Friedrich-Alexander-Universität
Erlangen-Nürnberg

Chair of Multimedia Communications and Signal Processing

Prof. Dr.-Ing. Walter Kellermann

Bachelor Thesis

Implementation and Evaluation of a Subspace Method as a Preprocessing Stage for Blind Source Separation

Daniel Kiesel

Supervisor: M.Sc. Hendrik Barfuss

Erlangen, September 2014

Lehrstuhl für Multimediakommunikation und Signalverarbeitung
Prof. Dr.-Ing. W. Kellermann

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Bachelorarbeit

für

Mr. cand. Ing. Daniel Kiesel

Implementierung und Evaluation einer Vorverarbeitungsstufe zur Verbesserung der Trennungsleistung eines blinden Quellentrennungsalgorithmus

Blinde Quellentrennung (*Blind Source Separation, BSS)* behandelt das Problem der Trennung einer Mischung von mehreren Signalen in Einzelsignale, die vor der Weiterverarbeitung - zum Beispiel mittels Spracherkennung - getrennt werden müssen. Bei BSS wird hierfür kein weiteres Wissen über die einzelnen Signale, wie zum Beispiel die Quellenposition, vorausgesetzt. Einzig wechselseitige statistische Unabhängigkeit der einzelnen Signale muss gegeben sein.

Nachhall hat einen großen Einfluss auf die Trennungsleistung eines BSS-Algorