acoustic waves introduced in a common field by man-made and natural sources. Hence, depending on the degree of correlation between noise signals at distinct spatial locations, different categories of noise fields can be defined for microphone array applications [34].

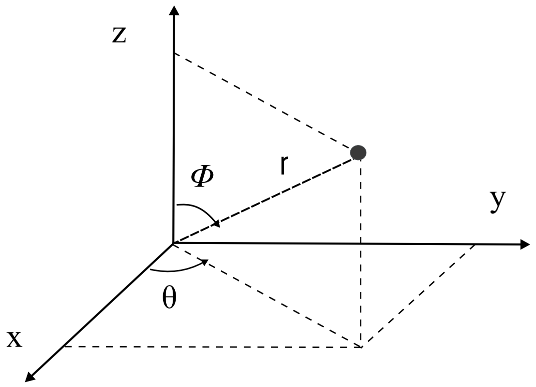


Figure 1: *Cartesian coordinates [x,y,z] and spherical coordinates [r,$\theta$,$\phi$] of a point in space.*

### Coherent versus Incoherent Noise Field

A coherent noise field corresponds to noise signals propagating from their source without undergoing reflection, dispersion or dissipation. It is characterized by a high correlation between received signals at different spatial locations. A coherent noise field results from a source in open air environments with no major obstacles to sound propagation. An incoherent noise field, in the other hand, is characterized by spatially uncorrelated noise signals. An example of incoherent noise is electrical noise in microphones, which is generally viewed as randomly distributed.

### Diffuse Noise Field

A diffuse noise field corresponds to noise signals propagating in all directions simultaneously with equal energy and low spatial correlation. In practice, many noise environments such as the noise in a car and an office can be characterized by a diffuse noise field, to some extend.

# 2 Acoustic Arrays

Acoustic sensor arrays consist of a set of acoustic sensors placed at different locations in order to receive a signal carried by propagating waves. Sensor arrays are commonly considered as spatially sampled versions of continuous sensors, also referred to as *apertures.* From this perspective, sensor array fundamentals can conveniently be derived from continuous aperture principles by means of the sampling theory.

## 2.1 Continuous Aperture

A continuous aperture is an extended finite area over which signal energy is gathered. The two major concepts in the study of continuous aperture are the *aperture function* and the *directivity pattern.*

***The aperture function*** defines the response of a spatial position along the aperture to a propagating wave. The aperture function, denoted in this text by $\omega(\mathbf{r})$, takes values between zero and one inside the region where the sensor integrates the field and is null outside the aperture area [10].

***The directivity pattern*** also known as *beam pattern* or *aperture smoothing function* [10], corresponds to the aperture response as a function of direction of arrival. It is related to the aperture function by the three dimensional Fourier transform relationship following

$$W(f, \boldsymbol{\alpha}) = \int_{-\infty}^{+\infty} \omega(\mathbf{r}) e^{j2\pi\boldsymbol{\alpha}^T\mathbf{r}} d\mathbf{r}, \tag{4}$$

where the direction vector $\boldsymbol{\alpha} = [\alpha_x, \alpha_y, \alpha_z]^T = \mathbf{k}/2\pi$.

### Linear Aperture

For a linear aperture of length $L$ along the x-axis centered at the origin of the coordinates ( i.e. corresponding to spatial points $\mathbf{r} = [x, 0, 0]$, with $-L/2 < x < L/2$), the directivity pattern can be simplified to

$$W(f, \alpha_x) = \int_{-L/2}^{L/2} \omega(x) e^{j2\pi\alpha_x x} dx. \tag{5}$$

For a uniform aperture function defined by

$$\omega(x) = \begin{cases} 1 & \text{when} \mid x \mid \leq L/2, \\ 0 & \text{when} \mid x \mid > L/2, \end{cases}$$

the resulting directivity pattern is given by

$$W(f, \alpha_x) = Lsinc(\alpha_x L). \tag{6}$$

The directivity pattern corresponding to a uniform aperture function, is illustrated in Fig. 2. It can be seen that zeros in the directivity pattern are located at $\alpha_x = m\lambda/L$. The beam width of the main lobe is given by $2\lambda/L = 2c/fL$. Thus, for a fixed aperture length, the main lobe is wider for lower frequencies. Considering only the horizontal directivity pattern, i.e. $\phi = \pi/2$, a polar plot is shown in Fig. 3.

## 2.2 Linear Sensor Array

A sensor array can be viewed as an aperture excited at a finite number of discrete points. For a linear array with $I$ identical equally spaced sensors, the far-field horizontal directivity pattern is given by

$$W(f, \theta) = \sum_{i=1}^{I} \omega_i \, e^{j \frac{2\pi f}{c} \, i \, d \, cos\theta}, \tag{7}$$

where $\omega_i$ is the element $i$ complex weighting factor and where $d$ is the distance between adjacent sensors. In the case of equally weighted sensors, $\omega_i = 1/I$, it can be seen from the evaluation of Eq. (7) for different values of the parameters $I$ and $d$ that increasing the number of sensors $I$ results in lower side lobes. On the other hand, for a fixed number of sensors, the beam width of the main lobe is inversely proportional to the sensor spacing $d$.

### Spatial aliasing

In an analogous manner to temporal sampling of continuous-time signals, spatial sampling can produce aliasing [10]. Spatial aliasing results in the appearance of spurious lobes in the directivity patters, referred to as *grating lobes*, as illustrated in Fig. 4. The requirement to fulfill the spatial sampling theorem, so to avoid spatial aliasing, is given by

$$d < \frac{\lambda_{min}}{2}, \tag{8}$$

where $\lambda_{min}$ is the minimum wavelength in the propagating signal. Hence, the critical spacing distance required for processing signals within the telephone bandwidth (300-3400 Hz) is $d = 5$ cm.

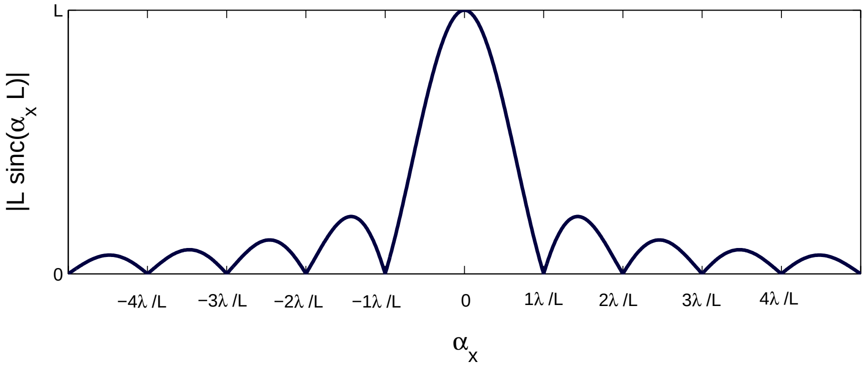
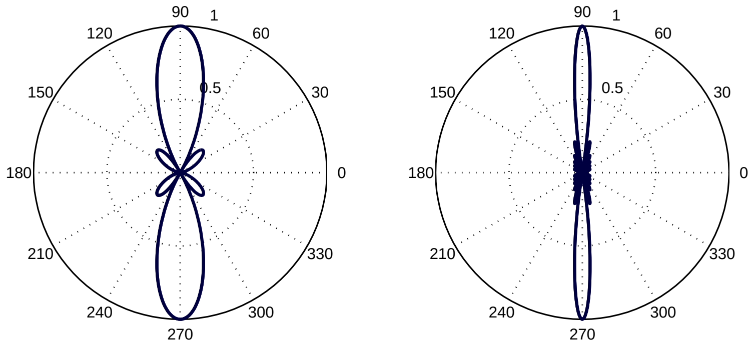


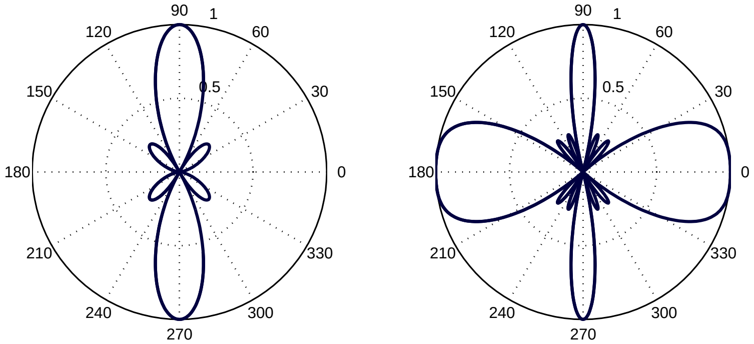
Figure 2: *Directivity pattern of a linear aperture.*



L / λ = 2

L / λ = 6

Figure 3: *Polar plot of the directivity pattern of a linear aperture as a function of the horizontal direction $\theta$, with $L/\lambda = 2$ (left) and $L/\lambda = 6$ (right). It can be seen that for a higher frequency, i.e. $L/\lambda$ higher (right) the main beam is narrower.*



d = λ/2

d = λ

Figure 4: *Spatial aliasing: Polar plot of the directivity pattern of a linear sensor array with four elements, as a function of the horizontal direction $\theta$; with a critical spatial sampling, $d = \lambda/2$ (left), and with aliasing effects for $d = \lambda$ (right).*

# 3 Microphone Array Beamforming techniques

Microphone arrays spatially sample the sound pressure field. When combined with spatio-temporal filtering techniques known as *beamforming*, they can extract the information from (spatially constrained) signals, of which only a mixture is observed.

In this section an introduction to the principle of beamforming is first given, followed by a description of the classical beamforming techniques: the Delay-and-Sum beamformer and the Filter-and-Sum beamformer. Post filtering is presented as an approach to improve the beamformer's performance. For optimal usability of the beamformer' structure in (temporally and spatially) non-stationary environments, adaptive beamforming techniques have been developed. The most popular of these algorithms are described next, including the Constrained Minimum Variance beamformer and its main variant the Frost's algorithm, the Generalized Sidelobe Canceller and the calibrated beamformer. Subband beamforming techniques are also introduced as an alternative to reduce the complexity of the beamforming filtering operation.

## 3.1 Beamforming and Classical Beamformers

The complex weighting element $\omega_i$ in the far-field horizontal directivity pattern of a linear sensor array can be expressed in terms of its magnitude and phase components as

$$\omega_i = a_i \, e^{j\varphi_i}. \tag{9}$$

The directivity pattern of Eq. (7) is reformulated as

$$W(f,\theta) = \sum_{i=1}^{I} a_i \, e^{j(\frac{2\pi f}{c} \, i \, d \, cos \, \theta + \varphi_i)}. \tag{10}$$

While the amplitude weights $a_i$ control the shape of the directivity pattern, the phase weights $\varphi_i$ controls the angular location of the response's main lobe. Beamforming techniques are algorithms for determining the complex sensor weights $\omega_i$ in order to implement a desired shaping and steering of the array directivity pattern.

### Delay-and-Sum Beamformer

The complex weights with frequency-dependent phase

$$\omega_i = \frac{1}{I} e^{j \frac{-2\pi f}{c} (i-1) \, d \, cos\theta_s}, \tag{11}$$

leads to the directivity pattern

$$W(f,\theta) = \frac{1}{I}\sum_{i=1}^{I} e^{j\frac{2\pi f}{c}\,(i-1)\,d\,(cos\theta - cos\theta_s)}, \tag{12}$$

such that an angular shift with angle $\theta_s$ of the directivity pattern's main lobe is accomplished.

By summing the weighted channels, the array output is given by

$$Y(f) = \frac{1}{I}\sum_{i=1}^{I} X_i(f) e^{j\frac{-2\pi f}{c}(i-1)\,d\,cos\theta_s}. \tag{13}$$

where $X_i(f)$ is the frequency representation of the sound field received at sensor $i$. The negative phase shift realized in the frequency domain corresponds to introducing a time delay of the sensor inputs, according to

$$y(t) = \frac{1}{I}\sum_{i=1}^{I} x_i(t-\tau_i), \tag{14}$$

where the delay for sensor $i$ is defined as $\tau_i = \frac{(i-1)\,d\,cos\theta_s}{c}$. This summarizes the formulation of the elementary beamformer known as the *Delay-and-Sum beamformer.*

### Filter-and-Sum Beamformer

In the Filter-and-Sum beamformer both the amplitude and the phase of the complex weights are frequency dependent, resulting in a filtering operation of each array element input signal. The filtered channels are then summed, according to

$$Y(f) = \sum_{i=1}^{I} \omega_i^{(f)} X_i(f). \tag{15}$$

The multiplications of the frequency-domain signals are accordingly replaced by convolutions in the discrete-time domain. The discrete-time output signal is hence expressed as

$$y(n) = \sum_{i=1}^{I}\sum_{l=0}^{L-1} \omega_i(l) x_i(n-l), \tag{16}$$

where $x_i(n)$ are sampled observations from sensor $i$, $\omega_i(l)$, $l = 0, ..., L-1$, are the filter weights for channel $i$, and $L$ is the filter length.

## 3.2 Post-Filtering

Post-filtering is a method to improve the performance of a filter-and-sum beamforming algorithm. This concept makes use of the information about the desired signal acquired by the spatial filtering, to achieve additional frequency filtering of the signal. A Wiener post-filter approach was suggested in [35]. It makes use of cross spectral density functions between channels, which improves the beamformer cancellation of incoherent noise as well as coherent noise, as long as it is not emanating from the look-direction. However, the effectiveness of this post filter has been shown to be closely linked to the beamformer performance [36].

## 3.3 Adaptive Beamforming

Adaptive beamforming techniques attempt to adaptively filter the received signals in order to pass the signal coming from a desired direction and suppress the unwanted signals coming from other directions. This is achieved by combining the classical beamforming principles with adaptive filter theory.

Least Mean-Square (LMS)-based beamforming focuses on the minimization of the mean-square error between a reference signal, highly correlated to the desired signal, and the input signal. This algorithm does not put any requirement on the signal's spatial characteristics, and it relies strictly on acquiring a reference signal with a good correlation to the desired signal. However, the LMS algorithm objective is solely to minimize the mean-square error, based on instantaneous correlation measures, without any condition upon the distortion of the signal. It results in a degradation of the desired signal, mainly in high noise environments. This limitation can be circumvented by the introduction of a constraint on the adaptive filter weights, based on adequate knowledge of the source, such to secure the passage of the desired signal. The filter optimization process can thus be viewed as a constrained least mean-square minimization.

### The Constrained Minimum Variance Beamformer

Constrained minimum variance beamforming is based on the concept of a constrained LMS array, which consists of minimizing the output from a sensor array while maintaining a constant gain towards the desired source. The most famous constrained minimum variance algorithm is the Frost's algorithm presented in [37]. The Frost's algorithm requires knowledge of the desired signal location and the array geometry in order to define a constraint on the filter weights such to ensure that the response to the signal coming from the desired direction has constant gain and linear phase. This is achieved in conjunction with a minimization of the received energy components originating from other directions. This structure presents a high sensitivity to steering vector errors.

### The Generalized Sidelobe Canceller

The Generalized Sidelobe Canceller (GSC) is an adaptive beamforming structure which is used to implement the Frost' algorithm as well as other linearly constrained minimum variance beamformers, in an unconstrained frame [10, 23]. The GSC relies on the separation of the beamformer into a fixed and an adaptive part. The fixed portion steers the array towards the desired direction such to identify the signal of interest. The desired signal is then eliminated from the input to the adaptive part by a blocking matrix, ensuring that the power minimization is done over the noise only.

In practice, it is rather difficult to achieve a perfect signal cancellation over a large frequency band. Thus, for broadband signals such as speech, the blocking matrix can not totally prohibit the desired signal from reaching the adaptive filters. This phenomenon known as the *superresolution problem* can cause the GSC algorithm to distort and even cancel the signal of interest.

### The Calibrated Microphone Array

The *In situ* Calibrated Microphone Array (ICMA) is an adaptive beamformer which relies on the use of calibration sequences, previously recorded in the environment of concern. In this way, the spatio-temporal characteristics of the environment are taken into account in the formulation of the system's response to the desired signal. This methodology does not require any knowledge of the desired source and the array positioning nor specifications.

The ICMA structure is built in two steps. In a pre-processing phase, calibration sequences are recorded for the desired speech position and the known interfering speech positions, separately, in a low noise environment. These calibration sequences which are stored in memory are added to the received array signals during processing. The target calibration sequence which is spatially correlated with the desired signal serves as a reference signal in the adaptive minimization of a mean-square error. A VAD is employed to limit the adaptation process to the time frames corresponding to silent speech.

The ICMA structure has been implemented based on the Normalized LMS

(NLMS) algorithm in [14], and on a Least-Squares (LS) solution in [15]. A Neural Network approach has also been investigated [16]. The theoretical limits of the ICMA scheme were established in [38], showing a robustness of this structure to perturbation errors and the superresolution problem.

## 3.4 Subband Beamforming

The use of a Recursive Least-Squares (RLS) algorithm, as an alternative to the LMS adaptive approach of a beamformer for speech processing, requires manipulating large matrices. The implementation of such complex algorithm can be made possible through a subband beamforming structure. Subband adaptive filtering principle consists on converting a high order full band filtering problem to a set of low order subband filtering blocks, with the aim of reducing complexity while improving performance [39]. Subband beamforming can be achieved by means of a frequency transform, which allows for the filter weight computation to be performed in the frequency domain. The time-domain convolutions of Eq. (16) are, thus, replaced by multiplications in the frequency domain following Eq. (15). The computational gain comes from the fact that the processing of narrow band signals requires lower sample rates [40]. Hence, in an efficient implementation, the frequency transform is followed by a decimation operation.

Fig. 5 illustrates the overall structure of a subband beamformer. The input signal for each microphone is decomposed into a set of narrow band signals using a multichannel subband transformation, also known as an analysis filter bank. The beamformer filtering operations are then performed for each frequency subband separately. A full band signal is reconstructed using a synthesis filter bank. It represents the output of the total system.

### Filter banks

Different frequency transformations can be used for subband beamforming applications. Among the most important elements in frequency transformation are the Discrete Fourier Transform (DFT) and its fast implementation the Fast Fourier Transform (FFT). These frequency transformations are often built in the form of a bank of filters and they should be constructed such to cancel aliasing effects introduced by decimation operations.

Many design methods for filter banks have been developed based on various optimization criteria. In multi-rate filter banks, in-band aliasing and imaging distortion are a major issue [40]. Another optimization parameter

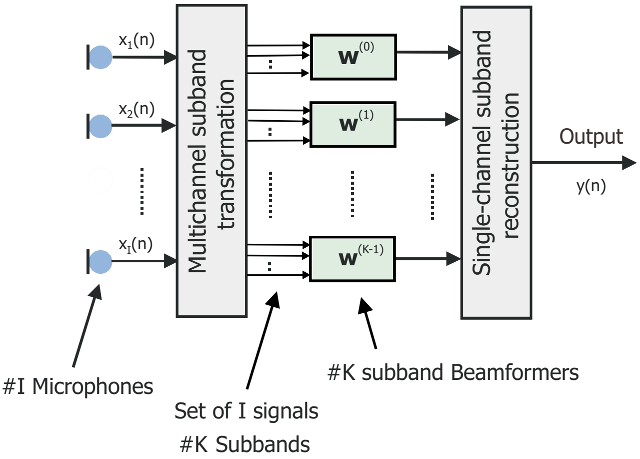


Figure 5: *Structure of the subband beamformer.*

to be considered is the delay introduced by the filter banks, where a tradeoff can be made between low delay and reduced complexity. Uniform and non-uniform methods based on delay specification, including delay-less structures, are presented in [41, 42, 43].

In a uniform filter bank design with modulated filter banks, a simplified structure is made available through the use of efficient polyphase implementation [40]. Exploiting this feature, an oversampled uniform DFT FIR filter bank design method was presented in [41], where aliasing and output signal distortion are minimized, under a pre-specified delay constraint.

# 4 Localization and Tracking

Speaker localization is of particular interest in the development of speech enhancement methods requiring information of the speaker position. Based on the localized speaker position, the microphone array can be steered towards the corresponding direction for effective speech acquisition. This approach is appropriate for speech enhancement applications with a moving speaker, such as in video-conferencing, where the speaker position can be provided to a video system in order to keep the speaker in focus of the camera [32]. A localization system may also be used in a multi-speaker scenario to enhance speech from a particular speaker with respect to others or with respect to noise sources.

The beamforming principle may be used as foundation for source localization by steering the array to various spatial points to find the peak in the output power. Localization methods based on the maximization of the Steered Response Power (SRP) of a beamformer, have been shown to be robust [32]. However, they present a high dependency on the spectral content of the source signal, which in most practical situations is unknown.

The most widely used source localization approach exploits time-difference of arrival (TDOA) information. A signal originated from a point in space is received by a pair of spatially distinct microphones with a time-delay difference. A specific delay can be mapped to a number of different spatial points along a hyperbolic curve as illustrated in Fig. 6. For a known array geometry, intersecting the hyperbolic curves, corresponding to the temporal disparity of a received signal relative to pairs of microphones, results in an estimate of the source location.

Acquiring a good time-delay estimation (TDE) of the received speech signals is, thus, essential to achieve an effective speaker localization. Most TDE methods have limited accuracy in the presence of background noise and reverberation effects [44]. The time-delay may be estimated by maximizing the cross-correlation between filtered versions of the received signals, which is the basis of the Generalized Cross Correlation (GCC) method. This approach is however impractical in high reverberant environment where the signal's spectral content is corrupted by the channel's multi-path [44]. This problem can be circumvented by equalizing the frequency-dependent weightings of the cross-spectrum components, such to obtain a peak corresponding to the dominant delay of the signal. The extreme case where the magnitude is flattened is referred to as the Phase Transform (PHAT).

A merge of the GCC-PHAT algorithm and the SRP beamformer, resulted in the so called SRP-PHAT algorithm. By combining the robustness of the steered beamformer and the insensitivity to received signal characteristics introduced by the PHAT approach, this algorithm has been shown to be robust and to provide reliable location estimates [32].

Other GCC-PHAT based-methods using pre-filtering, eigendecomposition or speech modelling have been presented in [30, 31].

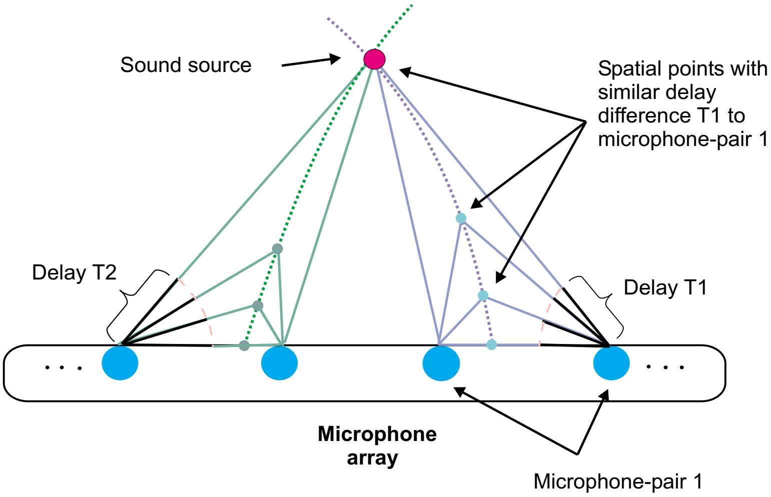


Figure 6: *Time-delay estimation of a point source signal relative to pairs of microphones.*

# 5 Thesis Overview

## PART I - A Constrained Subband Beamforming Algorithm for Speech Enhancement

This paper presents a comprehensive study of a calibrated subband adaptive beamformer for speech enhancement, in hands-free communication, which does not need the use of a VAD. Performance of the algorithm is evaluated on real data recordings conducted in typical hands-free environments. The beamformer is based on the principle of a soft constraint, formed from calibration data, rather than precalculated from free-field assumptions, as it is done in [45]. The benefit is that the real room acoustical properties will be taken into account. The algorithm recursively estimates the spatial information of the received data, while the initial precalculated source correlation estimates constitute a soft constraint in the solution. A subband beamforming scheme is used, where the filter banks are designed with the methodology described in [41], which minimizes in-band and reconstruction aliasing effects.

A real hands-free implementation with a linear array, under noisy conditions such as a crowded restaurant room and a car cabin in movement, shows good noise and interference suppression as well as low speech distortion.

## Part II - Spatial Filter Bank Design for Speech Enhancement Beamforming Applications

In this paper, a new spatial filter bank design method for speech enhancement beamforming applications is presented. The aim of this design is to construct a set of different filter banks that includes the constraint of signal passage at one position (and closing in other positions corresponding to known disturbing sources) as depicted in Fig. 7. By performing the directional opening towards the desired location in the fixed filter bank structure, the beamformer is left with the task of tracking and suppressing the continuously emerging noise sources. This algorithm has been tested on real speech recordings conducted in a car hands-free communication situation. Results show that a reduction of the total complexity can be achieved while maintaining the noise suppression performance and also reducing the speech distortion.

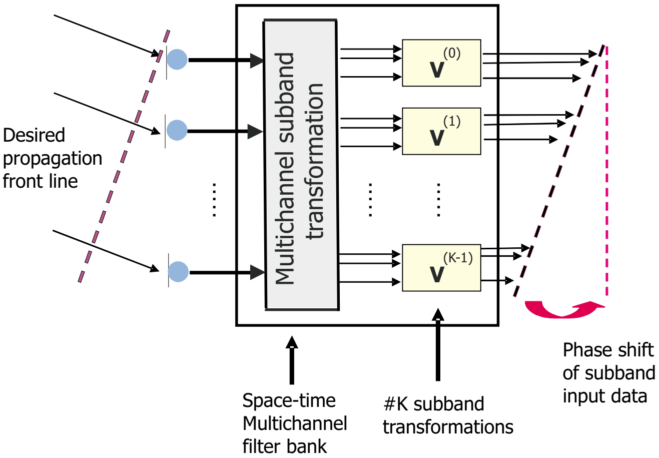


Figure 7: *Structure of the multidimensional space-time filter bank (The output data of the filter bank are phase-shifted to be in-phase for the source propagation direction, and out-of-phase for interference propagation directions).*

## Part III - Beamforming for Moving Source Speech Enhancement

To allows for source mobility tracking, a soft constrained beamforming approach with built-in speaker localization is proposed in this part. The beamformer is based on the principle of a soft constraint defined for a specified region corresponding to an estimated source location and a known array geometry, rather than formed from calibration data. An algorithm for sound source localization is used for speaker movement tracking. The source of interest is modelled as a cluster of stationary point sources and source motion is accommodated by revising the point source cluster. The source modelling and its direct exploitation in the beamformer through covariance estimates is presented. The choice of the point source cluster affects the updating of the covariance estimates when the source moves. Thus, a design tradeoff between tractability of updating and performance is considered in placement of these points. Real speech signals are used in the simulations and results show accurate speaker movement tracking with maintained noise and interference suppression of about 10-15 dB, when using a four-microphone array.

## References

[1] J.R. Deller Jr., J. G. Proakis, and J. H. L. Dudgeon, *Discete-Time Processing of Speech Signals*, Macmillan, 1993.

[2] D. G. Messerschmidt, "Echo Cancellation is Speech and Data Transmission," in *IEEE Journal Selected Area in Communications* , pp. 283–297, March 1984.

[3] M. Sondhi, and W. Kellerman, "Adaptive Echo Cancellation for Speech Signals," in *Advances in Speech Signal Processing*, S. Furui and M. Sondhi Eds., 1992.

[4] C. Breining, P. Dreiseitel, E. Hansler, A. Mader, B. Nitsch, H. Puder, T. Schertler, G. Schmidt, and J. Tilp, "Acoustic Echo Control -An Application of Very-High-Order Adaptive filters," in *IEEE Signal Processing Magazine*, pp. 42–69, July 1999.

[5] B. W. Gillespie, R. S. Malver, and D. A. F. Florencio, "Speech Dereverberation via Maximum-Kurtosis Subband Adaptive filtering," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001.

[6] M. Wu, and D. Wang, "A One-Microphone Algorithm for Reverberant Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 892–895, April 2003.

[7] S. Boll, "Suppression of Acoustical Noise in Speech Using Spectral Substraction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[8] P. Vary, "Noise suppression by Spectral Magnitude Estimation - Mechanism and Theoretical Limits-," in *Elsevier Signal Processing*, vol. 8, pp. 387–400, 1985.

[9] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 363–366, April 1993.

[10] D. Johnson, and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.

[11] Y. Kaneda, and J. Ohga, “Adaptive Microphone-Array System for noise reduction,” in *IEEE Transaction Conference on Acoustics, Speech and Signal Processing*, vol. 34, no. 6, pp. 1391–1400, December 1986.

[12] Y. Grenier, and M. Xu “An Adaptive Array for speech Input in Cars,” in *Proceedings of International Symposium of Automotive Technology and Automation*, 1990.

[13] S. Nordholm, I. Claesson, and B. Bengtsson, “Adaptive Array Noise Suppression of Hands-free Speaker Input in Cars,” in *IEEE Transactions in Vehicular Technology*, vol. 42, no. 4, pp. 514–518, November 1993.

[14] M. Dahl and I. Claesson, “Acoustic Noise and Echo Canceling with Microphone Array,” in *IEEE Transactions in Vehicular Technology*, vol. 48, no. 5, pp. 1518–1526, September 1999.

[15] N. Grbić, “Speech Signal Extraction - A Multichannel Approach,” University of Karlskrona/Ronneby, ISBN 91-630-8841-x, November 1999.

[16] N. Grbić, M. Dahl, and I. Claesson, “Neural network Based Adaptive Microphone Array System for Speech Enhancement,” in *IEEE World Congress on Computational Intelligence*, Anchorage, Alaska, USA, vol. 3, no. 5, pp. 2180–2183, May 1998.

[17] T. W. Lee, A. J. Bell,and R. rglmeister, “Blind Source Separation of Real World Signals,” in *IEEE International Conference in Neural Networks*, 1997.

[18] J. F. Cardoso, “Blind Source Separation: Statistical principles,” in *IEEE Proceedings, special issue on Blind System Identification and Estimation*, vol. 86, no. 10, pp. 2009–2025, October 1998.

[19] J. P. LeBlanc, and P. L. De Lon, “Speech Separation by Kurtosis Maximization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1029–1032, May 1998.

[20] N. Grbić, X. J. Tao, S. Nordholm, and I. Claesson, “Blind signal Separation Using Over-Complete Subband Representation,” in *IEEE Transaction on Speech and Audio Processing*, vol. 9, no. 5, pp. 524–533, July 2001.

[21] H. Saruwatari, S. Kajita, K. Takeda, and F. Itakura, "Speech Enhancement Using Nonlinear Microphone Array with Complementary Beamforming," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp.69–72, 1999.

[22] H. Gustafsson, S. Nordholm, and I. Claesson, "Spectral Substraction using Reduced Delay Convolution in Adaptive Averaging," in *IEEE Transactions on Speech and Audio Processing*, vol. 9, pp. 799–807, November 2001.

[23] J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Theoretical Noise Reduction Limits of the Generalized Sidelobe Canceler (GSC) for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 2965-2968, May 1999.

[24] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix using Constrained Adaptive Filters," in *IEEE Transactions on Signal Processing*, vol. 47, no. 10, pp. 2677-2684, June 1999.

[25] D. A. Florêncio, and H. S. Malvar, "Multichannel Filtering for Optimum Noise Reduction in Microphone Arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 197–200, May 2001.

[26] S. Affes and Y. Grenier, "A Signal Subspace Tracking Algorithm for Microphone Array Processing of Speech," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, no. 5, pp. 425–437, September 1997.

[27] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech Enhancement Based on the Subspace Method," *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 8, no. 5, pp. 497 – 507, September 2000.

[28] N. Grbić, "Optimal and Adaptive Subband Beamforming, Principles and Applications," Dissertation Series No 01:01, ISSN:1650-2159, Blekinge Institute of Technology, 2001.

[29] M. Brandstein and H. Silverman, "A practical Methodology for Speech Source Localization with Microphone Arrays," in *Computer, Speech and Language*, Vol. 11, pp. 91–126,April 1997.

[30] M. Brandstein and S. Griebel, “Time Delay Estimation of Reverberated Speech Exploiting Harmonic Structure,” in *Journal of Acoustic Society of America*, Vol. 105, no. 5, pp. 2914–2919, 1999.

[31] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, “Speaker localisation using Excitation Source Information in Speech,” *IEEE Transactions on Speech and Audio Processing.*

[32] M. Brandstein, and D. Ward (Eds.), “Microphone Arrays - Signal processing Techniques and applications,” Springer, 2001.

[33] L. Ziomek, “Fundamentals of Acoustic Field Theory and Space-Time Signal processing,” CRC Press, 1995.

[34] J. E. Hudson, “Array Signal Processing - Concepts and Techniques,” Prentice Hall, 1993.

[35] R. Zelinski, “A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms,” *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 2578 – 2581, 1988.

[36] C. Marro, Y. Mahieux, and K. Uwe Simmer, “Analysis of Noise Reduction and Dereverberation Techniques Based on Microphone Arrays with Post-Filtering,” *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 240 – 256, September 2000.

[37] O. Frost, “An Algorithm for Linearly Constrained Adaptive Array Processing,” IEEE Proceedings, vol. 60, 1972.

[38] S. Nordholm, I. Claesson, and M. Dahl, “Adaptive Microphone arrays Employing Calibration Signals. An analytical Evaluation,” in *IEEE Transaction on Speech and Audio Processing* vol. 7, no. 3, pp. 241–252, may 1999.

[39] S. Haykin, *Adaptive Filter Theory,* Prentice-Hall, 1996.

[40] P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Prentice-Hall, 1993.

[41] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, “Design of oversampled uniform dft filter banks with delay specifications using quadratic optimization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. VI, pp. 3633–3636, May 2001.

[42] N. Hirayama, H. Sakai, and S. Miyagi, "Delayless Subband Adaptive Filtering Using the Hadamard Transform," in *IEEE International Transactions on Signal Processing*, vol. 47, no. 6, pp. 1731–1736, June 1999.

[43] J. M. de Haan, L. O. Larson, I. Claesson, and S. Nordholm, "Filter Banks Design for Delayless Subband Adaptive Filtering Structures with Subband Weight Transformation," in *IEEE 10th Digital Signal Processing Workshop*, pp. 251–256, Pine Mountain, USA, October 2002.

[44] S. Bédard, B. champagne, and A. Stéphenne, "Effects of room reverberation on time-delay estimation performance," in *International Conference on Acoustics, Speech and Signal Processing*, vol. II, pp. 261–264, April 1994.

[45] N. Grbić, and S. Nordholm "Soft constrained subband beamforming for hands-free speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002, vol. I, pp. 885–888.

Part I

# A Constrained Subband Beamforming Algorithm for Speech Enhancement

**Part I is published as:**

Z. Yermeche, N. Grbić and I. Claesson, “A Constrained Subband Beamforming Algorithm for Speech Enhancement,” Research Report, ISSN: 1103-1581, December 2004.

Parts of this research report have been published as:

Z. Yermeche, P. Marquez, N. Grbić and I. Claesson, “A Calibrated Subband Beamforming Algorithm for Speech Enhancement,” published in Second IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, Washington DC, USA, August 2002.

# A Constrained Subband Beamforming Algorithm for Speech Enhancement

Z. Yermeche, N. Grbić and I. Claesson

## Abstract

This report presents a description and a study of a constrained subband beamforming algorithm constructed around the principle of an array calibration to the real acoustic environment. This method has been suggested for speech enhancement in hands-free communication, using an array of sensors. The proposed methodology is defined such to perform background noise and acoustic coupling reduction, while producing an undistorted filtered version of the signal originating from a desired location. The beamformer recursively minimizes a Least Squares error based on the continuously received data. This adaptive structure allows for tracking of the noise characteristics, such to accomplish its attenuation in an efficient manner. A soft constraint built from calibration data in high SNR conditions guarantees the integrity of the desired signal without the need of any speech detection. The computational complexity of the beamformer filters is substantially reduced by introducing a subband beamforming scheme.

This study includes an extensive evaluation of the proposed method in typical hands-free telephony environments, using real speech and noise recordings. Design issues of the subband beamformer are investigated and exploited in order to reach optimal usability. Measurements were performed in real acoustic environments, where the impact on the beamformer performance of different setups is considered. Results obtained in a crowded restaurant room as well as in a car cabin environment show a significant noise and hands-free interference reduction within the telephone bandwidth.

# 1 Introduction

Array processing involves the use of multiple sensors or transmitters to receive or transmit a signal carried by propagating waves. Sensor arrays have applications in a diversity of fields, such as telecommunications, sonar, radar and seismology [1, 2, 3, 4, 5, 6]. The focus of this report is on the use of microphone arrays to receive acoustic signals, and more specifically speech signals [7, 8, 9, 10]. The major applications for microphone arrays attempt to provide a good quality version of a desired speech signal, to localize the speech source or to identify the number of sources [7].

In the context of speech enhancement, microphone array processing has the potential to perform spatial selectivity, known also as directional hearing, via a technique known as beamforming, which reduces the level of directional and ambient noise signals, while minimizing distortion to speech from a desired direction. This technique is extensively exploited in hands-free communication technologies, such as video-conferencing, voice control and hearing-aids. In such environments, the transmitted speech signal is generated at a distance from the communication interface and thus it undergoes reverberations from the room response. Background noise and other interfering source signals also contribute to corrupt the signal actually conveyed to the far-end user.

This report presents a calibrated adaptive beamformer for speech enhancement, without the use of a voice activity detection (VAD), which was first introduced in [11]. Performance of the algorithm is evaluated on real data recordings conducted in different hands-free environments. The beamformer is based on the principle of a soft constraint, formed from calibration data, rather than precalculated from free-field assumptions as it is done in [12]. The benefit is that the real room acoustical properties will be taken into account. A subband beamforming implementation is chosen in order to allow the use of efficient, however, computationally demanding adaptive structures. A multichannel filter bank decomposes the input signal for each microphone into a set of narrow band signals, such to perform the beamformer filtering operations for each frequency subband separately. The full band output of the system is then reconstructed by a synthesis filter bank. The filter-banks are designed with the methodology described in [13], where in-band and reconstruction aliasing effects are minimized. The spatial characteristics of the input signal are maintained when using modulated filter-banks (analysis and synthesis), defined by two prototype filters, which leads to efficient polyphase realizations.

Information about the speech location is put into the algorithm in an initial acquisition by calculating source correlation estimates for microphone observations when the source signal of interest is active alone. The recording only needs to be done initially or whenever the location of interest is changed. The objective is formulated in the frequency domain as a weighted Recursive Least Squares (RLS) solution, which relies on the precalculated correlation estimates. In order to track variations in the surrounding noise environment, the adaptive beamformer continuously estimates the spatial information for each frequency band. The proposed algorithm updates the beamforming weights recursively where the initial precalculated correlation estimates constitutes a soft constraint. The soft constraint secures the spatial-temporal passage of the desired source signal, without the need of any speech detection.

Measurements were conducted in typical hands-free telephony environments, such as restaurant room and car cabin, where various setups were created. The choice of these environments was motivated by the extension of voice connectivity to personal computers, allowing the users to hold a distance-communication, in various environments such as offices, restaurants, trains, and other crowded public places, while being at a remote distance from the transmitting device, as well as by the automobile industry's effort to replace some hand-controlled functions with voice controls. The simulations were made with speech sequences from both male and female speakers. Results show a significant noise and interference reduction within the actual bandwidth. The influence of the design parameters on the performance of the proposed method was further investigated to achieve the optimal functionality of the proposed method.

# 2 Microphone Array Speech Enhancement

A microphone array consists of a set of acoustic sensors placed at different locations in order to spatially sample the sound pressure field. Hence, adaptive array processing of the spatial and temporal microphone samples allows time-variant control of spatial and spectral selectivity [1]. For instance, it allows us to separate signals that have overlapping frequency content but are originated from different spatial locations.

## 2.1 Signal Model

Consider an acoustic environment where a speech signal coexists with directional interfering signals (e.g. hands-free loudspeakers) and diffuse ambient noise. This sound field is observed by a microphone array with $I$ microphones, as depicted in Fig. 1.

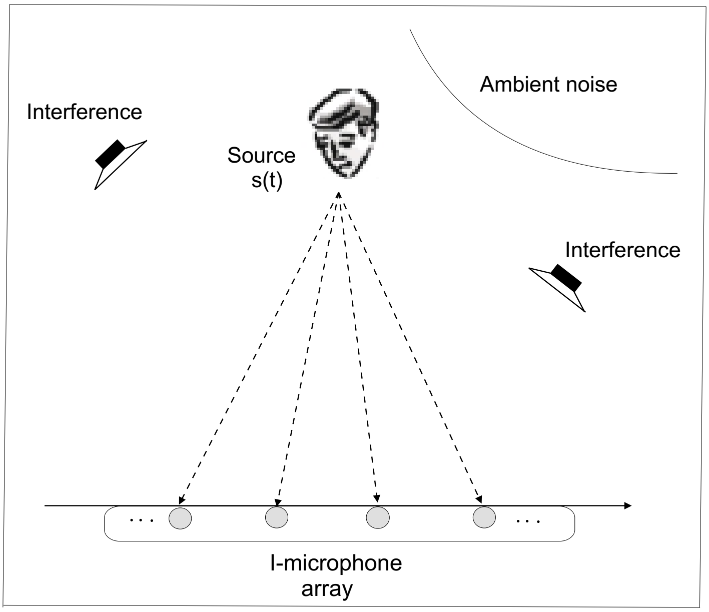


Figure 1: *Acoustic model.*

For a point source with a free-field propagation, the microphone input signal observed at the $i$th sensor, at time instant $t$, can be expressed as

$$x_{s,i}^{(\Omega)}(t) = a_i^{(\Omega)}\, s^{(\Omega)}(t - \tau_i), \tag{1}$$

where $s^{(\Omega)}(t)$ is the source signal component for the angular frequency $\Omega$, $\tau_i$ and $a_i^{(\Omega)}$ are the time-delay and the attenuation of the direct path from the point source to the $i$th sensor.

For a source located in the near-field of the array, the channel response $d_i^{(\Omega)}$ between the point source and the $i$th microphone can be expressed in complex-valued notation as

$$d_i^{(\Omega)} = a_i^{(\Omega)}\, e^{-j\Omega\tau_i} = \frac{1}{R_i}\, e^{-j\Omega\tau_i}, \tag{2}$$

where $R_i$ is the distance between the sound source and the sensor $i$ [10].

Using vector notation, Eq. (1) can be written as

$$\mathbf{x}_s^{(\Omega)}(t) = s^{(\Omega)}(t)\,\mathbf{d}^{(\Omega)}, \tag{3}$$

where

$$\mathbf{x}_s^{(\Omega)}(t) = [x_{s,1}^{(\Omega)}(t), \quad x_{s,2}^{(\Omega)}(t), \quad \ldots \quad, \quad x_{s,I}^{(\Omega)}(t)]^T$$

is the received microphone input vector for the point source signal and where the response vector elements are arranged as

$$\mathbf{d}^{(\Omega)} = [d_1^{(\Omega)}, \quad d_2^{(\Omega)}, \quad \ldots \quad, \quad d_I^{(\Omega)}]^T.$$

As multiple sources radiate in a common field, the corresponding propagating waves occur simultaneously without interaction, allowing for the superposition principle to apply. Consequently, the array input vector when all the sources are active simultaneously can be expressed as

$$\mathbf{x}^{(\Omega)}(t) = \mathbf{x}_s^{(\Omega)}(t) + \mathbf{x}_i^{(\Omega)}(t) + \mathbf{x}_n^{(\Omega)}(t), \tag{4}$$

where $\mathbf{x}_i^{(\Omega)}(t)$ and $\mathbf{x}_n^{(\Omega)}(t)$ are the received microphone input vectors generated by the interfering sources and the ambient noise, respectively.

## 2.2 Optimal Beamformer

The beamformer optimizes the array output by adjusting the weights of finite length digital filters so that the combined output contains minimal contribution from noise and interference. Consequently, the angle of the spatial passband is adjusted for each frequency. In a typical hands-free scenario, high order Finite Impulse Response (FIR) filters are required to achieve a reasonably good speech extraction, especially when it involves room reverberation suppression as well. Thus, in order to reduce the computational complexity and improve the overall performance of the filter, a subband beamforming structure is used. Each microphone input signal is first decomposed into narrow band signals, and then the filtering process is applied to each frequency subband.

### 2.2.1 Spatial Correlation

For the input vector $\mathbf{x}^{(\Omega)}(n)$ at discrete-time instant $n$, containing mainly frequency components around the center frequency $\Omega$, the spatial correlation matrix is given by

$$\mathbf{R}^{(\Omega)} = E[\mathbf{x}^{(\Omega)}(n)\,\mathbf{x}^{(\Omega)^H}(n)]. \tag{5}$$

The symbol $(^H)$ denotes the Hermitian transpose. Assuming that the speech signal, the interference and the ambient noise are uncorrelated, $\mathbf{R}^{(\Omega)}$ can be written as

$$\mathbf{R}^{(\Omega)} = \mathbf{R}_{ss}^{(\Omega)} + \mathbf{R}_{ii}^{(\Omega)} + \mathbf{R}_{nn}^{(\Omega)}, \tag{6}$$

where $\mathbf{R}_{ss}^{(\Omega)}$ is the source correlation matrix, $\mathbf{R}_{ii}^{(\Omega)}$ is the interference correlation matrix and $\mathbf{R}_{nn}^{(\Omega)}$ is the noise correlation matrix for frequency $\Omega$ defined as

$$\begin{aligned}
\mathbf{R}_{ss}^{(\Omega)} &= E[\mathbf{x}_s^{(\Omega)}(n)\,\mathbf{x}_s^{(\Omega)^H}(n)],\\
\mathbf{R}_{ii}^{(\Omega)} &= E[\mathbf{x}_i^{(\Omega)}(n)\,\mathbf{x}_i^{(\Omega)^H}(n)],\\
\mathbf{R}_{nn}^{(\Omega)} &= E[\mathbf{x}_n^{(\Omega)}(n)\,\mathbf{x}_n^{(\Omega)^H}(n)].
\end{aligned}$$

### 2.2.2 Wiener Solution

The optimal filter weight vector based on the Wiener solution [1] is given by

$$\mathbf{w}_{opt}^{(\Omega)} = \left[\mathbf{R}^{(\Omega)}\right]^{-1} \mathbf{r}_s^{(\Omega)}, \tag{7}$$

where the array weight vector, $\mathbf{w}_{opt}^{(\Omega)}$ is arranged as

$$\mathbf{w}_{opt}^{(\Omega)} = [w_1^{(\Omega)}, \quad w_2^{(\Omega)}, \quad \ldots \quad , w_I^{(\Omega)}]$$

and where $\mathbf{r}_s^{(\Omega)}$ is the cross-correlation vector defined as

$$\mathbf{r}_s^{(\Omega)} = E[\mathbf{x}_s^{(\Omega)}(n)\, s^{(\Omega)^H}(n)]. \tag{8}$$

The signal $s^{(\Omega)}(n)$ is the desired source signal at time sample $n$. The output of the beamformer is given by

$$y^{(\Omega)}(n) = \mathbf{w}_{opt}^{(\Omega)^H} \mathbf{x}^{(\Omega)}(n). \tag{9}$$

## 2.3 An Adaptive Structure

When working in a non-stationary environment, the weights of the beamformer are calculated adaptively in order to follow the statistical variations of the observable data. A fixed location of the target signal source always excites the same correlation patterns between microphones, and it therefore becomes spatially stationary. Hence, the corresponding statistics are constant and can be estimated from a data sequence gathered in an initial acquisition. In a similar manner, statistics for directional interfering sources can be initially estimated from calibration data. Consequently, the adaptive algorithm reduces to a time-varying filter, tracking the behavior of the noise, in order to suppress it.

The implementation of the Least Mean Squares (LMS) algorithm in such structure requires the use of large buffers to memorize source and interfering input data vectors, as has been suggested in [14]. Designing an adaptive beamformer based on the RLS algorithm, on the other hand, leads to saving in memory correlation matrices instead, as shown in [11], which results in significantly less memory usage.

In general, the RLS algorithm also offers a better convergence rate, mean-square error (MSE) and parameter tracking capabilities in comparison to the LMS algorithm [15]. Additionally, the LMS algorithm often exhibits inadequate performance in the presence of high power background noise [17].

The widespread acceptance of the RLS algorithm is nevertheless impeded by the numerical instability displayed when the input covariance matrix is close to singular [16]. By forcing the full rank of the input covariance matrix, this problem can be overcome. Experience, however, shows that the full rank of the matrix is most commonly ensured.

# 3 The Constrained RLS Beamformer

The constrained beamformer is based on the idea proposed in [11, 12]. In an initial acquisition, a calibration sequence emitted from the target source position and gathered in a quiet environment, is used to calculate the source statistics. Further, since in a realistic scenario the reference source signal information is not directly available, the received signal input of a selected sensor, with index $r$, is used instead. This calibration signal carries the temporal and spatial information about the source. Also, the interference statistics are calculated from a calibration signal sequence gathered when all the known directional interference sources are active simultaneously.

In order to track variations in the surrounding noise field, the proposed algorithm continuously estimates the spatial information of the acoustical environment and the update of the beamformer weights is done recursively where the initial precalculated correlation estimates constitute a soft constraint.

## 3.1 Least Squares Formulation

The objective in the proposed method is formulated as a Least Squares (LS) solution. The optimal weight vector $\mathbf{w}_{ls}^{(\Omega)}(n)$ at sample instant $n$ is given by

$$\mathbf{w}_{ls}^{(\Omega)}(n) = \left[\hat{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)} + \hat{\mathbf{R}}_{xx}^{(\Omega)}(n)\right]^{-1} \hat{\mathbf{r}}_{s}^{(\Omega)}, \tag{10}$$

where the source correlation estimates, i.e. the correlation matrix estimate, $\hat{\mathbf{R}}_{ss}^{(\Omega)}$, and the cross correlation vector estimate, $\hat{\mathbf{r}}_{s}^{(\Omega)}$, are pre-calculated in a calibration phase. For a data set of $N$ samples

$$\hat{\mathbf{R}}_{ss}^{(\Omega)} = \sum_{p=0}^{N-1} \mathbf{x}_{s}^{(\Omega)}(p)\ \mathbf{x}_{s}^{(\Omega)^{H}}(p), \tag{11}$$

$$\hat{\mathbf{r}}_{s}^{(\Omega)} = \sum_{p=0}^{N-1} \mathbf{x}_{s}^{(\Omega)}(p)\ x_{s,r}^{(\Omega)^{*}}(p), \tag{12}$$

where

$$\mathbf{x}_{s}^{(\Omega)}(p) = [x_{s,1}^{(\Omega)}(p), \quad x_{s,2}^{(\Omega)}(p), \quad \dots \quad x_{s,I}^{(\Omega)}(p)]^{T}$$

are digitally sampled microphone observations when the source signal of interest is active alone and $x_{s,r}^{(\Omega)}(p)$ constitutes the chosen reference signal.

In a similar manner the interference correlation matrix estimate, $\hat{\mathbf{R}}_{ii}^{(\Omega)}$, is pre-calculated for a data set of $N$ samples when all known disturbing sources are active alone, by

$$\hat{\mathbf{R}}_{ii}^{(\Omega)} = \sum_{p=0}^{N-1} \mathbf{x}_i^{(\Omega)}(p)\ \mathbf{x}_i^{(\Omega)^H}(p), \tag{13}$$

where

$$\mathbf{x}_i^{(\Omega)}(p) = [x_{i,1}^{(\Omega)}(p), \quad x_{i,2}^{(\Omega)}(p), \quad \ldots \quad x_{i,I}^{(\Omega)}(p)]^T.$$

Conversely, the correlation estimates, $\hat{\mathbf{R}}_{xx}^{(\Omega)}(n)$, are continuously calculated from observed data by

$$\hat{\mathbf{R}}_{xx}^{(\Omega)}(n) = \sum_{p=0}^{n} \lambda^{n-p}\mathbf{x}^{(\Omega)}(p)\ \mathbf{x}^{(\Omega)^H}(p), \tag{14}$$

where

$$\mathbf{x}^{(\Omega)}(p) = [x_1^{(\Omega)}(p), \quad x_2^{(\Omega)}(p), \quad \ldots \quad x_I^{(\Omega)}(p)]^T$$

and $\lambda$ is a forgetting factor, with the purpose of allowing for tracking variations in the surrounding noise environment.

## 3.2 Recursive Formulation

Given that the time-dependent parameter on the right side of Eq. (10) can be expressed recursively based on the available input data vector, $\mathbf{x}^{(\Omega)}(n)$ as

$$\hat{\mathbf{R}}_{xx}^{(\Omega)}(n) = \lambda\hat{\mathbf{R}}_{xx}^{(\Omega)}(n-1) + \mathbf{x}^{(\Omega)}(n)\mathbf{x}^{(\Omega)^H}(n), \tag{15}$$

a recursive solution for the update of the beamforming weight vector, $\mathbf{w}_{ls}^{(\Omega)}(n)$, is derived.

Since we are interested in the inverse of

$$\begin{aligned}\hat{\mathbf{R}}^{(\Omega)}(n) &= \hat{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)} + \hat{\mathbf{R}}_{xx}^{(\Omega)}(n) \\ &= \lambda\hat{\mathbf{R}}^{(\Omega)}(n-1) + \mathbf{x}^{(\Omega)}(n)\mathbf{x}^{(\Omega)^H}(n) + (1-\lambda)(\hat{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)}),\end{aligned} \tag{16}$$

the inversion of the total sum is required, including the precalculated correlation matrices, $\hat{\mathbf{R}}_{ss}^{(\Omega)}$ and $\hat{\mathbf{R}}_{ii}^{(\Omega)}$. An alternative approach in order to reduce the complexity of the problem is to simplify the representation of these matrices in Eq. (16), by applying the spectral theorem, to the form

$$\hat{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)} = \sum_{p=1}^{P} \gamma_p^{(\Omega)} \mathbf{q}_p^{(\Omega)} \mathbf{q}_p^{(\Omega)^H}, \tag{17}$$

where $\gamma_p^{(\Omega)}$ is the $p$-th eigenvalue and $\mathbf{q}_p^{(\Omega)}$ is the $p$-th eigenvector of the $I$-by-$I$ calibration correlation matrix sum, $\hat{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)}$, and $P$ is the dimension of the signal space, i.e. the effective rank of the matrix. This will result in adding scaled eigenvectors of the calibration correlation matrix sum to the update of Eq. (16), corresponding to several rank-one updates. Also, by sequentially adding one scaled eigenvector at each sample instant $n$, the complexity is further reduced while only affecting the scale of the problem, obtaining the following expression for the update of $\hat{\mathbf{R}}^{(\Omega)}(n)$,

$$\begin{aligned}\hat{\mathbf{R}}^{(\Omega)}(n) &= \lambda \hat{\mathbf{R}}^{(\Omega)}(n-1) + \mathbf{x}^{(\Omega)}(n)\mathbf{x}^{(\Omega)^H}(n) \\ &+ (1-\lambda)\left[\gamma_p^{(\Omega)} \mathbf{q}_p^{(\Omega)} \mathbf{q}_p^{(\Omega)^H}\right]_{p=n(modP)+1}\end{aligned} \tag{18}$$

where the notation *mod* represents the modulus function. This allows for the use of the Woodbury's identity in the inversion of $\hat{\mathbf{R}}^{(\Omega)}(n)$.

Since the statistical properties of the environmental noise can change abruptly, a smoothing of the weights may be appropriate. A first order AR-model for the smoothing with parameter $\eta$ is used and the weight update then becomes

$$\mathbf{w}_{ls}^{(\Omega)}(n) = \eta \mathbf{w}_{ls}^{(\Omega)}(n-1) + (1-\eta)[\hat{\mathbf{R}}^{(\Omega)}(n)]^{-1} \hat{\mathbf{r}}_s^{(\Omega)}. \tag{19}$$

## 3.3 Time-Frequency Filtering

In the beamforming algorithm previously described, the number of subbands is proportional to the length of the equivalent time-domain filters. The number of subbands is therefore the parameter controlling the temporal resolution of the algorithm. Each subband signal can be considered as a narrow band time signal sampled at a reduced sampling rate. Hence, a combined time-frequency filtering structure can be constructed by using consecutive time samples in the representation of the input vector $\mathbf{x}^{(\Omega)}(n)$ to compute the beamformer output of Eq. (9), according to

$$\mathbf{x}^{(\Omega)}(n) = [\mathbf{x}_1^{(\Omega)}(n), \quad \mathbf{x}_2^{(\Omega)}(n), \quad \dots \quad, \quad \mathbf{x}_I^{(\Omega)}(n)]^T,$$

with

$$\mathbf{x}_i^{(\Omega)}(n) = [x_i^{(\Omega)}(n), \quad x_i^{(\Omega)}(n-1), \quad \dots \quad, \quad x_i^{(\Omega)}(n-L_{sub}+1)],$$

for $i = 1, \dots, I$. This representation allows us to introduce an additional parameter, the subband filter length $L_{sub}$, controlling the algorithm's degrees of freedom.

Consequently, in the time-frequency representation, the weight vector for each subband is similarly extended by

$$\mathbf{w}^{(\Omega)} = [\mathbf{w}_1^{(\Omega)}, \quad \mathbf{w}_2^{(\Omega)}, \quad \dots \quad, \quad \mathbf{w}_I^{(\Omega)}]^T,$$

with

$$\mathbf{w}_i^{(\Omega)} = [w_{i,0}^{(\Omega)}, \quad w_{i,1}^{(\Omega)}, \quad \dots \quad, \quad w_{i,L_{sub}-1}^{(\Omega)}].$$

The size of the correlation matrices and vectors, eigenvectors and number of eigenvalues defined in previous section is correspondingly increased by a factor of $L_{sub}$.

## 4 Subband Beamforming

The broadband input signals are decomposed into sets of narrow band signals, such to perform the filtering operations on narrow band signals individually, requesting significantly smaller filters. With $K$ denoting the total number of subbands, the subband signals each have a bandwidth that is approximately $K$ times smaller in width than that of the full band input signal. This allows for the use of up till $K$ times lower sample rates and therefore reduces considerably the complexity of the overall filtering structure [18]. However, in order to reduce the aliasing between the subbands, it is preferable to achieve an over-sampled subband decomposition by using a down-sampling factor, $D$, also known as decimation factor, smaller than the number of subbands, $K$. This in fact means that more samples are carried by all the subband signals together than the original full band signal.

Fig. 2 illustrates the overall architecture of the microphone array speech enhancement system, based on the constrained adaptive subband beamformer.

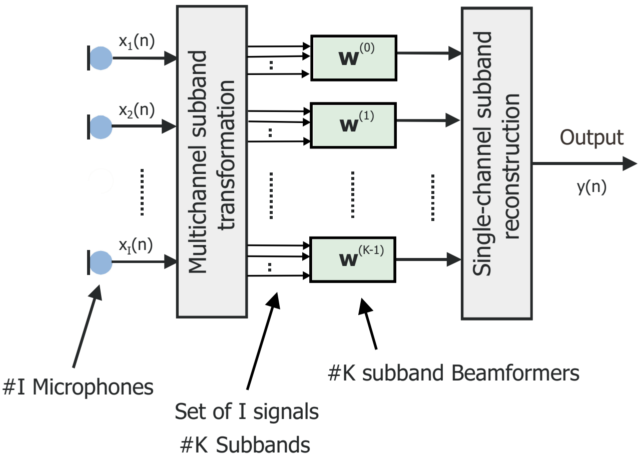


Figure 2: *Structure of the subband beamformer.*

The structure includes a multichannel analysis filter-bank used to decompose the received array signals into a set of subband signals, and a set of adaptive beamformers each adapting on the multichannel subband signals. The outputs of the beamformers are reconstructed by a synthesis filter bank in order to create a time-domain output signal.

## 4.1 Modulated Filter Banks

The spatial characteristics of the input signal are maintained when using the same modulated filter bank for all microphone signals [11]. The structure of the synthesis and analysis modulated filter bank is presented in Fig. 3. Modulated filter banks are defined by a low-pass prototype filter, $H_0(z)$, to which all filters, $H_k(z)$, are related by modulation,

$$H_k(z) = H_0(zW_K^k), \tag{20}$$

where $W_K = e^{\frac{-j2\pi}{K}}$. This definition holds for the synthesis modulated filters, $F_k(z)$, as well, which correspondingly are related to a synthesis prototype filter, $F_0(z)$. In other words, the filter bank consists of a set of frequency-shifted versions of the low-pass prototype filter, with each filter being centered at the frequencies $\frac{2\pi k}{K}$, $k = 0, ..., K-1$, covering the whole spectrum range. Fig. 4 shows the frequency responses for the analysis filters, when $K = 4$.

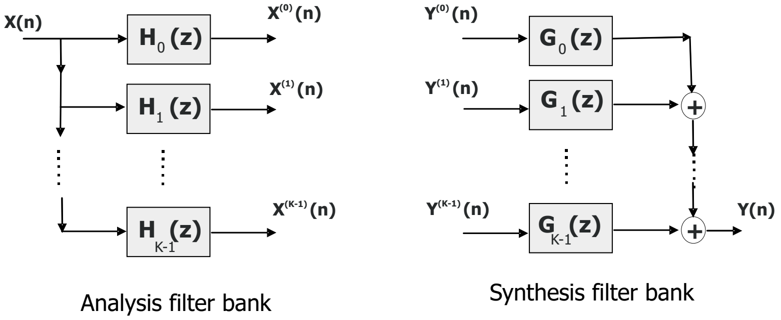


Figure 3: *Analysis and synthesis filter banks.*

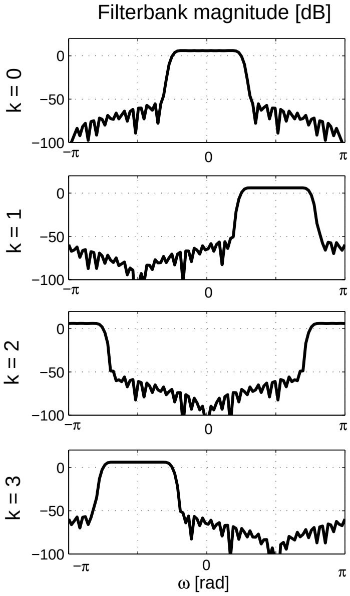


Figure 4: *Typical analysis modulated filter bank magnitude response for a number of subbands $K = 4$. The prototype filter used is a low pass Hamming window of length $L = 64$, and cutoff frequency $\pi/K$.*

## 4.2 The Polyphase Filter Bank Implementation

When the prototype filter of the modulated filter bank is an FIR filter the polyphase decomposition can be used to implement such a filter bank in an efficient manner [18]. The structure used in this evaluation is the polyphase filter bank realization factor-two-oversampled ($D = K/2$), in which the number of polyphase decompositions is chosen to be equal to the decimation factor, $D$.

### 4.2.1 Analysis filter bank structure

The polyphase decomposition of the analysis prototype filter, $H_0(z)$, is given by

$$H_0(z) = \sum_{n=-\infty}^{+\infty} h_0(n) z^{-n} = \sum_{l=0}^{D-1} z^{-l} E_l(z^D), \tag{21}$$

where $h_0(n)$ are the weights of the FIR prototype filter, and the type 1 polyphase components, $E_l(z)$, are given by

$$E_l(z) = \sum_{n=-\infty}^{+\infty} h_0(Dn + l) z^{-n}. \tag{22}$$

The type 1 polyphase decomposition of the prototype filter is presented in Appendix A. The $k$-th filter, $H_k(z)$, is then decomposed into $D$ polyphase components as

$$H_k(z) = \sum_{l=0}^{D-1} (zW_K^k)^{-l} E_l([zW_K^k]^D) = \sum_{l=0}^{D-1} z^{-l} W_K^{-kl} E_l(z^D W_K^{kD}). \tag{23}$$

Since

$$W_K^{kD} = e^{-j\pi k} = \begin{cases} +1, & \text{when } k \text{ is even,} \\ -1, & \text{when } k \text{ is odd,} \end{cases}$$

the decomposition of Eq. (23) is done separately for the analysis filters in the even-indexed subbands and the analysis filters in the odd-indexed subbands as

$$H_k(z) = \sum_{l=0}^{D-1} z^{-l} W_K^{-kl} E_l(z^D) \qquad \text{for } k \text{ even,} \tag{24}$$

$$H_k(z) = \sum_{l=0}^{D-1} z^{-l} W_K^{-kl} E_l^{'}(z^D) \qquad \text{for } k \text{ odd}, \tag{25}$$

where the polyphase components for even-indexed filters, $E_l(z)$, are defined in Eq. (22), and the polyphase components for odd-indexed filters, $E_l^{'}(z)$, are defined by

$$E_l^{'}(z) = \sum_{n=-\infty}^{+\infty} h_0(Dn+l)(-1)^n z^{-n}. \tag{26}$$

The analysis filter bank outputs are decimated by factor $D$ to insure a good trade-off between efficient implementation and a low level of aliasing between subbands. By applying the noble identity [18], the decimation operation can be performed prior to the filtering by the polyphase filters, leading to the filter $E_l(z^D)$ in Eq. (24) being replaced by $E_l(z)$ (correspondingly, $E_l^{'}(z^D)$ in Eq. (25) is replaced by $E_l^{'}(z)$ ).

Furthermore, since

$$W_K^{-kl} = e^{\frac{j2\pi kl}{K}} = \begin{cases} e^{\frac{j2\pi \frac{k}{2} l}{D}} & \text{when } k \text{ is even}, \\ e^{\frac{j2\pi \frac{k-1}{2} l}{D}} e^{\frac{j2\pi l}{K}} & \text{when } k \text{ is odd}, \end{cases}$$

for $l = 0, ..., D-1$. The summation with coefficients $W_K^{-kl}$ can be implemented efficiently using a D-length FFT operator for the even subbands and a D-length FFT operator preceded by a multiplication with the corresponding factor $e^{\frac{j2\pi l}{K}}$ for the odd subbands.

The $K$-analysis filter bank can consequently be implemented by using the structure illustrated in Fig. 5.

### 4.2.2 Synthesis filter bank structure

Similarly, if the synthesis prototype filter, $G_0(z)$, is an FIR filter defined by

$$G_0(z) = \sum_{n=-\infty}^{+\infty} g_0(n) z^{-n}, \tag{27}$$

the type 2 polyphase decomposition, with $D$ elements, of the synthesis filter in subband $k$ is

$$G_k(z) = \sum_{l=0}^{D-1} z^{-(D-l-1)} W_K^{kl} F_l(z^D) \qquad \text{for } k \text{ even}, \tag{28}$$

$$G_k(z) = \sum_{l=0}^{D-1} z^{-(D-l-1)} W_K^{kl} F_l^{'}(z^D) \qquad \text{for } k \text{ odd}, \tag{29}$$

where the type 2 polyphase components are given by

$$F_l(z) = \sum_{n=-\infty}^{+\infty} g_0(Dn - l - 1) z^{-n}, \tag{30}$$

$$F_l^{'}(z) = \sum_{n=-\infty}^{+\infty} g_0(Dn - l - 1)(-1)^n z^{-n}. \tag{31}$$

The type 2 polyphase derivation is presented in Appendix B.

The full-band output signal, $Y(z)$, of the synthesis filter bank can be expressed in terms of the interpolated subband signals, $Y^{(k)}(z)$, for $k = 0, ..., K-1$, corresponding to the synthesis filter bank inputs according to

$$\begin{aligned} Y(z) = &\sum_{\text{k even}} \sum_{l=0}^{D-1} W_K^{kl} Y^{(k)}(z) F_l(z^D) z^{-(D-l-1)} \\ &+ \sum_{\text{k odd}} \sum_{l=0}^{D-1} W_K^{kl} Y^{(k)}(z) F_l^{'}(z^D) z^{-(D-l-1)} \\ = &\sum_{l=0}^{D-1} \left[ \sum_{\text{k even}} W_K^{kl} Y^{(k)}(z) F_l(z^D) + \sum_{\text{k odd}} W_K^{kl} Y^{(k)}(z) F_l^{'}(z^D) \right] z^{-(D-l-1)}. \end{aligned} \tag{32}$$

Since the subband signals are interpolated with factor $D$, the noble identity can be invoked to simplify the implementation by applying the polyphase components prior to the interpolation operators [18].

As with the analysis filter bank an efficient implementation using the FFT algorithm can be deduced, which is illustrated in Fig. 6.

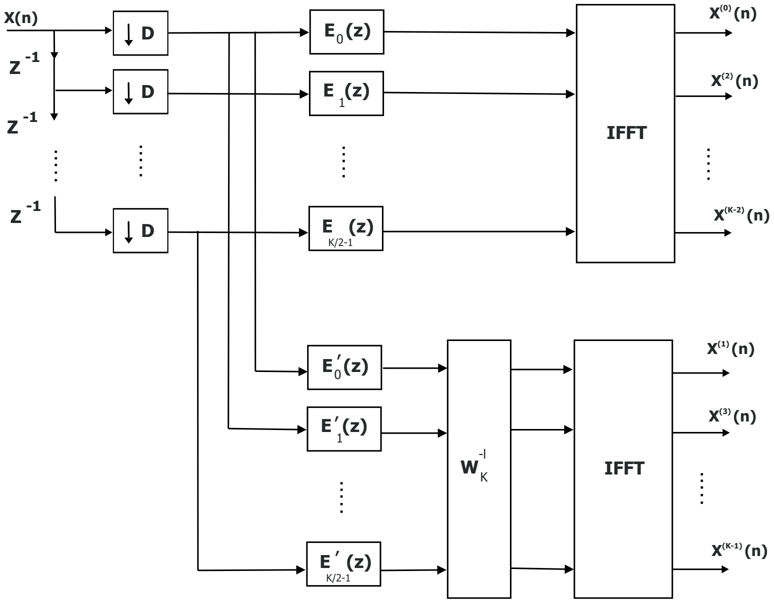


Figure 5: *Polyphase implementation of a two-times over-sampled uniform analysis filter bank.*

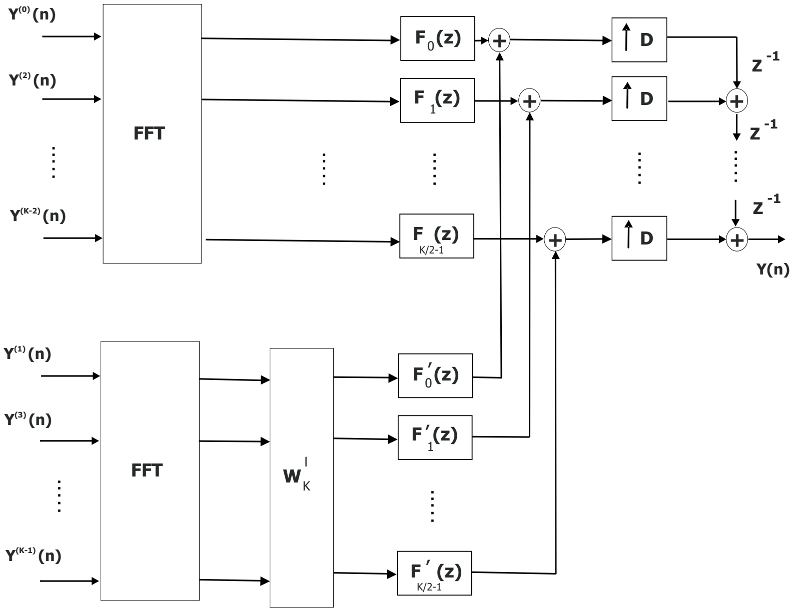


Figure 6: *Polyphase implementation of a two-times over-sampled uniform synthesis filter bank.*

# 5 Algorithm Implementation

The array sensor input signals are sampled with a frequency $F_s$ and decomposed each into a set of $K$ corresponding subband signals. These narrow-band signals constitute the inputs to a set of $K$ subband beamformers.

We assume that the source signal and the additional known disturbing sources with fixed location (e.g. fixed loudspeakers) are available during a training phase preceding the online processing. Consequently, the estimated source correlation matrix, $\hat{\mathbf{R}}_{ss}^{(k)}$, the estimated source cross-correlation vector, $\hat{\mathbf{r}}_s^{(k)}$, and the estimated interference correlation matrix , $\hat{\mathbf{R}}_{ii}^{(k)}$, are made available from this initial acquisition, for each subband indexed $k = 0, 1, ..., K-1$. The index $k$ refers to the frequency subband centered at the angular frequency $\Omega = 2\pi F_s k/K$.

During the online processing the algorithm is stated as an iterative procedure, individually for each subband. It is run sequentially with the steps described in the operation phase below.

## Calibration phase:

- Calculate the estimated source correlation matrix $\hat{\mathbf{R}}_{ss}^{(k)}$, and the estimated cross correlation vector, $\hat{\mathbf{r}}_s^{(k)}$, according to Eqs. (11) and (12) when the source of interest is active alone.

- Calculate the observed data correlation matrix when known disturbing sources are active alone, i.e. the interference correlation matrix $\hat{\mathbf{R}}_{ii}^{(k)}$, according to Eqs. (13).

- The correlation matrices are saved in memory in a diagonalized form:

$$\mathbf{Q}^{(k)^H}\mathbf{\Gamma}^{(k)}\mathbf{Q}^{(k)} = \left(\alpha\hat{\mathbf{R}}_{ss}^{(k)} + \beta\hat{\mathbf{R}}_{ii}^{(k)}\right)$$

  The eigenvectors are denoted:

$$\mathbf{Q}^{(k)} = [\mathbf{q}_1^{(k)}, \quad \mathbf{q}_2^{(k)}, \quad \dots \quad \mathbf{q}_I^{(k)}]$$

  The eigenvalues are denoted:

$$\mathbf{\Gamma}^{(k)} = diag([\gamma_1^{(k)}, \quad \gamma_2^{(k)}, \quad \dots \quad \gamma_I^{(k)}])$$

The parameters $\alpha$ and $\beta$ are weighting factors for the precalculated correlation estimates, controlling the relative amount of sources amplification/attenuation. They are chosen in accordance to the requirements of the application, and they provide means for trading the level of interference suppression with the level of speech distortion.

- Initialize the weight vector ,$\mathbf{w}_{ls}^{(k)}(n)$, as a zero vector.
- Initialize the inverse of the total correlation matrix, $\hat{\mathbf{R}}^{(k)^{-1}}$, denoted as $\mathbf{P}^{(k)}(0) = \sum_{p=1}^{P} \gamma_p^{(k)^{-1}} \mathbf{q}_p^{(k)} \mathbf{q}_p^{(k)^H}$, and define the same size dummy variable matrix, $\mathbf{D}$.
- Choose a forgetting factor, $0 < \lambda < 1$, and a weight smoothing factor, $0 < \eta < 1$.

## Operation phase:

For n =1, 2, ...

- When any of the sources are active simultaneously, update the inverse total correlation matrix, as

$$\mathbf{D} = \lambda^{-1}\mathbf{P}^{(k)}(n-1) - \frac{\lambda^{-2}\mathbf{P}^{(k)}(n-1)\mathbf{x}(n)^{(k)}\mathbf{x}(n)^{(k)^H}\mathbf{P}^{(k)}(n-1)}{1+\lambda^{-1}\mathbf{x}(n)^{(k)^H}\mathbf{P}^{(k)}(n-1)\mathbf{x}(n)^{(k)}}$$

$$\mathbf{P}^{(k)}(n) = \mathbf{D} - \frac{\gamma_p(1-\lambda)\mathbf{D}\mathbf{q}_p^{(k)}\mathbf{q}_p^{(k)^H}\mathbf{D}}{1+\gamma_p(1-\lambda)\mathbf{q}_p^{(k)^H}\mathbf{D}\mathbf{q}_p^{(k)}}$$

  where the index of the eigenvalues and eigenvectors[1] is $p = n(mod\, I)+1$.

- Calculate the weights, for each sample instant, as

$$\mathbf{w}(n)^{(k)} = \eta\mathbf{w}(n-1)^{(k)} + (1-\eta)\mathbf{P}^{(k)}(n)\hat{\mathbf{r}}_s^{(k)}$$

- Calculate the output for the subband $k$ as:

$$y^{(k)}(n) = \mathbf{w}(n)^{(k)^H}\mathbf{x}(n)^{(k)}$$

The output from all subband beamformers are used in the reconstruction filter bank to create the time-domain output.

[1] In this way, we insure the full rank of the total correlation matrix by taking into account the contribution from all $I$ eigenvectors successively, regardless of the eigenvalue spread.

# 6 Evaluation Conditions and Performance Measures

In order to evaluate the proposed beamforming method, recordings where performed in real-world environments. The performance measures presented in Sec. 6.3 are based on these real speech and noise recordings.

## 6.1 Equipment and Settings

The recording system is illustrated in Fig. 7. A generator is used to produce different acoustic sequences. For speech utterances, a CD player is utilized as generator to play high SNR speech sentences previously recorded on a disc. The acoustical source is embodied by a loudspeaker receiving the generator output and is moreover used to simulate a real person speaking. The acoustical sensor-array consists of a (commercialized) array of four microphones. Due to the existence of a 2 V DC component at each single microphone observation, a pre-filtering have been used to eliminate it. The array output data is gathered on a multichannel DAT-recorder with a sampling rate $F_s = 12$ kHz and with a 5 kHz bandwidth for each channel, in order to avoid temporal aliasing.

The reference microphone observation is chosen to be microphone number two in the array, throughout the evaluation.

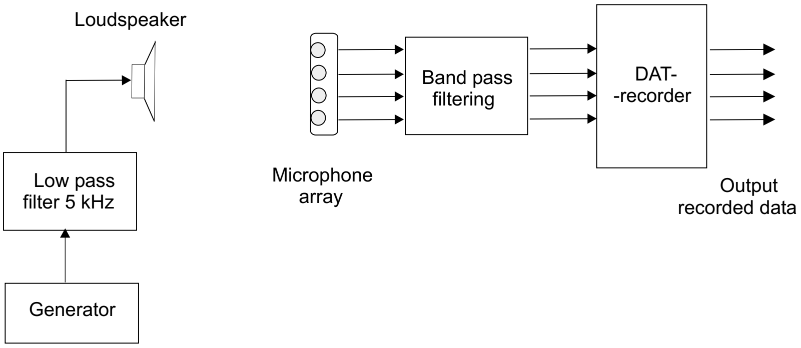


Figure 7: *Data-recording equipment setting.*

### 6.1.1 Microphone Configuration

The used microphone array in this evaluation is the DA-400 2.0 Desktop Array from Andrea Electronics, which is an uniformly spaced linear array based on four sensors. The sensors are pressure gradient microphone elements with 5 cm spacing. Andrea recommends to place the microphone unit approximately at eye level and to maintain a distance of 45 cm to 60 cm (optimum operating distance), when using the DA-400 incorporated beamformer, as it is a device built to sit on top of a computer monitor.

## 6.2 Environment Configuration

For evaluation purposes, recordings where performed in different environments. For each environment several scenarios were defined corresponding to different settings of position and type of sound for the target, the interference and the ambient noise sources.

### 6.2.1 Isolated-room Environment

Recordings were first carried out in an isolated-room with hard walls. All acoustical sources were simulated and therefore have predefined known characteristics. Thus, recordings made in this controlled environment were used to investigate the optimal working conditions of the algorithm.

The isolated-room environment is shown in Fig. 8, where the - simulated - sound sources are represented by the symbols of loudspeaker. Every scenario is defined by the distance and the angle formed by the - target and interfering - speakers, placed at the microphone height, and the sensor array. A surrounding diffuse noise is simulated by four loudspeakers situated at the corners of the room. Initially, recordings with white noise signals and real speech signals, from both the target and the interfering loudspeakers, were recorded individually and used as calibration signals for the target source position and the interfering source position, respectively. Real speech sequences emitted by the artificial speakers and recorded individually serve as performance measure signals. In order to evaluate the performance of the beamformer under different noise conditions, noise, speech and music signals, emitted by the four surrounding loudspeakers, were recorded simultaneously. It should be noted that the surrounding noise sources properties and placement are unknown, in regard to the beamformer.

### 6.2.2 Restaurant Environment

Recordings were performed in a restaurant room in order to evaluate the algorithm performance in a crowded environment. The recording equipment was situated in a corner of the room of size $[5 \times 10 \times 3]m$. The target and interfering speakers were simulated by two loudspeakers, as in the isolated-room environment case, while the surrounding noise consisted of the ambient noise recorded at busy hours. The restaurant noise environment consists of a number of sound sources, mostly persons holding discussions, moving objects such as chairs and colliding items, e.g. glasses and plates. Recordings of both the target source speech and the interference source speech were made individually in a silent room. These recordings serve as calibration signals respectively for each one of the source positions.

### 6.2.3 Car Environment

The performance of the beamformer was evaluated in a car hands-free telephony environment with the linear microphone array mounted on the visor at the driver's side, see Figs. 9 and 10. The measurements were performed in a Volvo station wagon where the speech originating from the driver's position constitutes the desired source signal. The microphone-array was positioned at a distance of 35 cm from the speaker. A loudspeaker was mounted at the passenger seat to simulate a real person engaging a conversation. In some scenarios, the speaker on the passenger side is regarded as an interfering source. It is often convention in a car hands-free installation to use the existing audio system for the far-end speaker. Thus, two loudspeakers were positioned at the back of the car, to simulate loudspeakers commonly placed at this location. Speech signals emanating from the driver's seat are recorded in a non-moving car with the engine turned off, and used as target source calibration signals. Similarly, speech sequences emitted from the artificial talkers in a silent environment and recorded individually serve as calibration signals for the corresponding positions.

In order to gather background noise signals, recordings were made when the car was running at 100 km/hour on a normal paved road. The car cabin noise environment consists of a number of unwanted sound sources, mostly with a broad spectral content, e.g. wind and tire friction as well as engine noise.

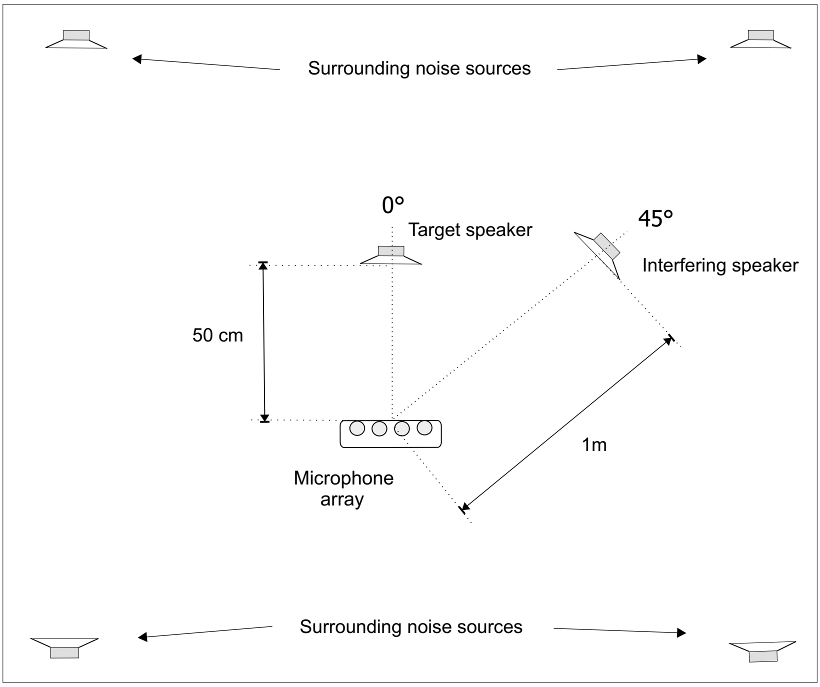


Figure 8: *Typical scenario setting for recordings performed in the isolated-room environment. Loudspeakers located at different positions relatively to the microphone array simulate the sound sources simultaneously active.*

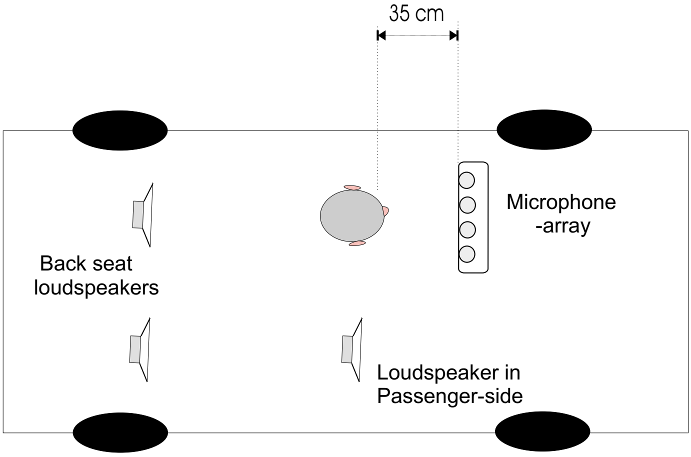


Figure 9: *Placement of microphone array and loudspeakers in the car cabin.*

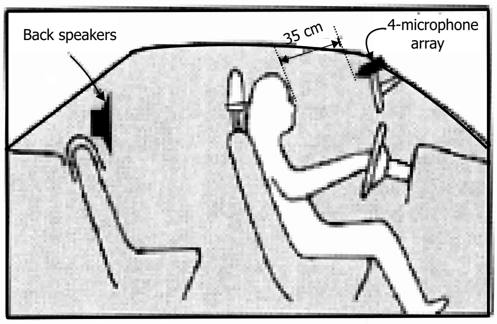


Figure 10: *Placement of microphone array and loudspeakers in the car cabin.*

## 6.3 Performance Measures

The performance evaluation is based on three measures, the distortion caused by the beamforming filters measured by the spectral deviation between the beamformer output and the source signal, and the noise and interference suppression. In order to measure the performance, the normalized distortion quantity, $\mathcal{D}$, is introduced as

$$\mathcal{D} = \frac{1}{2\pi} \int_{-\pi}^{\pi} |C_d \hat{P}_{y_s}(w) - \hat{P}_{x_s}(w)| dw \tag{33}$$

where $w = 2\pi f$, and $f$ is normalized frequency. The constant, $C_d$, is defined as

$$C_d = \frac{\int_{-\pi}^{\pi} \hat{P}_{x_s}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{y_s}(w)dw} \tag{34}$$

where $\hat{P}_{x_s}(w)$ is a power spectral density (PSD) estimate of a single sensor observation and $\hat{P}_{y_s}(w)$ is the PSD estimate of the beamformer output, when the source signal is active alone. The constant $C_d$ normalizes the mean output PSD estimate to that of the single sensor mean PSD estimate. The single sensor observation is chosen as the reference microphone observation, i.e. microphone number two in the array. The measure of distortion in Eq. (33), is the mean output PSD deviation from the observed single sensor power spectral density estimate. Ideally, the distortion is zero ($-\infty$ dB).

In order to measure the noise suppression the normalized noise suppression quantity, $S_N$, is introduced as

$$S_N = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_N}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_N}(w)dw} \tag{35}$$

and the normalized interference suppression quantity, $S_I$, as

$$S_I = C_s \frac{\int_{-\pi}^{\pi} \hat{P}_{y_I}(w)dw}{\int_{-\pi}^{\pi} \hat{P}_{x_I}(w)dw} \tag{36}$$

where

$$C_s = \frac{1}{C_d} \tag{37}$$

and where, $\hat{P}_{y_N}(w)$ and $\hat{P}_{x_N}(w)$ are PSD estimates of the beamformer output and the reference sensor observation, respectively, when the surrounding noise is active alone. In the same way $\hat{P}_{y_I}(w)$ and $\hat{P}_{x_I}(w)$ are PSD estimates when the interference signal/signals are active alone. Both the noise- and the interference- suppression measures are normalized to the amplification/attenuation caused by the beamformer to the reference sensor observation when the source signal is active alone, i.e. if the beamformer attenuates the source signal by a specific amount, the noise- and interference- suppression quantities are reduced with the same amount.

# 7 Design Parameters

In order to use the designed algorithm a number of parameters needs to be determined. The most crucial parameters are the number of subbands $K$, the prototype filter length $L$ and the subband filter length $L_{sub}$. The forgetting factor $\lambda$ of the RLS algorithm, and the weight smoothing factor $\eta$, also needs to be chosen. The design parameters listed above influence both the performance and the complexity as will be shown in the following sections. Other factors that can be adjusted for optimal performance are the calibration parameters, $\alpha$ and $\beta$, controlling the balance between the power of the calibration signals and the power of the processed signals, i.e. inputs of the beamformer. Furthermore, the length and the type of the calibration signals acquired during the training phase are also important for the output of the beamformer to be optimal.

## 7.1 Optimization of Design Parameters

All performance evaluations presented in this chapter have been done for the standard scenario on data recorded in the isolated-room environment. In this scenario, the target sound source is situated 50 cm in front of the linear sensor array. A loudspeaker situated at a distance of 1 m with a 45° angle to the array is considered as the unwanted interfering source. Prerecorded female speech is emitted from the target artificial speaker, while male speech is emitted from the speaker corresponding to the interfering source. A subset of these recordings serve as calibration signals for the corresponding positions, while the other speech sequences are used as the beamformer input signals. The unwanted ambient noise is colored noise emitted by the four surrounding loudspeakers, as given in Fig. 8. If not specified, the subband filter length is chosen to be $L_{sub} = 1$.

### 7.1.1 Filter Bank Parameters

The complexity of the filter bank depends on both the number of subbands, $K$, and the prototype filter length, $L$. The prototype filter length is chosen to be multiples of the decimation factor, giving rise to an efficient polyphase implementation. Fig. 11 illustrate the performance variation of the subband beamformer as well as the increase in computational cost for different values of the prototype filter length. The number of subbands is fixed to $K = 64$. It can be seen that the increase in noise and interference suppression as well as the increase in speech distortion as a function of the filter length stabilizes around a filter length $L = 4 \times K = 256$, to the expense of a relatively small computational cost. The delay introduced by the subband structure, i.e. the delay of the analysis and synthesis filter banks, is linearly increasing with the prototype filter length.

The relation between the number of subbands, $K$, and the performance of the subband beamformer as well as its complexity is presented in Fig. 12. The filter length is set to $L = 256$. The figures show that the noise and interference suppression increases rapidly with $K$ while the complexity of the algorithm decreases. The suppression increase and computational reduction become smaller for higher values of $K$ starting from $K = 64$. These performances are however associated to a relative increase in speech distortion as well as a significant increase in system delay.

### 7.1.2 Performance related to subband filter length

The total filter length of the subband beamformers is given by

$$L_{eff} = L_{sub} \times D \times I = L_{sub} \times K/2 \times I, \tag{38}$$

where $L_{sub}$ denotes the filter length of a single subband beamformer. Fig. 13 illustrates the algorithm performance variation when increasing the order of the subband FIR filters, $L_{sub}$, for $K = 64$ and $L = 256$. A higher noise and interference suppression is achieved with longer subband filters, to the expense of more speech distortion.

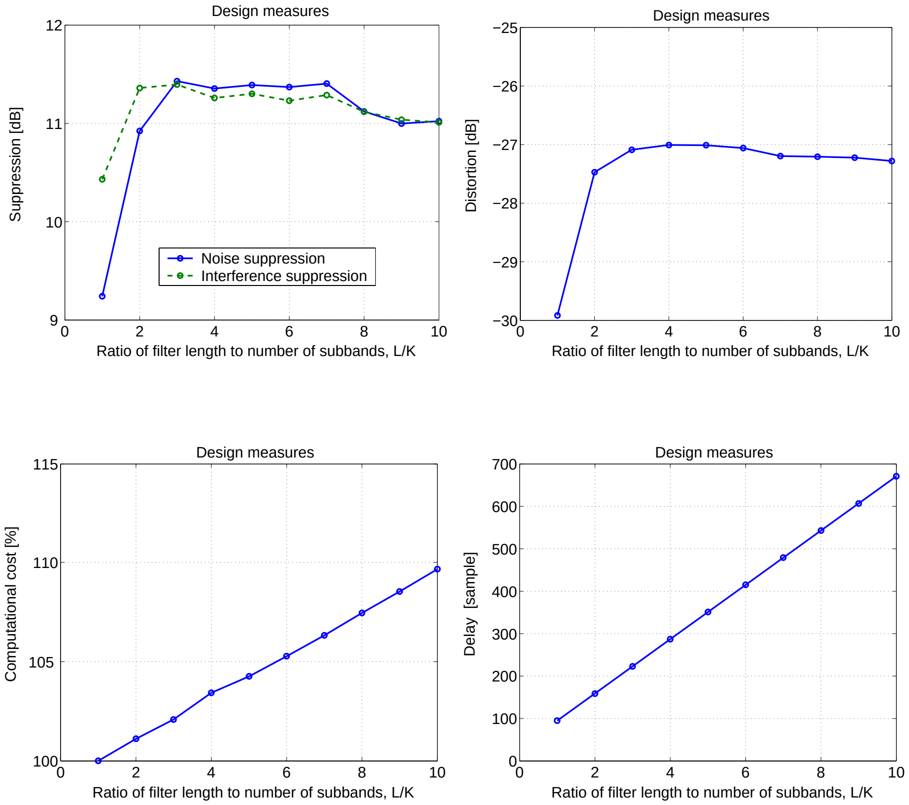


Figure 11: *Subband beamformer performance and complexity measures as a function of the prototype filter length. The number of subbands is* $K = 64$ *and the prototype filter length varies between* $L = K, ..., 10 \times K$*. The bottom left figure shows the relative complexity increase of the algorithm with longer prototype filter lengths in comparison to an* $L = K$ *based algorithm. The bottom right figure gives the delay introduced by the filter banks when increasing the prototype filter length.*

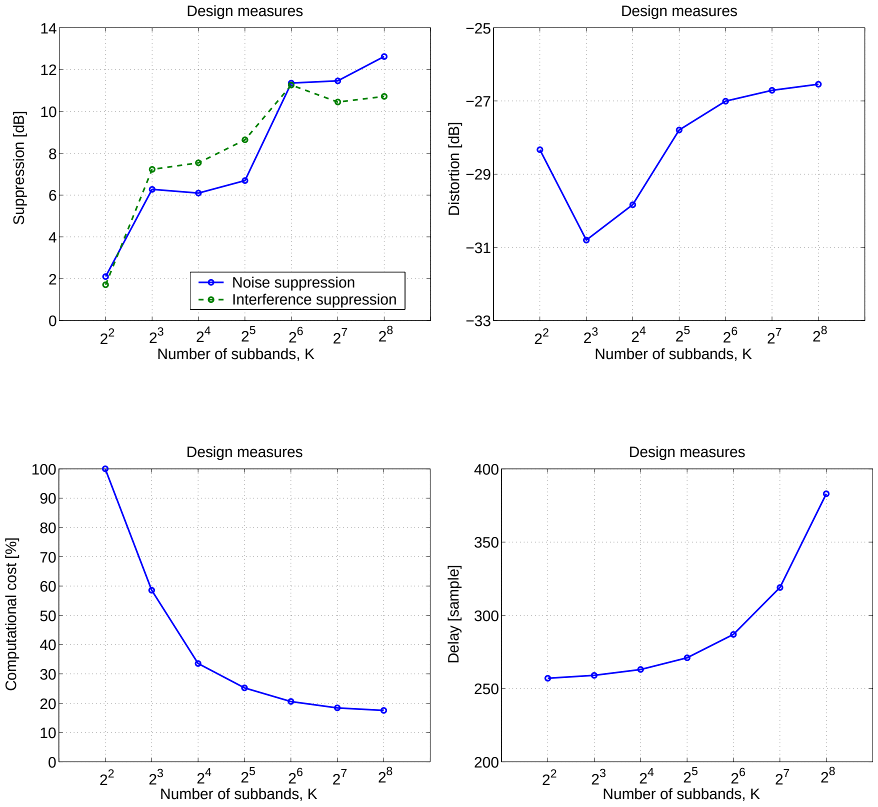


Figure 12: *Subband beamformer performance and complexity measures as well as the delay introduced by the subband structure, as a function of the number of subbands. The filter length is set to* $L = 256$.

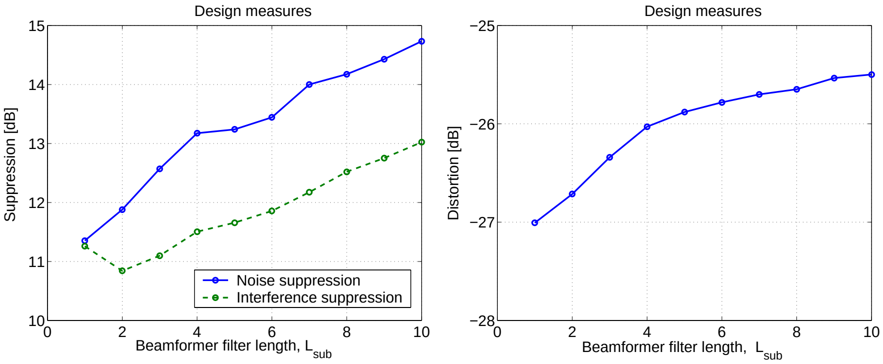


Figure 13: *Subband beamformer performance measures as a function of the subband filter length, for I=4.*

In Fig. 14 the algorithm is evaluated for a fixed effective filter length, $L_{eff} = 128$. The number of subbands varies between $K = 2, 2^2..., 2^6$, while the length of the subband FIR filters is correspondingly reduced from $L_{sub} = 2^5, 2^4..., 1$. The noise and interference suppression as well as the speech distortion measures are shown as a function of the number of subbands $K$. Results show that the number of subbands has a greater influence on the performance than the subband filter length.

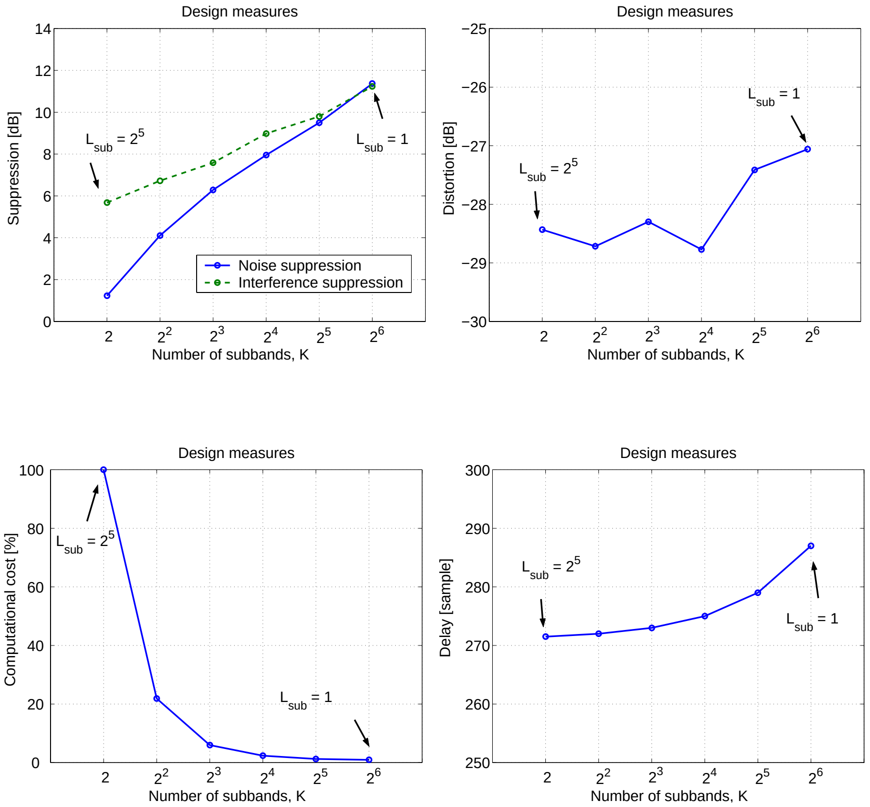


Figure 14: *Subband beamformer performance measure and complexity. The number of subbands varies between* $K = 2, 2^2, ..., 2^6$*, while the length of the subband beamformer filters is correspondingly reduced from* $L_{sub} = 2^5, 2^4..., 1$*.* $L_{eff}$ *is fixed to 128.*

### 7.1.3 Weighting in the RLS Filter

The forgetting factor $\lambda$ and the weight smoothing factor $\eta$ can be adjusted to optimize the tracking capabilities of the RLS beamformer to the observable data's statistical changes.

#### Performance versus Forgetting Factor

The smaller $\lambda$ is the less is the adaptive algorithm relying on the statistics of previous data, hence the more it follows the statistical variations of the observable data. Fig. 15 illustrates the variation of the algorithm performance for different values of $\lambda$, when processing a sound sequence composed of a target speech and interfering speech sequences, together with a colored additive noise sequence.

It can be seen that higher values of $\lambda$ achieve a considerably higher noise suppression with a relatively small increase in speech distortion. However, they also generate a high sudden variation in the noise suppression and interference suppression performances, as can be noticed for the performance displayed around time instant $t = 3.7s$.

#### Performance versus Smoothing Factor

The smoothing factor $\eta$, regulates the speed of change for the filter weights. Fig. 16 illustrates the variation of the algorithm performance for different values of $\eta$, given the same input signals as in the previous experiment. It can be observed that a high value of $\eta$ lowers the fluctuations of the algorithm performance due to the abruptly changing observable data statistics, which correspond to the speech fluctuations of the target source sequence.

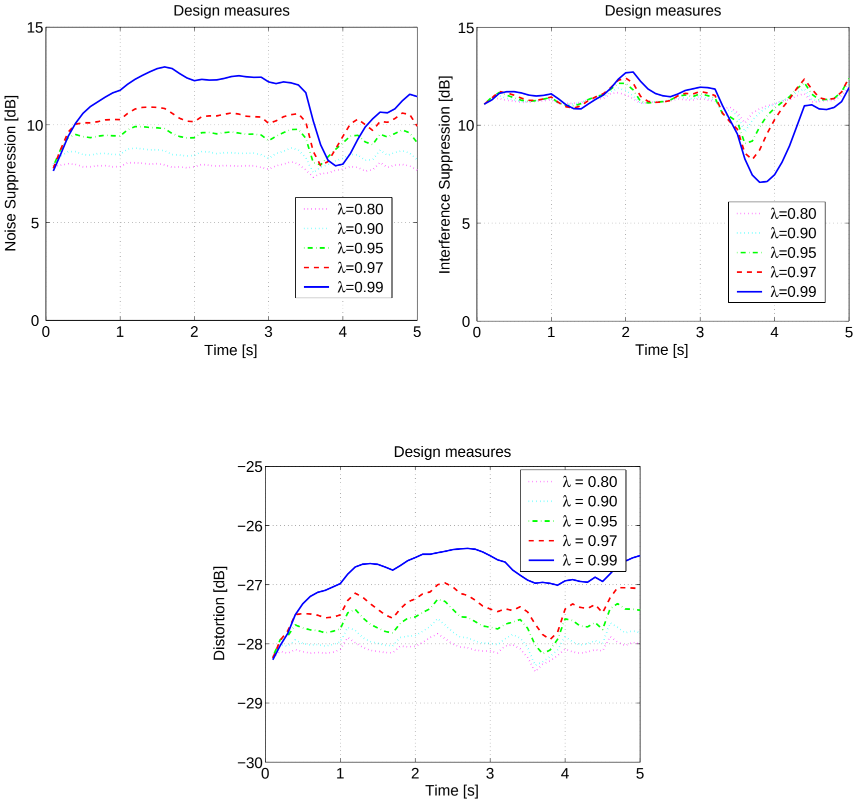


Figure 15: *Subband beamformer performance measures for different values of the weighting factor $\lambda$. Here $\eta = 0.99$.*

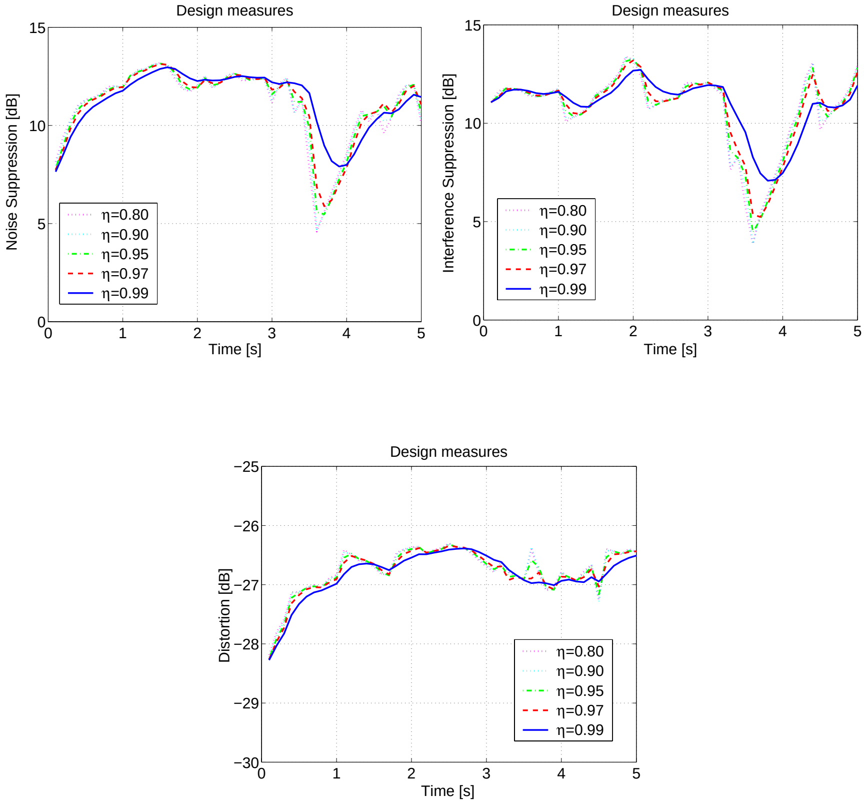


Figure 16: *Subband beamformer performance measures for different values of the smoothing factor $\eta$. Here $\lambda = 0.99$.*

## 7.2 Calibration Parameters

During the operation phase, the balance between the memorized inputs (i.e. data gathered during the calibration phase) and the microphone inputs to the adaptive beamformer can be controlled by:

- Factor $\alpha$, which corresponds to the target source calibration signal amplification/attenuation.
- Factor $\beta$, which corresponds to the interference calibration signal amplification/attenuation.

The mix of these components will control the relative amount of sources amplification or suppression. Suitable values of $\alpha$ and $\beta$ should achieve a reasonable noise and interference suppression without degrading the target signal speech. These optimal values are however different depending on the signal-to-noise and signal-to-interference ratios (SNR, SIR).

### Performance versus factor $\alpha$

Fig. 17 illustrates the beamformer performance as a function of the source calibration parameter $\alpha$, with $\beta = 1$. The plots on the left column of graphs correspond to different SNRs of the input signals, while the plots of the right column graphs correspond to different SIRs of the input signals. It can be observed that a low value of $\alpha$ increases the noise suppression and interference suppression, mainly when the SNR and SIR are low. However, this is associated to an increase in speech distortion. In a relatively noisy environment, the noise suppression is crucial to obtain a reasonable understanding of the transmitted speech, thus, a low value of $\alpha$ is more appropriate to use. In a low noise environment, on the other hand, the focus would rather be to minimize the speech distortion introduced by the beamformer. This implies the use of a higher value of $\alpha$, putting more emphasis on preserving low distortion than on achieving high noise suppression.

### Performance versus factor $\beta$

Fig. 18 illustrates the beamformer performance as a function of the interference calibration parameter $\beta$, with $\alpha = 1$, for different SNRs and SIRs. The variation of the parameter $\beta$ does not have a significant impact on the noise suppression. Higher values of $\beta$, though, increase the interference suppression at the cost of more speech distortion, especially for low SIRs. In a similar manner as with the parameter $\alpha$, it can be deduced that a high value of $\beta$ is more appropriate to use when the unwanted interfering noise is high relatively to the target signal. By this choice, more weight is put on the interference suppression ability of the beamformer.

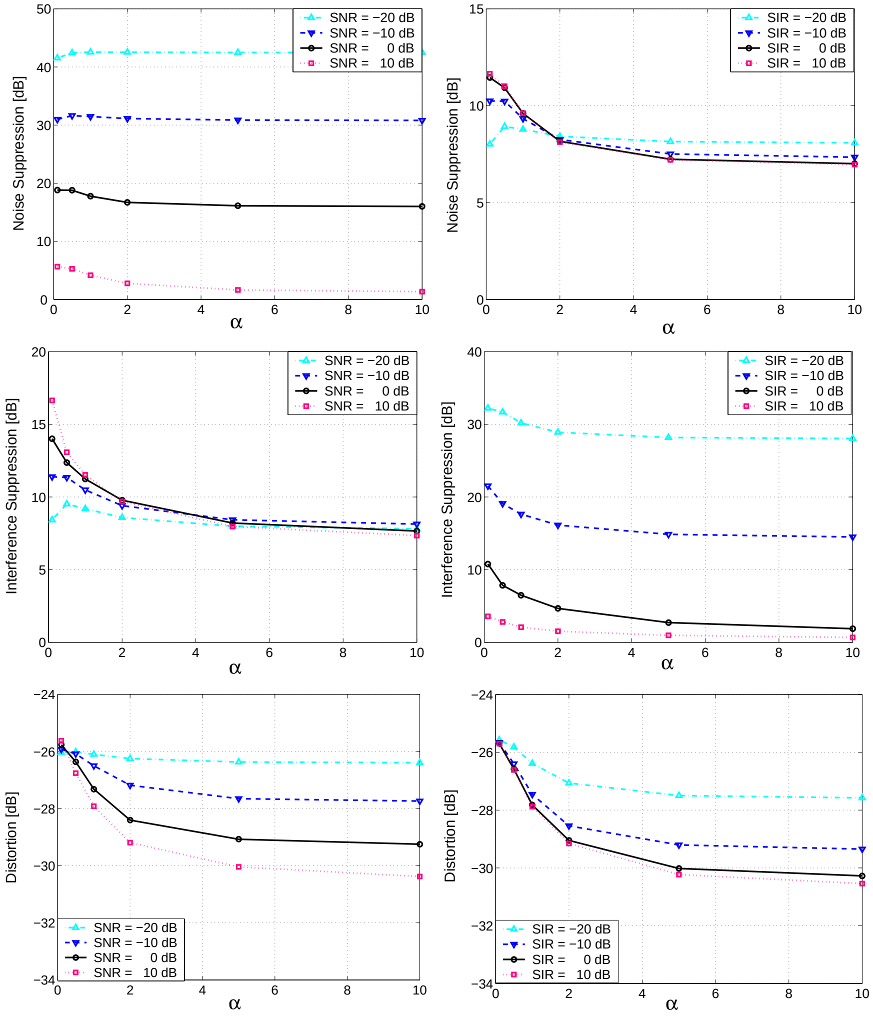


Figure 17: *Subband beamformer performance measures as a function of the target source calibration parameter* $\alpha$*, for different SNR and SIR,* $\beta = 1$*.*

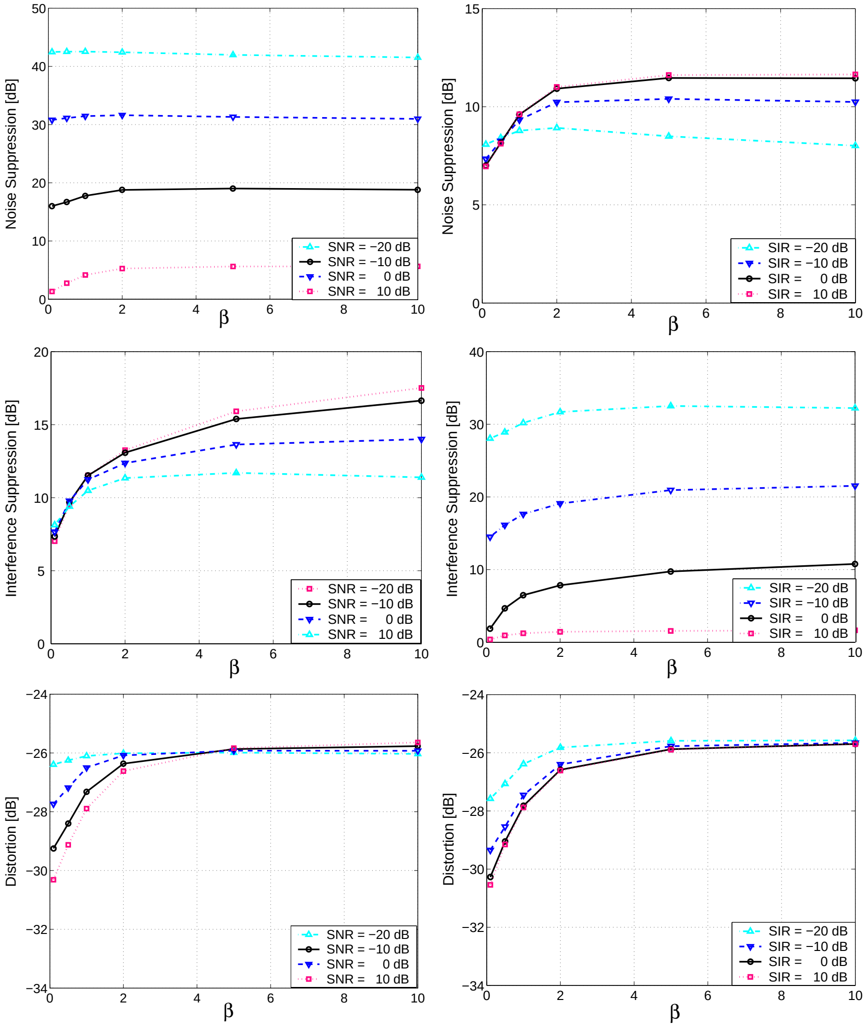


Figure 18: *Subband beamformer performance measures as a function of the target source calibration parameter $\beta$, for different SNR and SIR, $\alpha = 1$.*

## 7.3 Training Phase Parameters

The main parameters in the training phase which influence the performance of the subband constrained beamformer algorithm are the duration and the type of the training sequence.

### Performance versus Training Sequence Length

Fig. 19 illustrates the relation between the training sequence length and the algorithm performance. The noise and interference suppression reaches a peak when using a training sequence of 1800 samples, i.e. a time duration of 1.5 s. A longer training sequence results in less speech distortion, with a decrease in the noise suppression as well. This is due to the fact that the use of more data samples in the calibration phase calculations increases the power of the pre-calculated correlation matrix estimates, $\hat{\mathbf{R}}_{ss}^{(\Omega)}$ and $\hat{\mathbf{R}}_{ii}^{(\Omega)}$, which is comparable to the use of higher values of the calibration parameters, $\alpha$ and $\beta$ (see Sec. 7.2).

### Performance versus Training Sequence Type

To assess the influence of the training sequence type on the beamformer function, a white gaussian noise sequence as well as speech sequences, corresponding to male and female speakers, emitted from the artificial talkers were recorded. These sequences alternatively served as the desired and interfering sound source calibration signals. The scenarios simulated are defined by the calibration sequences emitted from the target and interfering loudspeakers as stated in Table 1 and the corresponding results are shown in Fig. 20.

It can be seen from Fig. 20 that using a source calibration sequence with a similar spectral content as the processed signal generated from the same position (scenarios 1, 2, 8, 9) gives a high noise and interference suppression, with a low speech distortion. The performance considerably decreases when using a sequence with a different spectral content, in the worst case a white noise sequence, to calibrate the beamformer. Additionally, it can be noticed that when using the adequate calibration sequences, the extraction of female speech provides generally a better overall performance than the processing of male speech (up to 5 dB more noise suppression and 3 dB less speech distortion).

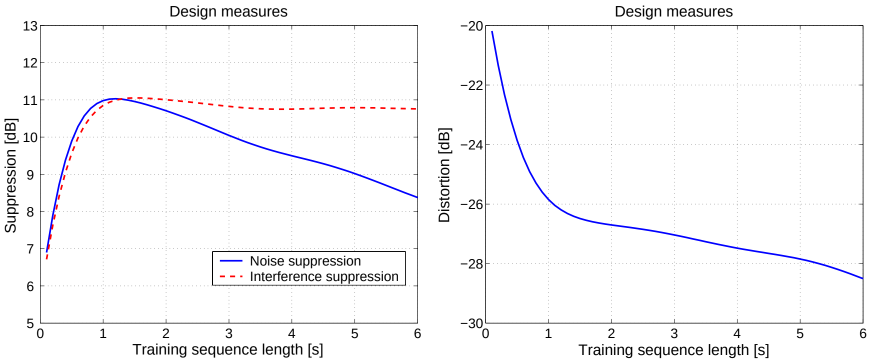


Figure 19: *Subband beamformer performance measures as a function of the training sequence length.*

| Speaker | Calibration phase | | Operational phase | |
|---|---|---|---|---|
| | Target | Interfering | Target | Interfering |
| Scenario (1) | female | male | female | male |
| Scenario (2) | female | female | female | male |
| Scenario (3) | male | male | female | male |
| Scenario (4) | male | female | female | male |
| Scenario (5) | white noise | white noise | female | male |
| Scenario (6) | female | male | male | female |
| Scenario (7) | female | female | male | female |
| Scenario (8) | male | male | male | female |
| Scenario (9) | male | female | male | female |
| Scenario (10) | white noise | white noise | male | female |

Table 1: *Type of sound used in the calibration and operation phase for different scenarios.*

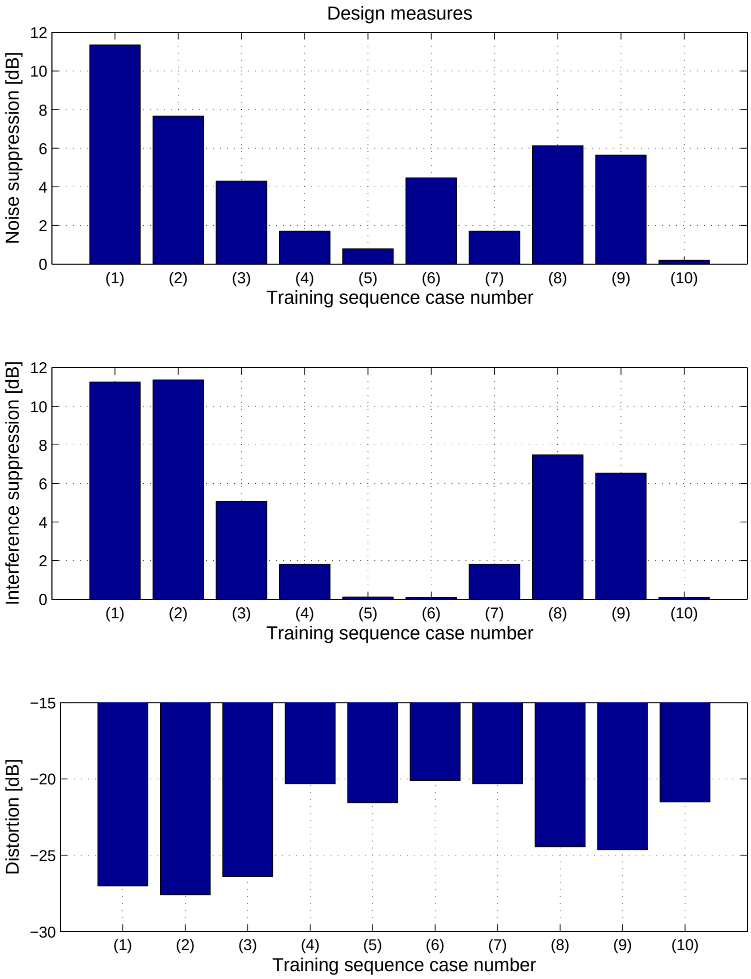


Figure 20: *Subband beamformer performance measures for the scenarios described in Table 1, and corresponding to different types of sound sequences used.*

# 8 Performance Evaluation

In this chapter, the constrained subband beamformer is evaluated for hands-free speech enhancement applications, using real speech signals with different settings and in different environments. The algorithm is run for a one tap subband filter length, i.e. $L_{sub} = 1$, with the following parameter settings: $K = 64$, $L = 256$, $\lambda = \eta = 0.99$, $\alpha = \beta = 1$ and training sequences of 2400 samples originating from the same speaker used at the corresponding position during the online processing.

## 8.1 Variation of Position

Measurements were conducted in the isolated-room environment, where the impact of a different set up of source positions on the beamformer performance is investigated.

### Performance versus Target Source Position

A setup consisting on varying the position of the target speaker relative to the microphone array, is illustrated in Fig. 21. This experiment was conducted in order to evaluate the influence of the speech source position on the speech enhancement performance of the beamformer. Each position is defined by its spherical coordinates with the origin situated at the center of the array, i.e. each point is represented by its distance to the origin and the angle from the perpendicular line to the array passing by its center. The speaker was first placed at a distance of 50 cm in front of the array, referred to as direction with angle 0°. This location corresponds to the standard position of a person using a microphone array mounted at eye level in front of him/her. In scenarios (2)-(8) depicted in Fig. 21, different angles of the speaker's position with a fixed distance of 50 cm were simulated. Scenarios (9) and (10) were aimed at comparing the beamformer's performance for different distances of the speaker to the array. Both female and male speech sequences were emitted from each position, while background noise corresponding to recorded computer fan noise was generated by the four surrounding loudspeakers. The beamformer was run for each sequence, at each position, individually.

The beamformer's noise suppression and speech distortion levels for this setup are presented in Fig. 22. Results achieved in this environment show up till 13 dB noise suppression. It can be seen that by increasing the angle of arrival of the speech, when moving out of the angular view of the array (above

60°), the noise suppression decreases, reaching a relatively low suppression level when the speaker is located behind the sensor array. This also comes with an increase in speech distortion. On the other hand, the algorithm's performance increases when the target source is closer, in radius, to the array.

The difference in performance for a target source signal corresponding to female speech in comparison to the enhancement of male speech is significant when it comes to both the noise suppression ($\sim$ 2-5 dB) as well as the speech distortion ($\sim$ 3-4 dB). This can be explained by a larger frequency overlapping of the noise spectrum with the male speech spectrum than with the female speech spectrum, as can be seen in Fig. 23. The power spectrum of the simulated background noise is more represented in the low frequencies where the male speech signal displays more power than the female speech signal.

### Performance versus Interference Source Position

A second setup was arranged with the ambition of investigating the impact of the interference position compared to the target source on the beamformer output. Ten scenarios have been simulated with a fixed position of the target source, situated at 50 cm in front of the array, and different positions of the interfering loudspeaker, relative to the microphone array as illustrated in Fig. 24. The ambient noise composed of computer fan and air conditioner noise was generated by the surrounding loudspeakers and constituted the background noise.

Results presented in Fig. 25 show that a small angular separation between the target source and the interference source (below 30°), reduces the noise and the interference suppression. However, there is a relative invariance of the noise suppression with angles above 30° degrees between the two sources. The relative increase in interference suppression displayed for separation angles around 30°-45°, accompanied by a slight improvement of speech distortion, can be explained by the fact that different positions of a sound source result in a different set of reflections which are received by the microphone array. The interference is relatively better suppressed ($\sim$1-2 dB) when generated from a closer position in radius to the sensor array.

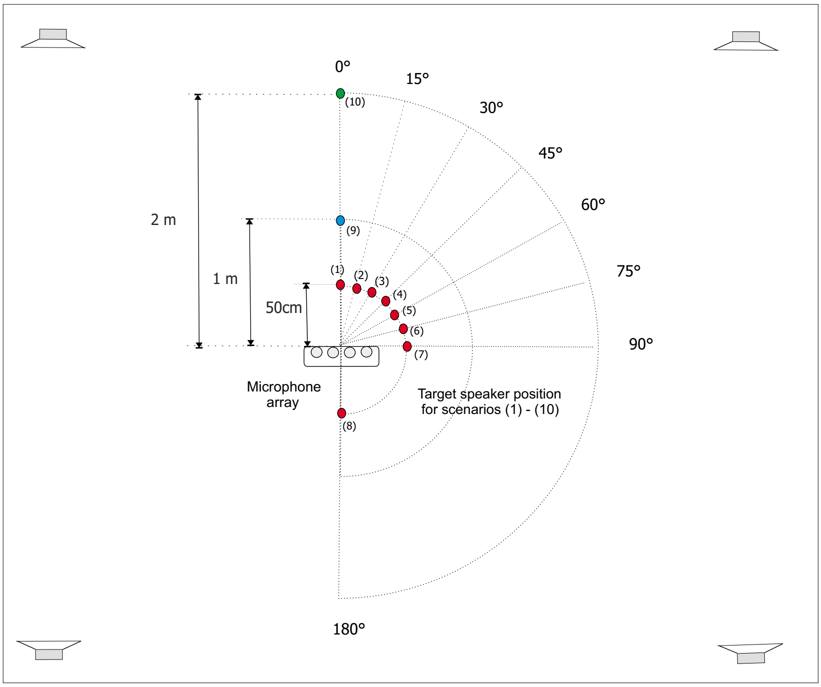


Figure 21: *Data acquisition scenarios: The target speaker has a specific position for each scenario. In scenarios (1)-(8) it is at a distance of 50cm from the array center but with a different angle at each scenario. In scenarios (9) and (10) the target speaker is positioned in front of the array, at a distance of 1m and 2m respectively. The ambient noise of the room generated by the four loudspeakers, situated at each corner, is included in all scenarios.*

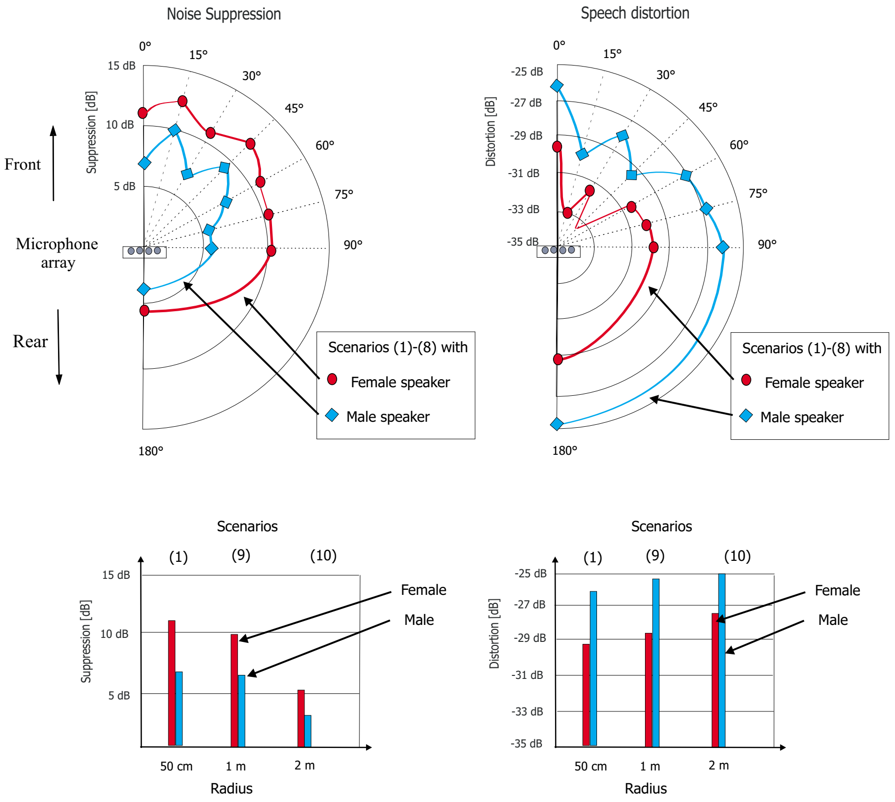


Figure 22: *Subband beamformer performance measures achieved for the scenarios presented in Fig. 21, using female an male speech as target source signals. The top plots correspond to the angle variation (scenarios (1)-(8)) and the bottom plots to the radius variation (scenarios (1),(9) and (10)).*

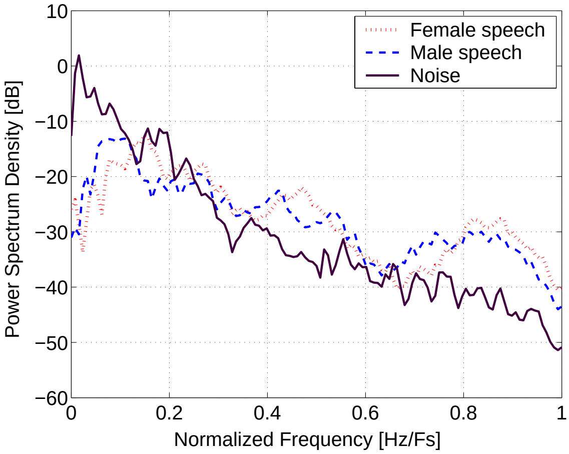


Figure 23: *Power Spectrum Density comparison for the female and male target speech signals and the background noise.*

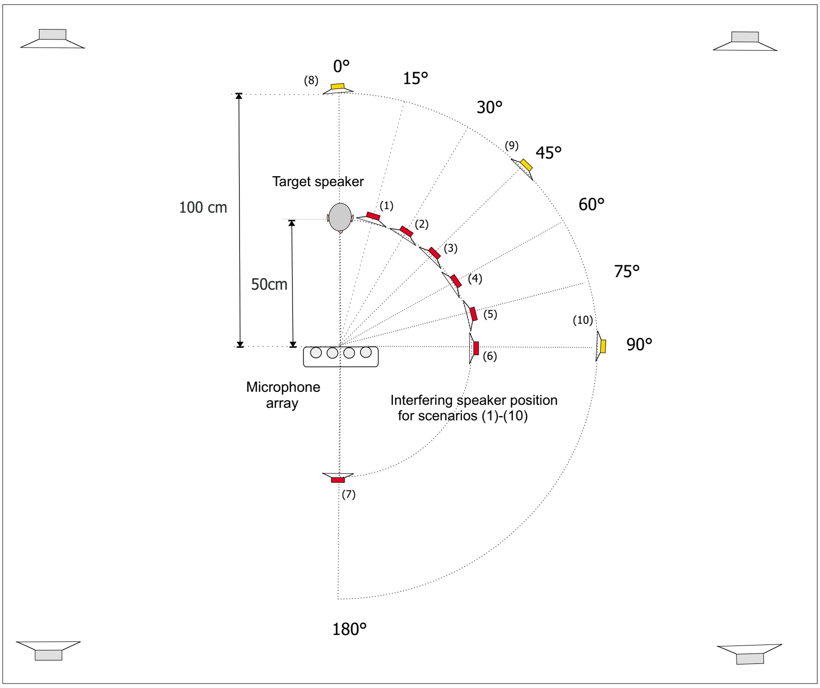


Figure 24: *Data acquisition scenarios: The target speaker is at a distance of 50cm from the microphone array, and has a fixed position for all scenarios. The interfering speaker in scenarios (1)-(7) is also at a distance of 50cm from the array center but makes a different angle with the target source position at each scenario. In scenarios (8)-(10) the interfering speaker is positioned further away from both the array (1m) and the target source. The ambient noise of the room generated by the four loudspeakers, situated at each corner, is included in all scenarios.*

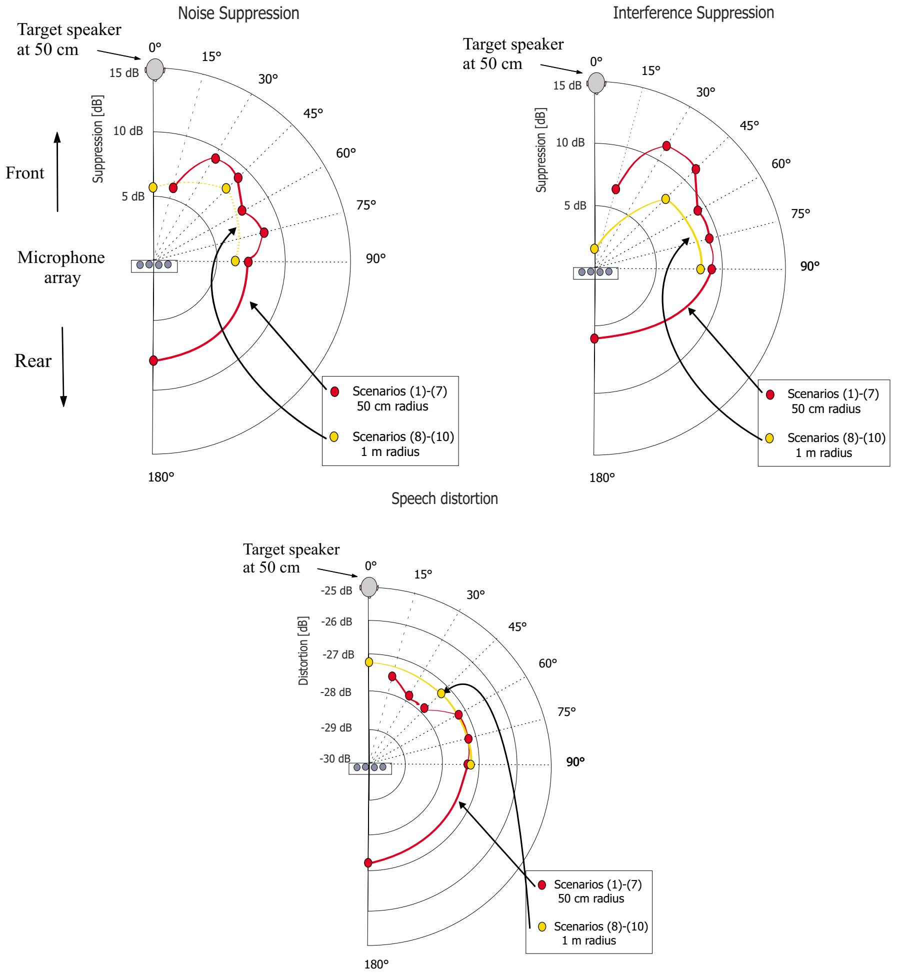


Figure 25: *Subband beamformer performance measures achieved for the scenarios presented in Fig. 24. The red points correspond to a 50cm radius, and the yellow points to a 1m radius, for the position of the interference source.*

## 8.2 Variation of Background Noise

Different background noises were simulated to evaluate the performance of the beamformer, for a typical scenario. The surrounding loudspeakers from the isolated-room emitted successively fan noise, recorded in the vicinity of a computer fan, music and speech.

The proposed beamformer shows a higher noise suppression capability when the unwanted noise is from a computer fan, see Fig. 26. The noise suppression drops considerably when the unwanted noise consists of speech. The power spectral content of the different noise sources simulated as well as for the desired speech source are compared in Fig. 27. As expected, the beamformer displays better performance when the noise has a different spectral content than the speech signal.

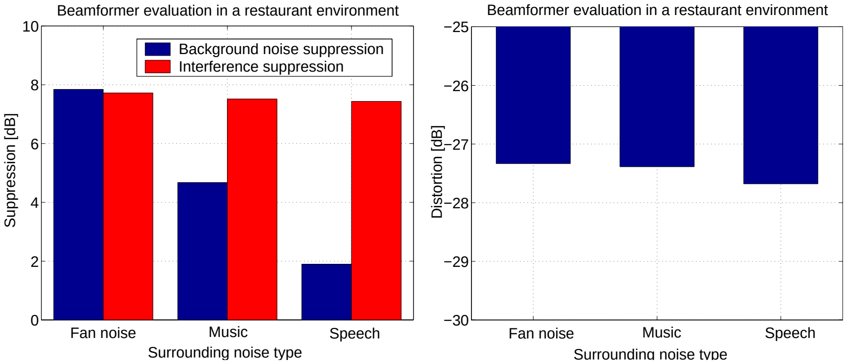


Figure 26: *Subband beamformer performance measures for different types of background noise.*

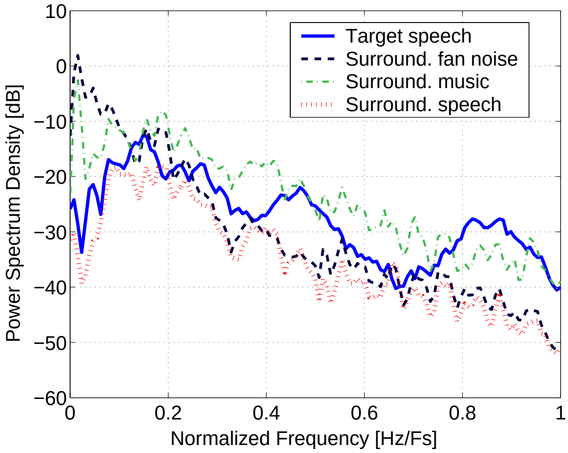


Figure 27: *Power Spectrum Density of the desired speech signal and the different background noise sources.*

## 8.3 Variation of Environment

### Restaurant Environment

Measurements were conducted in a restaurant environment, with different setup. The ambient noise generated by the activity in the restaurant constitutes the background noise. Similarly to the previous setup, a number of scenarios have been evaluated with a fixed position of the target source, situated in front of the array at a distance of 50 cm, and different positions for the interfering loudspeaker, relative to the microphone array. These scenarios are described in Fig. 28. Both male and female speech sequences were used as target signal in the simulations.

The beamformer performance for each of the scenarios illustrated in Fig. 28 are presented in Fig. 29. The interference suppression is affected by the position of the interfering source relative to the sensor array as well as its distance to the target source. A sufficient angular separation (above 30°) between the two sources is required for the beamformer to distinguish them and achieve a reasonably good interference suppression ($\sim$ 10 dB). However, the interference suppression level drops with the interfering source moving out of the vicinity of the beamformer. In such environment, up to 7 dB noise suppression level has been achieved.

The spectral comparison of the background noise, generated by the activity in the restaurant, to the female and male speech sequences used as desired signals, in Fig. 30, exhibits a higher frequency similarity of the noise to the male speech in the low frequency range (below $f = 0.2$). The beamformer therefore achieves better results when enhancing female speech.

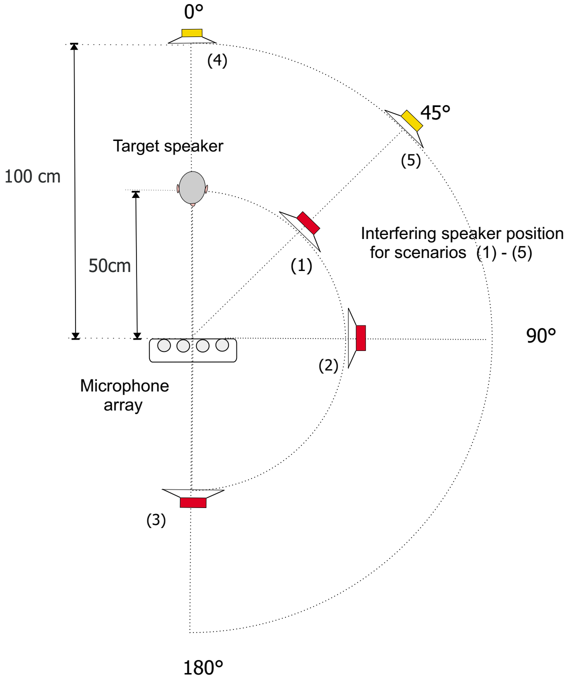


Figure 28: *Data acquisition scenarios in the restaurant environment: The target speaker is at a distance of 50cm from the microphone array, and has a fixed position for all scenarios. The interfering speaker in scenarios (1) to (3) is also at a distance of 50cm from the array center but makes a different angle with the target source position at each scenario. In scenarios (4) and (5) the interfering speaker is positioned further away from both the array (1m) and the target source. The same ambient noise from the restaurant at busy hours is included in all scenarios.*

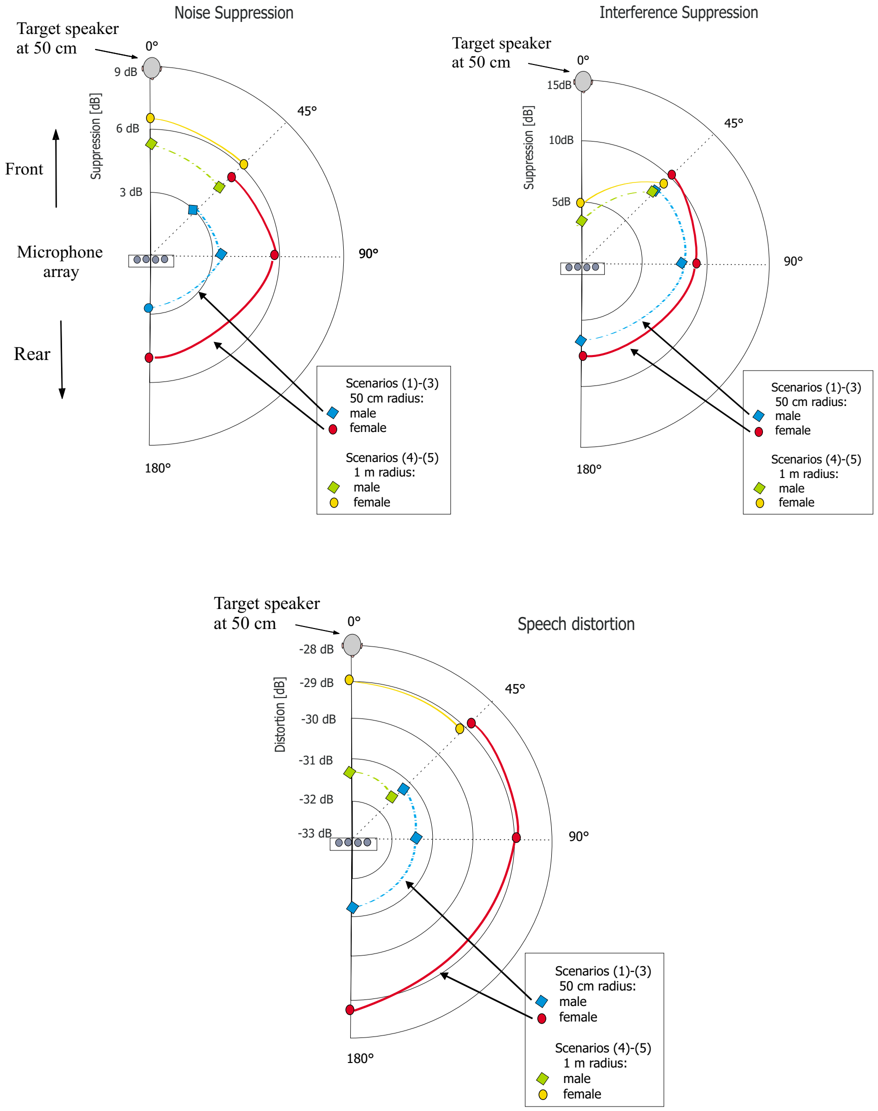


Figure 29: *Subband beamformer performance measures for the scenarios of Fig. 28. The dots correspond to results with female speech desired signals and the diamonds to the use of male speech desired signals. The interference consisted of a mix of male and female speech sequences.*

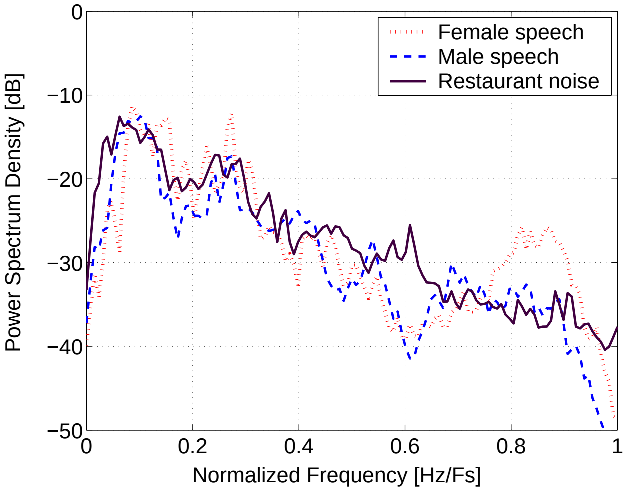


Figure 30: *Power Spectrum Density comparison for the female and male target speech signals and the restaurant background noise.*

### Car Environment

Measurements where conducted in the car environment described in Sec. 6.2.3. Different scenarios where arranged to estimate the suppression capabilities of the proposed beamformer in such environment. In all scenarios, the aim is to enhance the speech of a person sitting in the driver position. Recordings of the speech of individuals (from both genders) sitting at the driver seat were performed with the engine off, in a quite environment. Recordings in a moving car as well as recordings of the different loudspeakers (described in Fig. 9) emitting speech were recorded individually. For performance evaluation purposes, scenarios with several sound sources active simultaneously were simulated by summing for each microphone the individual outputs corresponding to each of the sources active alone. For each scenario simulated the unwanted sound sources are as follow:

**(1)** The car cabin noise when the car is moving at a fixed speed of 100 km/h.

**(2)** Same car cabin noise as in (**1**). Female speech is emitted from the loudspeaker positioned by the passenger seat on the front of the car, and is considered as a known interfering source with fixed position.

**(3)** Same car cabin noise as in (**1**). Male speech is emitted from the loudspeaker positioned by the passenger seat on the front of the car, and is considered as a known interfering source with fixed position.

**(4)** Same car cabin noise as in (**1**). The known interfering source consists of the back loudspeakers emitting different speech sequences (male and female) simultaneously.

Performance results are given in Fig. 31 for each scenario. Results show that more than 5 dB noise suppression and 10 dB interference suppression are accomplished for all scenarios. A relatively lower noise suppression ($\sim$ 1 dB) and slightly higher speech distortion ($\sim$ 0.5 dB) can be noticed for scenario (**4**), when the back loudspeakers are active. It can also be noticed that the cancellation of a jammer is more effective ($\sim$ 4 dB) when the generated interfering signal is composed of female speech in comparison to the suppression of male speech interference.

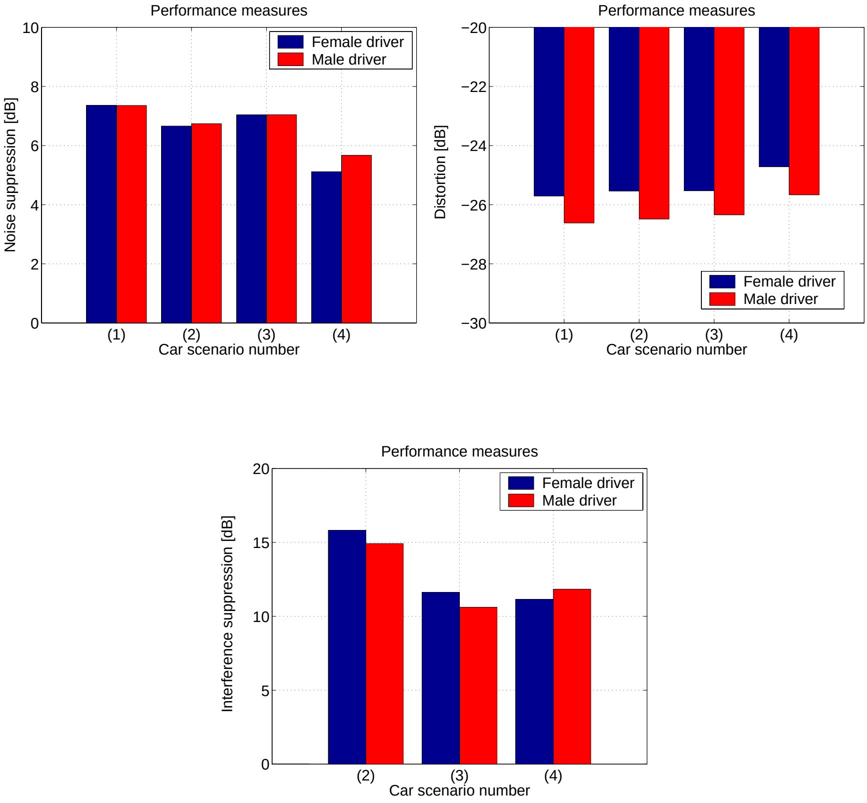


Figure 31: *Subband beamformer performance measures for scenarios performed in a car cabin environment.*

# 9 Conclusion

In this paper, a subband calibrated beamforming algorithm for speech enhancement has been presented and evaluated in real hands-free acoustic environments.

The solution is based on the principle of a soft constraint, formed from calibration data. The algorithm recursively estimates the spatial information of the received data, while the initial precalculated source correlation estimates constitute a soft constraint in the solution.

The pre-specified design parameters of the subband beamformer, related to the subband structure and to the RLS algorithm, are adjusted such to achieve optimal performance of the presented method. Additionally, weighting factors, $\alpha$ and $\beta$ can be used to control the relative amplification of the memorized calibration data in comparison to the input data during processing. By varying the parameters $\alpha$ and $\beta$, more emphasis can be put on improving the noise and interference suppression or on reducing the speech distortion. The results obtained are however dependent on the SNR and SIR of the input data.

A hands-free implementation with real signals using a linear array, under noisy conditions such as a crowded restaurant room and a car cabin in movement, shows up to 15 dB noise and interference suppression in the restaurant room and more than 15 dB suppression in the car cabin, achieved with low speech distortion. However, since spatial separation between the source and the jammer as well as their positioning relative to the array, control the achievable level of speech distortion and interference suppression, the placement of the hands-free loudspeaker should be made with this fact in mind.

## References

[1] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.

[2] Y. Kameda and J. Ohga, "Adaptive Microphone-Array System for Noise Reduction," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, pp. 1391–1400, December 1986.

[3] L. C. Godara, "Application of Antenna Arrays to Mobile Communications. II Beamforming and Direction-of-Arrival Considerations," in *Proceedings of IEEE*, vol. 85, pp. 1195–1245, August 1997.

[4] A. B. Gershman, V. I. Turshin, and V. A. Zverev, "Experimental Results of Localization of Moving Underwater Signal by Adaptive Beamforming," in *IEEE Transactions on Signal Processing*, vol. 44, pp. 2605–2611, October 1996.

[5] A. B. Gershman, E. Nemeth, and J. F. Böhme, "Experimental Performance of Adaptive Beamforming in a Sonar Environment with a Towed Array and Moving Interfering sources," in *IEEE Transactions on Signal Processing*, vol. 48, pp. 246–250, January 2000.

[6] J. Capon, R. J. Greenfield, and R. J. Kolker, "Multidimensional Maximum-Likelihood Processing for a Large Aperture Seismic Array," in *Proceedings of IEEE*, vol. 55, pp. 192–211, February 1967.

[7] M. Brandstein and D. Ward, *Microphone Arrays, Signal Processing Techniques and Applications*, Springer, 2001.

[8] D. A. Florêncio and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 197–200, May 2001.

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

[10] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.

[11] N. Grbić, "Optimal and Adaptive subband beamforming, Principles and applications" Dessertation Series No 01:01, ISSN:1650-2159, Blekinge Institute of Technology, 2001.

[12] N. Grbić and S. Nordholm, "Soft constrained subband beamforming for hands-free speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 885–888, May 2002.

[13] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, "Design of oversampled uniform dft filter banks with delay specifications using quadratic optimization," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. VI, pp. 3633–3636, May 2001.

[14] S. Nordholm, I. Claesson and M. Dahl, "Adaptive Microphone Array Employing Calibration Signals: An Analytical Evaluation," in *IEEE Transactions on Speech andAudio processing*, vol. VII, pp. 241–252, May 1999.

[15] E. Eweda, "Maximum and Minimum Tracking Performances of Adaptive Filtering Algorithms over Target Weight Cross Correlations," in *IEEE Transactions on Circuits and Systems*, vol. 45, No. 1, pp. 123–132, January 1998.

[16] E. Eweda, "Convergence Analysis of Adaptive Filtering Algorithms with singular Data Covariance Matrix," in *IEEE Transactions on Signal Processing*, vol. 49, No. 2, pp. 334–343, February 2001.

[17] N. Grbić M. Dahl, and I. Claesson, "Acoustic Echo Cancelling and Noise Suppression with Microphone arrays," Research report 1999:5, ISSN:1103-1581, University of Karlskrona/Ronneby, April 1999.

[18] P. P. Vaidyanathan, *Multirate Systems and Filter Banks,* Prentice-Hall, 1993.

# Appendix

## A Type 1 Polyphase Decomposition

If the filter $H(z)$ is an FIR filter defined by

$$H(z) = \sum_{n=-\infty}^{+\infty} h(n) z^{-n}, \tag{39}$$

the type 1 polyphase decomposition of $H(z)$, with $D$ elements, is

$$\begin{aligned} H(z) = & \sum_{n=-\infty}^{+\infty} h(Dn) z^{-Dn} \\ & + z^{-1} \sum_{n=-\infty}^{+\infty} h(Dn+1) z^{-Dn} \\ & + ... \\ & + z^{-D+1} \sum_{n=-\infty}^{+\infty} h(Dn+D-1) z^{-Dn} \\ = & \sum_{l=0}^{D-1} z^{-l} E_l(z^D), \end{aligned} \tag{40}$$

where $E_l(z)$, $l = 0, ..., D-1$, are the type 1 polyphase components given by

$$E_l(z) = \sum_{n=-\infty}^{+\infty} h(Dn+l) z^{-n}. \tag{41}$$

## B Type 2 Polyphase Decomposition

The type 2 polyphase decomposition of $H(z)$, with $D$ elements, is

$$\begin{aligned} H(z) &= \sum_{l=0}^{D-1} z^{-l} F_{D-l-1}(z^D) \\ &= \sum_{l=0}^{D-1} z^{-(D-l-1)} F_l(z^D), \end{aligned} \tag{42}$$

where $F_l(z)$, $l = 0, ..., D-1$, are the type 2 polyphase components of the filter $H(z)$, given by

$$F_l(z) = \sum_{n=-\infty}^{+\infty} h(Dn - l - 1) z^{-n}. \tag{43}$$

# Part II

# Spatial Filter Bank Design for Speech Enhancement Beamforming Applications

**Part II is published as:**

Z. Yermeche, P. Cornelius, N. Grbić and I. Claesson, “Spatial Filter Bank Design for Speech Enhancement Beamforming Applications,” published in Third IEEE Sensor Array and Multichannel Signal Processing Workshop Proceedings, Sitges, Spain, July 2004.

# Spatial Filter Bank Design for Speech Enhancement Beamforming Applications

Z. Yermeche, P. Cornelius, N. Grbic and I. Claesson

## Abstract

In this paper, a new spatial filter bank design method for speech enhancement beamforming applications is presented. The aim of this design is to construct a set of different filter banks that would include the constraint of signal passage at one position (and closing in other positions corresponding to disturbing sources). By performing the directional opening towards the desired location in the fixed filter bank structure, the beamformer is left with the task of tracking and suppressing the continuously emerging noise sources.

This algorithm has been implemented in MATLAB and tested on real speech recordings conducted in a car hands-free communication situation. Results show that a reduction of the total complexity can be achieved while maintaining the noise suppression performance and reducing the speech distortion.

# 1 Introduction

Microphone arrays can be exploited for speech enhancement in order to extract a speaker while suppressing interfering speech and background noise. These arrays are used in conjunction with digital beamforming, a technique providing spatial selectivity to separate signals that have overlapping frequency content but are originated from different spatial locations. A microphone array consists of a set of acoustic sensors placed at different locations in order to spatially sample the sound pressure field. It offers a directivity gain proportional to the number of sensors. Thus, adaptive array processing, i.e. beamforming, of the spatial microphone samples allows time-variant control of spatial and spectral selectivity [1, 2].

Several beamforming techniques have been suggested in order to enhance a desired speech source [3, 4, 5]. A Constrained Adaptive Subband Beamformer has been evaluated in [6] for speech enhancement in hands-free communication situations. The adaptive beamformer optimizes the array output by adjusting the weights of finite length digital filters so that the combined output contains minimal contribution from noise and interference. A soft constraint, formed from calibration data, secures the spatio-temporal passage of the desired source signal, without the need of any speech detection. The computational complexity of the finite impulse response filters is substantially reduced by introducing a subband beamforming scheme [7].

The weight update equation for the constrained adaptive beamformer implies the calculation and the use of a combined covariance matrix at each iteration. From the observation that the combined covariance matrix comprises a pre-calculated fixed part and a recursively updated part, the beamforming problem can be divided into a fixed part and an adaptive part. The objective in this paper is to transfer the a-priori known portion of the optimization process into the filter bank structure (fixed part of the system).

Information about the desired speech location is used in the filter bank design by adding a spatial decomposition of the multichannel data for each subband. This spatial decomposition takes the form of a spatial transformation matrix, and it is extracted from correlation function estimates. Such structure is shown to reduced the total complexity of the system while improving its performance. The main improvement being a faster convergence speed and less speech distortion.

## 2 Subband Beamforming

Figure 1 illustrates the overall architecture of the microphone array speech enhancement system, based on the constrained adaptive subband beamformer. The structure includes a multichannel uniform over-sampled analysis filter-bank used to decompose the received array signals into a set of subband signals and a set of adaptive beamformers, each adapting on the multichannel subband signals. The outputs of the beamformers are reconstructed by a synthesis filter-bank in order to create a time domain output signal. The spatial characteristics of the input signal are maintained when using the same modulated filter bank for all channels. The filter banks are defined by two prototype filters, which leads to efficient polyphase realizations [8].

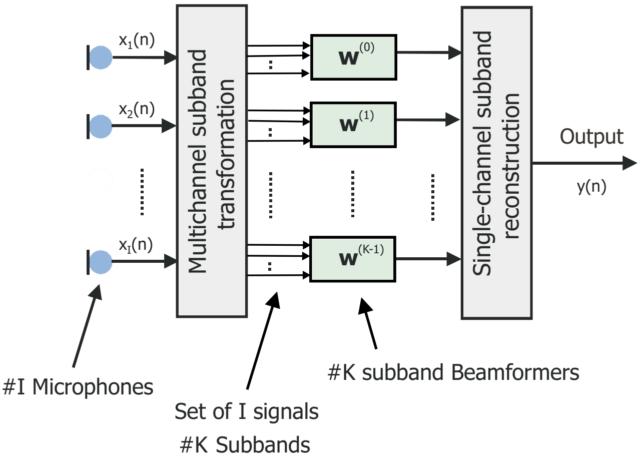


Figure 1: *Structure of the subband beamformer.*

The source is assumed to be a wide-band source, as in the case of a speech signal, located in the near-field of a uniform linear array of number $I$ microphones. The filtering operations of the beamformer are formulated in the frequency domain as multiplications with $I$ complex frequency domain representation weights, $w_i^{(f)}(n)$, for each frequency.

For a specific frequency, $f$, the output is given by

$$y^{(f)}(n) = \sum_{i=1}^{I} w_i^{(f)}(n)\ x_i^{(f)}(n) \tag{1}$$

where the signals, $x_i^{(f)}(n)$ are digitally sampled microphone observations and $y^{(f)}(n)$ corresponds to the beamformers output. These time domain signals are narrow band, containing essentially components with frequency $f$.

The objective of the beamformer is formulated in the frequency domain as a calibrated weighted recursive least square solution, where the optimal weight vectors $\mathbf{w}_{ls,opt}^{(f)}(n)$ are calculated by

$$\mathbf{w}_{ls,opt}^{(f)}(n) = \left[\hat{\mathbf{R}}_{ss}^{(f)} + \hat{\mathbf{R}}_{ii}^{(f)} + \hat{\mathbf{R}}_{xx}^{(f)}(n)\right]^{-1} \hat{\mathbf{r}}_s^{(f)} \tag{2}$$

where an initial calibration procedure is used to calculate source correlation estimates, i.e. the correlation matrix estimate $\hat{\mathbf{R}}_{ss}^{(f)}$ and the cross correlation vector estimate $\hat{\mathbf{r}}_s^{(f)}$, for microphone observations when the source signal of interest is active alone, as well as the interference correlation matrix estimate, $\hat{\mathbf{R}}_{ii}^{(f)}$, when the known source interferences are active alone [6].

Conversely, the correlation estimates, $\hat{\mathbf{R}}_{xx}^{(f)}(n)$, are continuously calculated from observed data by

$$\hat{\mathbf{R}}_{xx}^{(f)}(n) = \sum_{p=0}^{n} \lambda^{n-p} \mathbf{x}^{(f)}(p)\ \mathbf{x}^{(f)^H}(p) \tag{3}$$

where

$$\mathbf{x}^{(f)}(n) = [x_1^{(f)}(n), \quad x_2^{(f)}(n), \quad \ldots \quad x_I^{(f)}(n)]^T$$

and $\lambda$ is a forgetting factor, with the purpose of tracking variations in the surrounding noise environment. The initially precalculated correlation estimates constitutes a soft constraint in the recursive update of the beamforming weights.

# 3 Suggested method

In the previous structure of the constrained subband beamformer, both the fixed pre-calculated source correlation estimates and the updated data correlation estimates are used at each iteration for the update of the beamformers weight vectors (see Eq. (2)).

In this section a more efficient design method is introduced. The spatial information carried by the source correlation estimates is used to process the data prior to the beamformers, through a matrix transformation, based on matrices $\mathbf{V}^{(f)}$ (see Fig. 2), undergoing a spatial decomposition.

The resulting subband signal vector $\mathbf{x}'^{(f)}(n)$ is given by

$$\mathbf{x}'^{(f)}(n) = \mathbf{V}^{(f)H}\mathbf{x}^{(f)}(n). \tag{4}$$

By this method, the spatial information carried by the input vector to the beamformer is transformed in such a way to direct the array towards the source position, and close its opening in directions of known interfering sources.

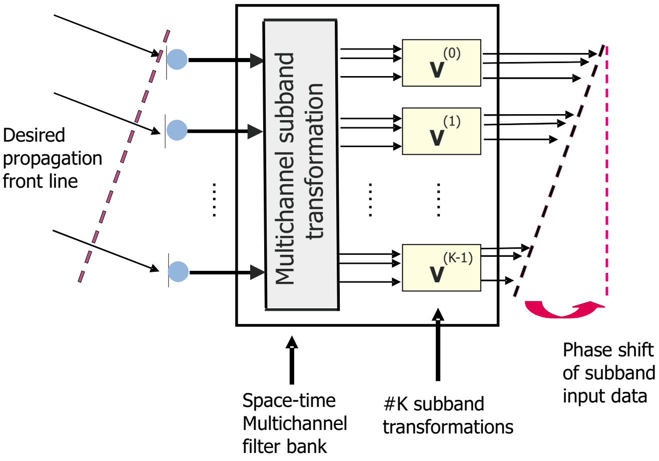


Figure 2: *Structure of the multidimensional space-time filter bank (The output data of the filter bank are phase-shifted to be in-phase for the source propagation direction, and out-of-phase for interference propagation directions).*

The objective is therefore to maximize the quadratic ratio between the source signal power and the interference signal power according to

$$\mathbf{v_{max}}^{(f)} = \arg\max_{\mathbf{v^{(f)}}} \left\{ \frac{\mathbf{v}^{(f)H}\hat{\mathbf{R}}_{ss}^{(f)}\mathbf{v}^{(f)}}{\mathbf{v}^{(f)H}\hat{\mathbf{R}}_{ii}^{(f)}\mathbf{v}^{(f)}} \right\}. \tag{5}$$

The source correlation matrix, $\hat{\mathbf{R}}_{ss}^{(f)}$, and the interference correlation matrix, $\hat{\mathbf{R}}_{ii}^{(f)}$, are estimated from received data when each component, source and interference, are individually active.

The solution of this optimization problem can be found from the eigenvectors, $\mathbf{v}^{(f)}$, complying with

$$(\hat{\mathbf{R}}_{ii}^{(f)^{-1/2}})^H\, \hat{\mathbf{R}}_{ss}^{(f)}\, \hat{\mathbf{R}}_{ii}^{(f)^{-1/2}} \mathbf{v}^{(f)} = \lambda \mathbf{v}^{(f)}, \tag{6}$$

in the order of decreasing corresponding eigenvalues, i.e. the optimal solution is the eigenvector belonging to the maximum eigenvalue. Hence, the transformation matrix $\mathbf{V}^{(f)}$ is chosen to be the eigenvector matrix of the matrix $(\hat{\mathbf{R}}_{ii}^{(f)^{-1/2}})^H\, \hat{\mathbf{R}}_{ss}^{(f)}\, \hat{\mathbf{R}}_{ii}^{(f)^{-1/2}}$.

One way to reduce the complexity of the problem, without any significant loss of information, is to reduce the number of eigenvectors in $\mathbf{V}^{(f)}$ such that the most significant eigenvectors are used. As a result, the dimension of the input vector, and consequently the dimension of the correlation matrix and weight vector in Eq. (2), is reduced.

# 4 Simulations and Results

The performance of the beamformer was evaluated in a car hands-free telephony environment with a linear six-microphone array mounted on the visor at the passenger side, see Fig. 3. The measurements were performed in a Volvo station wagon. The speech originating from the passenger position constitutes the desired source signal and the hands-free loudspeaker emission is the source interfering signal, while the ambient noise received in a moving car constitutes the background noise. A loudspeaker was mounted at the passenger seat to simulate a real person engaging a conversation. The sensors used in this evaluation were six high quality Sennheiser microphones uniformly spaced in-line with 5 cm spacing. The microphone-array was positioned at a distance of 35 cm from the artificial speaker.

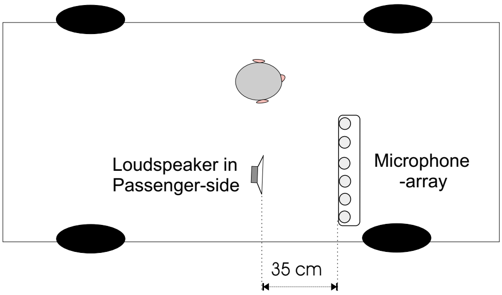


Figure 3: *Placement of the microphone-array in the car. The distance between microphone centers is 5 cm.*

Data was gathered on a multichannel DAT-recorder with a sampling rate of 12 KHz, and with a 300-3400 Hz bandwidth.

The desired source calibration signals were initially recorded when a speech sequence was emitted from the artificial talker, in a non-moving car with the engine turned off. Similarly, interference calibration signals were recorded by emitting a different speech sequence, from the hands-free loudspeaker alone, within the bandwidth.

In order to evaluate the proposed beamformer's new structure, a set of weights were calculated according to Eq. (2), based on correlation estimates calculated from source input data. The performance evaluation includes source speech distortion and suppression of both background noise and hands-free loudspeaker interference as well as computational complexity.

In the particular case of a car scenario, the interference sound power originated from the hands-free loudspeaker is low compared to the noise generated by the wind, the car engine and tire friction. Based on this observation, the interfering source signal is considered as part of the surrounding noise, and the optimization of (5) is simplified by replacing the interference correlation matrix estimate $\hat{\mathbf{R}}_{ii}^{(f)}$ by the identity matrix of size $I$. Consequently, the transformation matrix $\mathbf{V}^{(f)}$ is chosen to be the eigenvector matrix of the source calibration signals, $\hat{\mathbf{R}}_{ss}^{(f)}$.

Fig. 4 presents the results of a comparison between the original Constrained Subband beamformer and the reduced-rank beamformer based on the proposed space-time filter-bank structure described in Sec. 3, when it comes to noise suppression (top subplot), interference suppression (middle subplot) and speech distortion (bottom subplot). The effect of reducing the rank of the beamformer, with one or two dimensions, on the resulting performances is also presented. The figure shows that by using the spatial decomposition of the input data (following Eq. (4)) prior to the beamforming process, the distortion of the speech is considerably decreased while the noise and interference suppression is relatively unchanged. Furthermore, the reduction from six to four dimensions for this scenario maintains the performance of the algorithm.

In Fig. 5, the computational complexity gain of the proposed method is evaluated when reducing the rank of the beamformer algorithm. A considerable reduction of computational complexity is obtained when less dimensions are used.

# 5 Conclusion

A new beamforming algorithm based on a spacial filter bank design method has been presented and evaluated on real-world recordings in a car hands-free situation.

Results with the new method were compared to the ones obtained from the original constrained subband beamformer. The results clearly show that by directing the beamformer input-vector towards the source propagation direction, prior to the beamformer, the proposed method maintains the noise and interference suppression performances of the original subband beamformer, while decreasing the speech distortion with approximately 5 dB and reducing its computational complexity significantly.

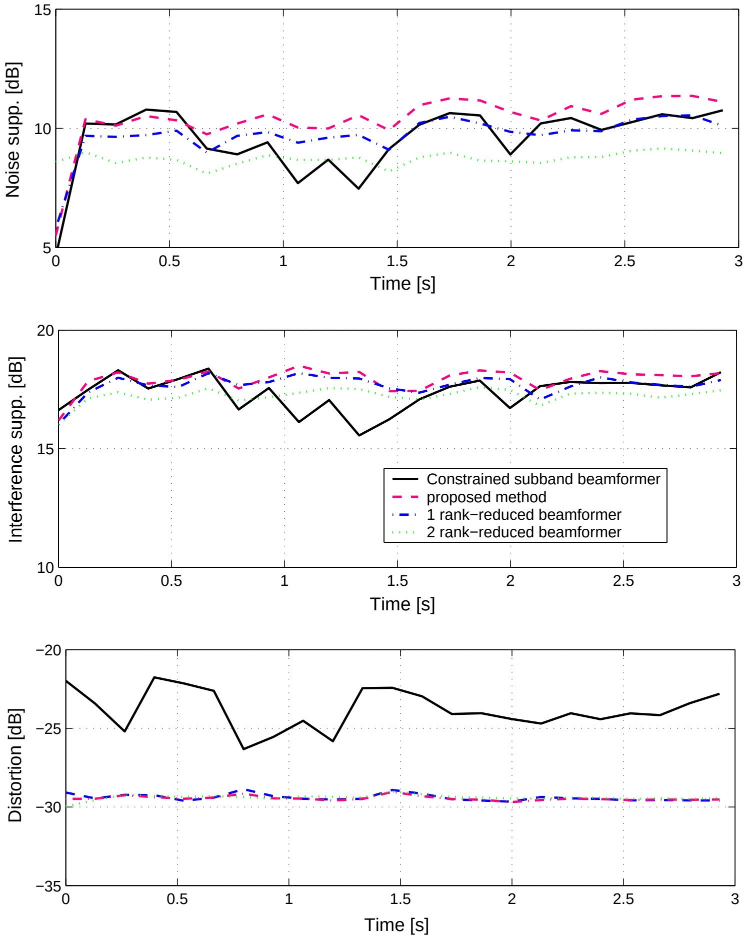


Figure 4: *Performance evaluation of the spatial subband beamformer, using an array of six microphones, in the case of 0, 1 and 2 dimensions reduced in the beamformer calculations (i.e. using a transformation matrix composed of, respectively, the 6, 5 and 4 most significant eigenvectors).*

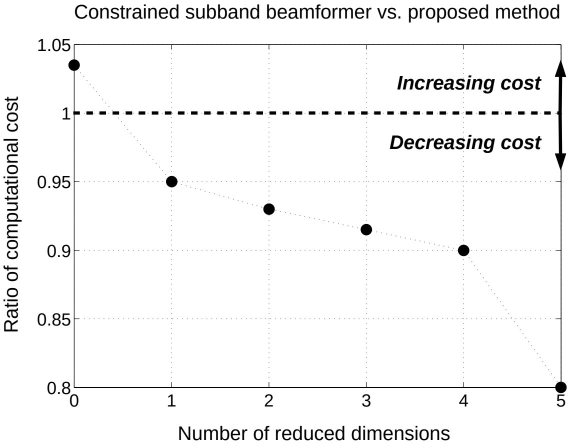


Figure 5: *Ratio of the computational cost between the reduced-rank proposed beamformer and the Constrained Subband beamformer. It can be seen that the proposed subband beamformer even with full-rank presents less than 5 per cent increase in computational cost, when compared to the original subband beamformer, while gaining considerably in performance.*

# References

[1] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.

[2] M. Brandstein and D. Ward, *Microphone Arrays, Signal Processing Techniques and Applications*, Springer, 2001.

[3] D. A. Florêncio and H. S. Malvar, “Multichannel filtering for optimum noise reduction in microphone arrays,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 197–200, May 2001.

[4] S. Affes and Y. Grenier, “A signal subspace tracking algorithm for microphone array processing of speech,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 5, no. 5, pp. 425–437, Sep. 1997.

[5] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, “Speech enhancement based on the subspace method,” *IEEE Trans. Acoust. Speech Signal Processing*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.

[6] Z. Yermeche, P Márquez Garcia, N. Grbić and I. Claesson, “A Calibrated Subband Beamforming Algorithm for Speech Enhancement,” in *IEEE Sensor Array and Multichannel Signal Processing Workshop*, August 2002.

[7] N. Grbić and S. Nordholm, “Soft constrained subband beamforming for hands-free speech enhancement,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp.885-888, May 2002.

[8] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, “Design of oversampled uniform dft filter banks with delay specifications using quadratic optimization,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. VI, pp. 3633–3636, May 2001.

Part III

# Beamforming for Moving Source Speech Enhancement

**Part III is submitted in its original version as:**

Z. Yermeche, N. Grbić and I. Claesson, “Beamforming for Moving Source Speech Enhancement,” submitted to IEEE Transactions on Speech and Audio Processing, December 2004.

# Beamforming for Moving Source Speech Enhancement

Z. Yermeche, N. Grbic and I. Claesson

## Abstract

This paper presents a new constrained subband beamforming algorithm to enhance speech signals generated by a moving source in a noisy environment. The beamformer is based on the principle of a soft constraint defined for a specified region corresponding to the source location. The soft constraint secures the spatial-temporal passage of the desired source signal in the adaptive update of the beamforming weights. The source of interest is modelled as a cluster of stationary point sources and source motion is accommodated by revising the point source cluster. The source modelling and its direct exploitation in the beamformer through covariance estimates are presented. An algorithm for sound source localization is used for speaker movement tracking and this information is exploited to update the spatial distribution in the source model.

Evaluation in a real environment with a moving speaker shows a significant noise and hands-free interference suppression within the conventional telephone bandwidth. This is achieved with a negligible impact on target signal distortion.

# 1 Introduction

Microphone arrays in conjunction with digital beamforming techniques have been extensively exploited for speech enhancement in hands-free applications, such as conference telephony, speech recognition and hearing aid devices [1, 2, 3]. In a hands-free environment, microphones are placed at a remote distance from the speakers causing problems of room reverberation, noise and acoustic feed-back. Successful microphone array processing of speech should achieve speech dereverberation, efficient noise and interference reduction, and should also provide an adaptation capacity to speaker movement.

In the microphone array literature many algorithms address these issues separately. The Generalized Sidelobe Canceller (GSC), predominantly used for noise suppression, has proven to be sensitive to reverberation [4, 5]. Other beamforming techniques using optimal filtering or signal subspace concepts have been suggested [6, 7, 8, 9]. Many of these algorithms rely on voice activity detection. This is needed in order to avoid source signal cancellation effects [1], which may result in unacceptable levels of speech distortion.

This paper proposes a novel constrained subband beamformer based on the principle of a soft constraint calculated from an estimated source position and a known array geometry [10, 11, 12], rather than formed from calibration data [13, 14]. This approach allows for an efficient adaptation of the beamformer to speaker movement by using a tracking algorithm for sound source localization. A SRP-PHAT algorithm is therefore used in conjunction with the filtering operations [15].

The computational complexity of the beamformer is substantially reduced by introducing a subband beamforming scheme [10, 13, 16]. The objective is formulated in the frequency domain as a weighted recursive least squares solution. In order to track variations in the surrounding noise environment, the proposed algorithm continuously estimates the spatial information for each frequency band, based on the received data. The update of the beamforming weights is done recursively where the initially pre-calculated correlation estimates constitute a soft constraint.

Performance of the proposed algorithm is evaluated on real data recordings conducted in an office environment with a moving source.

## 2 Signal model

Consider an acoustical environment where a speech signal coexists with interfering hands-free signals and diffuse ambient noise. This sound field is observed by an array with $I$ microphones. The speech source is considered as a spatially spread source and it is modelled as an infinite number of stationary and independent point sources clustered closely in space within a range of radii $[R_a, R_b]$ and inside the range of arrival angles $[\theta_a, \theta_b]$, see Fig. 1.

The direct path from a point source $m$, positioned at radius $R_m$ and angle $\theta_m$ from the center of the microphone array, has a response vector given by

$$\mathbf{d}(\Omega, R_m, \theta_m) = \left[ \frac{1}{R_{m,1}} e^{-j\Omega\tau_1(R_m,\theta_m)}, ..., \frac{1}{R_{m,I}} e^{-j\Omega\tau_I(R_m,\theta_m)} \right] , \quad (1)$$

where $R_{m,i}$ is the distance between the sound source $m$ and the sensor $i$, $\tau_i(R_m, \theta_m)$ is the time delay from the point source to the sensor $i$ and $\Omega$ is the angular frequency.

Thus, the received microphone input vector generated by the speech signal is expressed as

$$\mathbf{x}_s^{(\Omega)}(t) = \int_{R_a}^{R_b} \int_{\theta_a}^{\theta_b} s_m^{(\Omega)}(t) \mathbf{d}(\Omega, R_m, \theta_m) dR_m d\theta_m \, , \quad (2)$$

where $s_m^{(\Omega)}(t)$ is the signal component of the point source with spherical coordinates $(R_m, \theta_m)$, for a frequency $\Omega$.

The spatial correlation matrix is then given by

$$\begin{aligned} \mathbf{R}_{ss}^{(\Omega)} &= E[\mathbf{x}_s^{(\Omega)}(t) \mathbf{x}_s^{(\Omega)^H}(t)] \\ &= P(\Omega) \int_{R_a}^{R_b} \int_{\theta_a}^{\theta_b} \mathbf{d}(\Omega, R_m, \theta_m) \mathbf{d}(\Omega, R_m, \theta_m)^H \, dR_m d\theta_m \, , \quad (3) \end{aligned}$$

where $P(\Omega) = E[s_m^{(\Omega)}(t) s_m^{(\Omega)}(t)^H]$ is the source power spectral density, PSD, at frequency $\Omega$.

The speech source area is defined as a pie slice region as depicted in Fig. 1. This configuration is appropriate to contain the consequences of errors in the response vector, caused by the misplacement and gain variations of the microphones as well as by the error in the source position estimate. It has been shown that small errors in the response vector cause large radial errors in the corresponding source location [16], for sources that are outside the extreme near-field. In the extreme near-field, i.e. when the source is within a radius smaller than the array width, we can expect large angular errors in the corresponding source location.

## 2.1 Discrete-Space Formulation

To accommodate for the speaker's mobility, the pre-calculated spatial correlations should be updated whenever a movement of the speaker is detected. This update is more conveniently implemented by using a discrete-space model of the speech source rather than the continuous-space model described in the previous section. By using a finite number $M$ of stationary and independent point sources clustered closely within the constraint area, see Fig. 2, the spatial correlation matrix of Eq. (3) can be approximated by

$$\tilde{\mathbf{R}}_{ss}^{(\Omega)} = P(\Omega) \sum_{m=1}^{M} \mathbf{d}(\Omega, R_m, \theta_m) \mathbf{d}(\Omega, R_m, \theta_m)^H \, \Delta A_m, \tag{4}$$

where $\sum_{m=1}^{M} \Delta A_m$ covers the whole constraint area.

The error introduced by this approximation can be made negligible by using a high enough number $M$ of point sources in the calculation. The discrete-space formulation of the speech source correlations in Eq. (4) requires far fewer computations in comparison to the continuous-space formulation of Eq. (3) and is therefore more simply updated.

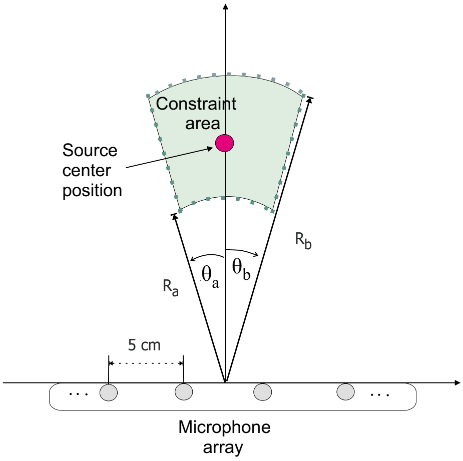


Figure 1: *Constraint region defined by the radii* $[R_a, R_b]$ *and the angles* $[\theta_a, \theta_b]$.

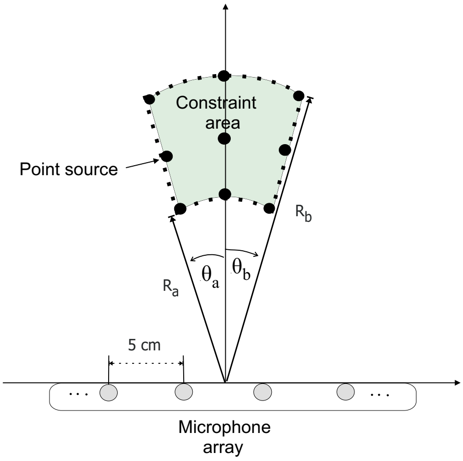


Figure 2: *Constraint region defined by the radii* $[R_a, R_b]$ *and the angles* $[\theta_a, \theta_b]$, *with 9 point sources.*

# 3 Soft constrained Beamforming

## 3.1 Objective

The objective of the soft constrained beamformer is formulated in the frequency domain as a combination of Least Squares and the Wiener solution, where the source covariance matrix obtained for a finite number of points in a specified constraint region constitutes a soft constraint. The constraint region denotes the area in which the speech source should be located.

An SRP-PHAT algorithm for sound source localization, described in [15], is used in this paper. It has shown to be robust and accurate in localizing speech for office applications and exhibits a fast tracking of speaker movement.

## 3.2 SRP-PHAT Algorithm

The most widely used source localization approach exploits time-difference of arrival (TDOA) information of a signal originated from a point in space, received by a pair of spatially separated microphones [2]. The time-delay can be estimated by maximizing the cross-correlation between filtered versions of the received signals, which is the basis of the Generalized Cross Correlation (GCC) method. This approach is however impractical when the signal of interest is corrupted by noise and reverberation [17]. This problem can be circumvented by equalizing the frequency-dependent weightings of the cross-spectrum components. The extreme case where the magnitude is flattened is referred to as the Phase Transform (PHAT). The GCC-PHAT algorithm computes the TDOA, $\tau_s$, for the source signal received by the microphone pair $l$ and $k$ as

$$\tau_s = \arg\max_{\tau_{\mathbf{lk}}} \left\{ \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{X_l(\omega)X_k(\omega)^H}{|X_l(\omega)X_k(\omega)^H|} e^{jw\tau_{\mathbf{lk}}}\, d\omega \right\} . \tag{5}$$

A merge of the GCC-PHAT algorithm and the Steered Response Power (SRP) beamformer, resulted in the so called SRP-PHAT algorithm [15]. The goal of the SRP-PHAT algorithm is to combine the robustness of the steered beamformer and the insensitivity to received signal characteristics introduced by the PHAT approach. The SRP-PHAT algorithm is formulated as

$$\mathbf{q}_s = \arg\max_{\mathbf{q}} \left\{ \sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} \frac{X_l(\omega)X_k(\omega)^H}{|X_l(\omega)X_k(\omega)^H|} e^{jw\Delta_{lk}(\mathbf{q})}\, d\omega \right\} , \tag{6}$$

where $\mathbf{q}_s$ is the true spatial location of the source, and where $\Delta_{lk}(\mathbf{q})$ is the TDOA for the signal generated from the spatial location $\mathbf{q}$, and received by the microphone pair $l$ and $k$.

Assuming a far-field scenario (i.e. a propagating plane wave), the TDOA of a sound signal for a pair of microphones from a linear array, can be expressed as a multiple of the corresponding time-delay between adjacent microphones. Hence, the far-field SRP-PHAT algorithm is derived from Eq. 6 as [15]

$$\tau_s = \arg\max_{\tau} \left\{ \sum_{l=1}^{I} \sum_{k=1}^{I} \int_{-\infty}^{+\infty} \frac{X_l(\omega) X_k(\omega)^H}{|X_l(\omega) X_k(\omega)^H|} e^{jw\tau[l-k]} \, d\omega \right\} . \tag{7}$$

## 3.3 Beamformer Algorithm

The filtering operations of the beamformer are formulated in the frequency domain as multiplications with number $I$ complex frequency domain weights, $w_j^{(\Omega)}(n)$. For a specific frequency $\Omega$, and at a sample instant $n$, the output is given by

$$y^{(\Omega)}(n) = \sum_{l=1}^{I} w_l^{(\Omega)}(n) x_l^{(\Omega)}(n) \, , \tag{8}$$

where the signals $x_l^{(\Omega)}(n)$ are digitally sampled and bandpass filtered microphone observations and $y^{(\Omega)}(n)$ corresponds to the beamformer output. These time domain signals are narrow-band, containing essentially components with frequency $\Omega$.

Given a known array geometry and a corresponding constraint region, the objective of the beamformer is formulated in the frequency domain as a calibrated weighted recursive least squares solution, where the optimal array weight vector is given by

$$\mathbf{w}_{ls,opt}^{(\Omega)}(n) = \left[ \tilde{\mathbf{R}}_{ss}^{(\Omega)} + \hat{\mathbf{R}}_{ii}^{(\Omega)} + \hat{\mathbf{R}}_{xx}^{(\Omega)}(n) \right]^{-1} \tilde{\mathbf{r}}_s^{(\Omega)} \, , \tag{9}$$

with the array elements arranged as

$$\mathbf{w}_{ls,opt}^{(\Omega)}(n) = [w_1^{(\Omega)}(n), \quad w_2^{(\Omega)}(n), \quad \ldots \quad w_I^{(\Omega)}(n)]^T .$$

The spatial source covariance matrix, $\tilde{\mathbf{R}}_{ss}^{(\Omega)}$, is given in Eq. (4), while the cross covariance vector, $\tilde{\mathbf{r}}_s^{(\Omega)}$, is given by the response vector of the $M$ point sources and the source PSD

$$\tilde{\mathbf{r}}_s^{(\Omega)} = P(\Omega) \sum_{m=1}^{M} \mathbf{d}(\Omega, R_m, \theta_m). \tag{10}$$

An initial calibration procedure is used to calculate the interference correlation matrix estimate, $\hat{\mathbf{R}}_{ii}^{(\Omega)}$, for $P$ sample observations when the known source interferences are active alone as

$$\hat{\mathbf{R}}_{ii}^{(\Omega)} = \sum_{p=1}^{P} \mathbf{x}_i^{(\Omega)}(p) \mathbf{x}_i^{(\Omega)^H}(p)\,, \tag{11}$$

where $\mathbf{x}_i^{(\Omega)}(p)$ is the received interference array data vector. Here, we assume that the interfering sources will maintain a fixed position in relation to the microphone array.

Conversely, the received signal correlation estimates, $\hat{\mathbf{R}}_{xx}^{(\Omega)}(n)$, are continuously calculated from observed data by

$$\hat{\mathbf{R}}_{xx}^{(\Omega)}(n) = \sum_{p=0}^{n} \lambda^{n-p} \mathbf{x}^{(\Omega)}(p) \mathbf{x}^{(\Omega)^H}(p)\,, \tag{12}$$

where $\mathbf{x}^{(\Omega)}(p)$ is the received array data vector with frequency $\Omega$ and $\lambda$ is a forgetting factor, with the purpose of tracking variations in the surrounding noise environment.

The update of the beamforming weights is done recursively as in [13], by iteratively using the matrix inversion lemma [1].

## 3.4 Movement tracking

A judicious choice for the configuration of the $M$ point sources in the constraint area allows for an efficient recursive update of the received source covariance matrix and cross covariance vector consequent to the speaker's mobility. As a change in source position occurs, the set of points in Eqs. (4) and (10) is altered in such a way that new points are added while previously calculated points are subtracted, to reflect this change, see Fig. 3.

One way to achieve a smooth and efficient adaptation to speaker movement of the constraint area statistics is to introduce a weighing factor on the received source covariance matrix. At the same time, the contribution from new points corresponding to the actual position of the speech source is added.

By introducing a uniformly spread configuration of the $M$ point sources in the constraint area, i.e. a radius and angular equidistant point structure, the movement of the source can be expressed as a function of the angular distance $\Delta\theta$ and the radius distance $\Delta R$ between adjacent point sources. When a source is moving from an angular position $\theta$ with an angle shift $\Delta\theta$, as depicted in Fig. 3, the correlation matrix is updated by the quantity

$$\Delta\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta+\Delta\theta) = P(\Omega) \sum_{m\in M_{\theta+\Delta\theta}} \mathbf{d}(\Omega, R_m, \theta+\Delta\theta)\mathbf{d}(\Omega, R_m, \theta+\Delta\theta)^H \,, \tag{13}$$

where $M_{\theta+\Delta\theta}$ is the cluster of points having an angle $(\theta+\Delta\theta)$ and contained in the constrained area. Additionally, a forgetting factor, $0 < \zeta < 1$, is used in the update of the correlation matrix during movement following

$$\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta+\Delta\theta) = \zeta\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta) + \Delta\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta+\Delta\theta)\,. \tag{14}$$

Thus, for a movement of the speech source from a position with angle $\theta_{old}$ to a position with angle $\theta_{new}$, see Fig. 4, the update of the correlation matrix is formulated as

$$\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta_{new}) = \zeta^L\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta_{old}) + \sum_{l=1}^{L}\zeta^{L-l}\Delta\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta+l\Delta\theta) =$$
$$\zeta^L\tilde{\mathbf{R}}_{ss}^{(\Omega)}(R,\theta_{old}) + P(\Omega)\sum_{l=1}^{L}\zeta^{L-l}\sum_{m\in M_l}\mathbf{d}(\Omega,R_m,\theta_{old}+l\Delta\theta)\mathbf{d}(\Omega,R_m,\theta_{old}+l\Delta\theta)^H. \tag{15}$$

Here $M_l$ refers to the cluster of points with angle $(\theta_{old}+l\Delta\theta)$ in the constraint area and $L$ is the number of angular steps $\Delta\theta$ covered by the speaker's movement. Similarly, the received cross correlation vector is updated by

$$\tilde{\mathbf{r}}_{s}^{(\Omega)}(R,\theta_{new}) = \zeta^L\tilde{\mathbf{r}}_{s}^{(\Omega)}(R,\theta_{old}) + P(\Omega)\sum_{l=1}^{L}\zeta^{L-l}\sum_{m\in M_l}\mathbf{d}(\Omega,R_m,\theta_{old}+l\Delta\theta). \tag{16}$$

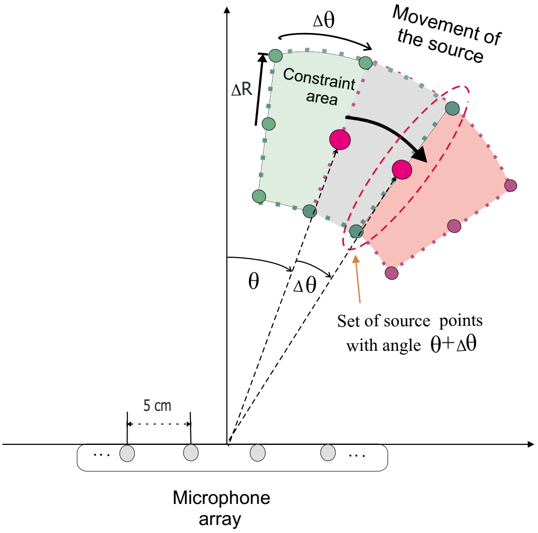


Figure 3: *Constraint region defined for a moving source by $\Delta\theta$, with a uniformly spread configuration of 9 point sources in the constraint area.*

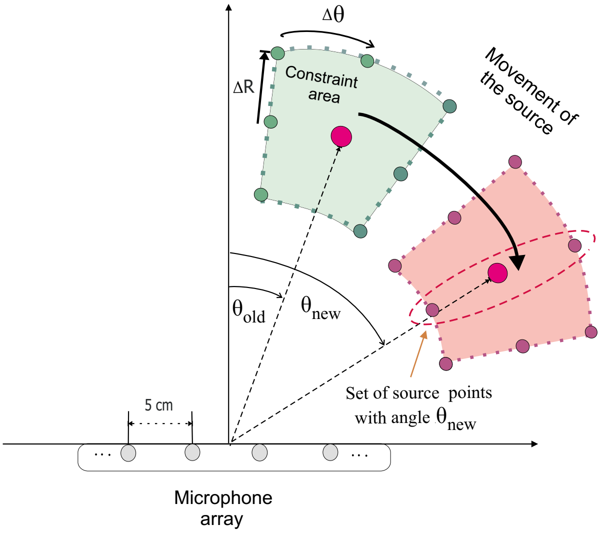


Figure 4: *Constraint region defined for a moving source from a position with angle $\theta_{old}$ to a position with angle $\theta_{new}$, with a uniformly spread configuration of 9 point sources in the corresponding constraint area.*

## 3.5 Subband Beamforming

The approach derived in Sec. 3.3 requires a frequency decomposition of each microphone input signal into a set of narrow-band signals. A multichannel uniform over-sampled analysis filter bank is used to decompose the received array signals into narrow band signals, prior to the beamformer filtering operations. The outputs of the subband beamformers are reconstructed by a synthesis filter bank in order to create a full band time domain output signal. The analysis and synthesis filter banks constitute a modulation of two prototype filters, which leads to efficient polyphase realization [18]. The spatial characteristics of the input signal are maintained by using the same modulated filter bank at each microphone.

# 4 Simulations

Simulations are constructed such to include influence of background noise, hands-free interference as well as room reverberation. The data was acquired with a linear array of four microphones uniformly spaced with 5 cm spacing and was gathered on a multichannel DAT-recorder with a sampling rate of 12 KHz. The signal at each microphone was bandlimited to 300-3400 Hz. All simulations were performed with 32 subbands.

## 4.1 Office Environment

The room used in the experiment is a room of size (3 × 4 × 3 m), with the microphone array placed in the center of the room. Fig. 5 illustrates the arrangement for the sound data acquisition. To simulate ambient noise, four loudspeakers were positioned at the corners of the room. The emitted noise was colored noise prerecorded in a typical office room and corresponding to the noise generated by a computer fan and an air conditioner. Two experimental setups where considered. The first setup consisted of a target source, simulated by a loudspeaker, positioned at a 50 cm radius in front of the array and an interference source at a radius of 1 m from the array and making an angle of 45° to the source. The sound used as target source was female speech while the interfering loudspeaker emitted male speech. The second setup was defined for a moving target source from the position described above, along a circular path centered at the reference point to describe a quarter of a circle. The numbered crosses depicted in Fig. 5 correspond to the speaker positions where the update of the SRP-PHAT algorithm takes place. The time-interval to move between two adjacent speaker positions was one second.

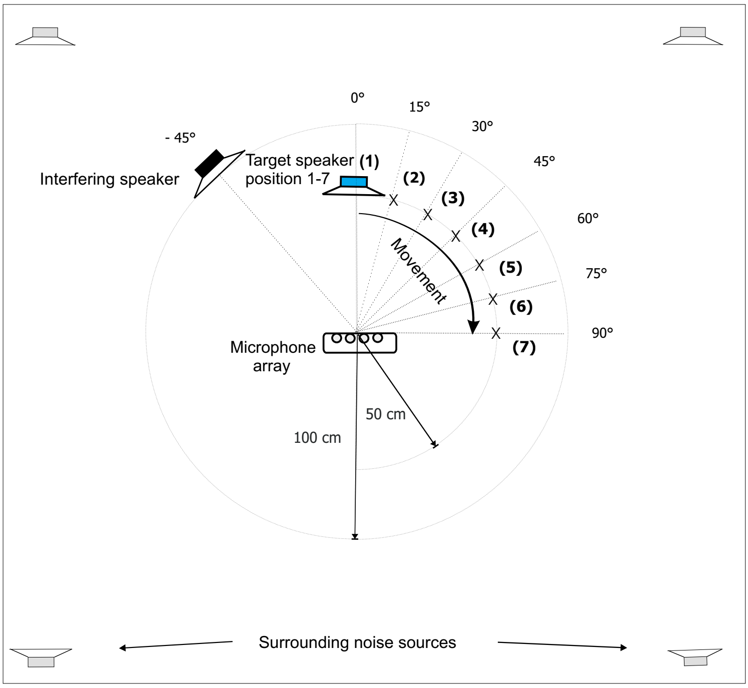


Figure 5: *Configuration of microphone array and sound sources in an office room. The source movement path is represented by the numbered crosses, which are passed at a speed of one second per step.*

# 5 Evaluation

In order to evaluate the proposed beamformer, the weights were calculated for the first setup according to Eq. (9). Interference calibration signals were initially recorded by emitting a speech sequence from the interfering loudspeaker alone within the bandwidth. These recordings were used to estimate the subband interference correlation matrix given in Eq. (11), for all subbands. The performance evaluation includes source speech distortion and suppression of both background noise and hands-free loudspeaker interference.

### Performance versus Number of Point Sources

The source correlations were obtained for a constraint region defined by a range of radii [40 cm, 60 cm] and inside the range of arrival angles [-5°, 5°], i.e. radius spread of 20 cm and angle spread of 10°, centered at the position of the source given by the tracking algorithm. Table 1 shows the algorithm performance by varying the number of points, $M$, in the constraint region used to calculate the received spatial source covariance estimates from Eqs. (4) and (10). It can be seen that the number of points used does not severely affect the performance of the algorithm. For the chosen constraint region size, increasing $M$ above 9 points does not improve the performance of the beamformer.

| Number of points | Noise sup. [dB] | Interf. sup. [dB] | Distortion [dB] |
|---|---|---|---|
| 4 | 8.6 | 17.6 | -28.7 |
| 9 | 10.3 | 19.1 | -28.1 |
| 16 | 8.1 | 14.9 | -29.8 |
| 25 | 7.3 | 16.8 | -28.5 |
| 36 | 8.4 | 16.5 | -28.8 |

Table 1: *Distortion and suppression levels for different number of source points.*

### Performance versus Angle of Source Spread

The performance of the algorithm is given in Fig. 6 as a function of angle spread of the constraint region. The number of point sources used is $M = 9$

and the radius spread of the constraint region is 20 cm. It can be seen that the beamformer achieves optimal performance for an angle spread of 10°. A small angle spread (below 10°) does not cover both the source spread and the localization uncertainty, resulting in reduced performance. For a larger angle spread (above 10°), the opening of the beamformer is wider and can therefore allow passage for a bigger portion of the diffuse noise within the corresponding angle spread.

### Performance versus Radius of Source Spread

The dependency on the constraint region radius spread is shown in Fig. 7. The performance evaluation is given for an angle spread of 10° and 30°. Results show that the optimum, reached for a certain radius spread of the constraint region, depends on the angle spread. A constraint region with bigger angle spread requires a bigger radius spread in order to reach optimum. It can be seen that the performance slightly decreases with higher values of radius spread, due to less noise being suppressed in the extended constraint region. However, the proposed beamformer exhibits a relatively higher sensitivity to angle spread than that to radius spread.

### Performance versus Movement of Speaker

The impact on the beamformer performance of the source movement is evaluated using the second setup. A constraint area of size (20 cm, 30°) with 9 points is used. It is centered at each coordinate of the source obtained by the SRP-PHAT speech localization algorithm. Fig. 8 shows the speech distortion and noise suppression measures for the different positions given by the tracking algorithm and corresponding to the speaker's movement. The performance measures are plotted for different values of the weighting factor $\zeta$. The proposed solution exhibits a good tracking capability when the speaker is moving. The noise suppression is slightly improved with higher values of the weighting factor $\zeta$. As the angle of the source increases, corresponding to a movement away from the front view of the sensor array, more noise is suppressed.

### Performance versus Source Tracking Error

An error may be introduced by the tracking algorithm when evaluating the position of the speaker. The algorithm performance is given in Fig. 9 as a function of the source location error. Here, only the angle error has been considered. Due to the small sensitivity of the beamformer to the constraint region radius spread, a large radius spread can be used to compensate for the radial error in the estimation of the source position. The source correlations were obtained for a constraint region of size (20 cm, 10°). The beamformer has proven to be robust to position tracking errors. It presents a relative small decrease in performance ($\sim$ 2.5 dB) with an error increase of 10°.

# 6 Conclusion

A new adaptive subband beamformer for moving source speech enhancement and its evaluation in an office room has been presented. The proposed beamformer is a recursive least squares algorithm using a soft constraint defined for a specified region corresponding to the speech source location. A speech localization algorithm is combined with the beamformer to allow for speaker movement. The update of the soft constraint due to source mobility is performed recursively. Results show up to 9 dB noise reduction and 15 dB hands-free suppression. Furthermore, the algorithm provides good speaker movement tracking, while keeping an approximately constant level of speech distortion as for a non-moving source.

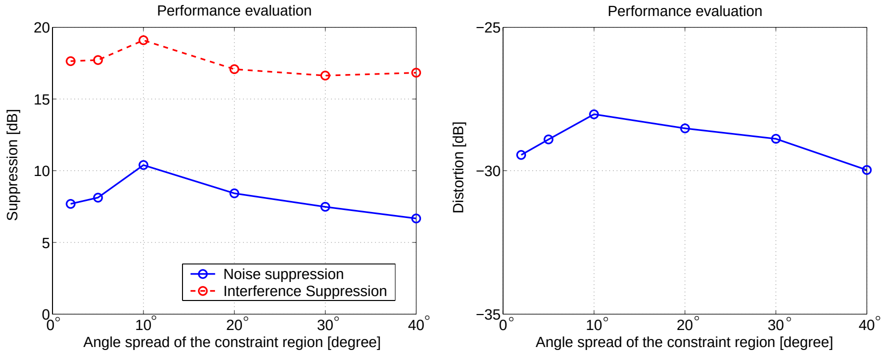


Figure 6: *Performance measures of the proposed beamformer with a different angle spread of the constraint region. Here, the radius spread is 20 cm and the number of source points considered is M=9.*

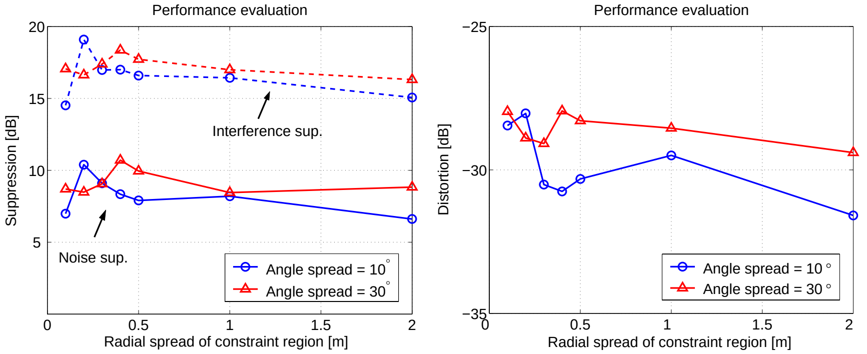


Figure 7: *Performance measures of the proposed beamformer with a different radius spread of the constraint region. The number of source points is M=9.*

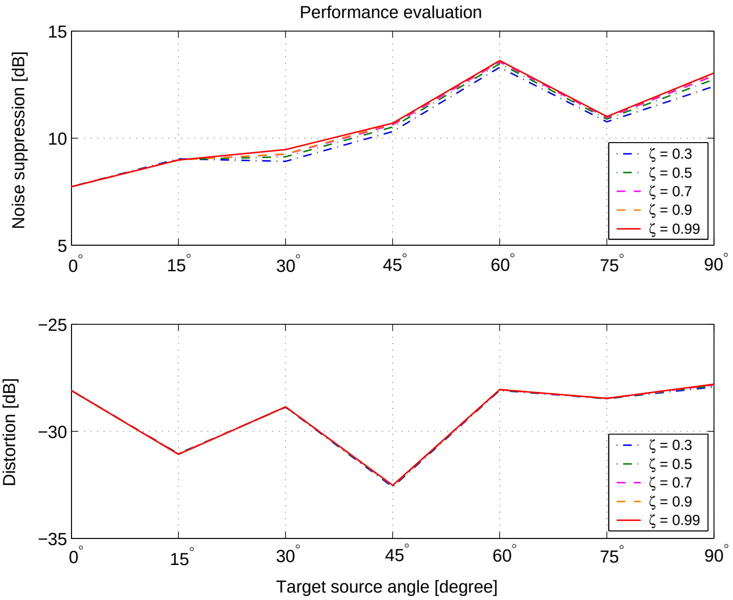


Figure 8: *Performance measures of the proposed beamformer versus speaker angular position, with a different weighting factor* $\zeta$.

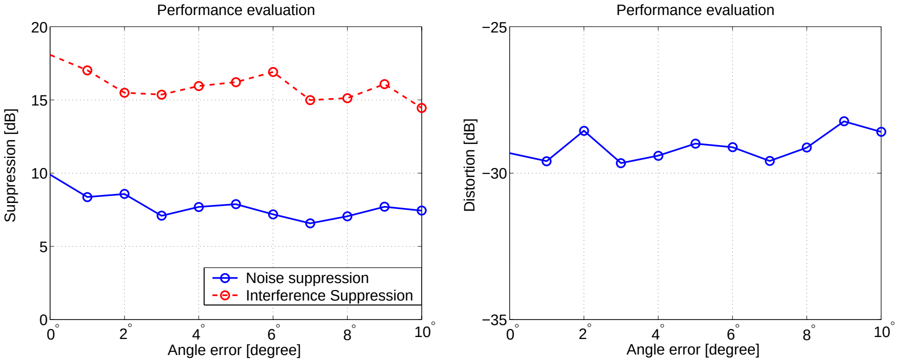


Figure 9: *Performance measures of the proposed beamformer as a function of the angle error in the estimation of the source position.*

# References

[1] D. Johnson and D. Dudgeon, *Array Signal Processing - Concepts and Techniques*, Prentice Hall, 1993.

[2] M. Brandstein, and D. Ward (Eds.), "Microphone Arrays - Signal processing Techniques and applications," Springer, 2001.

[3] X. Zhang, and J. H. L. Hansen, "CSA-BF: Novel Constrained Switched Adaptive Beamforming for Speech Enhancement and Recognition in Real Car Environment," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2003.

[4] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constraind adaptive filters," in *IEEE Transactions on Signal Procrocessing*, vol. 47, no. 10, pp. 2677-2684, June 1999.

[5] J. Bitzer, K.U. Simmer, and K.D. Kammeyer, "Theorethical noise reduction limits of the generalized sidelobe canceller (GSC) for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 5, pp. 2965-2968, May 1999.

[6] D. A. Florêncio, and H. S. Malvar, "Multichannel filtering for optimum noise reduction in microphone arrays," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, pp. 197–200, May 2001.

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

[8] F. Asano, S. Hayamizu, T. Yamada, and S. Nakamura, "Speech enhancement based on the subspace method," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 8, no. 5, pp. 497 – 507, Sep. 2000.

[9] D. D. Feldman, and L. J. Griffiths, "A projection Approach to Robust Adaptive Beamforming," *IEEE Transactions on Signal Processing*, vol. 42, pp. 867 – 876, April 1994.

[10] S. Y. Low, N. Grbić, and S. Nordholm, "Speech enhancement using Multiplt soft Constrained Beamformers and Non-Coherent Technique," in

*IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2003.

[11] S. Y. Low, N. Grbić, and S. Nordholm, “Subband Generalized Sidelobe Canceller - A Constrained Region Approach,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, October 2003.

[12] H. Q. Dam, S. Y. Low, H. H. Dam, and S. Nordholm, “Space Constrained Beamforming with Source PSD Updates,” in *International Conference on Acoustics, Speech and Signal Processing*, May 2004.

[13] Z. Yermeche, P. M. Garcia, N. Grbić, and I. Claesson, “A Calibrated Subband Beamforming Algorithm for Speech Enhancement,” in *IEEE Sensor Array and Multichannel Sig. Proc. Workshop*, August 2002.

[14] S. Shahbazpanahi, A.B. Gershman, Z. Q. Luo, and K. M. Wong, “Robust adaptive beamforming for general-rank signal model,” in *IEEE Transactions on Signal Processing*, Vol. 51 , no. 9 , pp. 2257 2269, September 2003.

[15] A. Johansson, N. Grbić, and S. Nordholm, “Speaker Localisation Using the Far-field SRP-PHAT in Conference Telephony,” in *IEEE International Symposium on Intelligent Sig. Proc. and Comm. Systems*, November 2002.

[16] N. Grbić, and S. Nordholm, “Soft constrained subband beamforming for hands-free speech enhancement,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. I, pp. 885–888, May 2002.

[17] S. Bédard, B. champagne, and A. Stéphenne, “Effects of room reverberation on time-delay estimation performance,” in *International Conference on Acoustics, Speech and Signal Processing*, vol. II, pp. 261–264, April 1994.

[18] J. M. de Haan, N. Grbić, I. Claesson, and S. Nordholm, “Design of oversampled uniform DFT filter banks with delay specifications using quadratic optimization,” in *IEEE International Conf. on Acoust., Speech and Sig. Proc.*, vol. VI, pp. 3633–3636, May 2001.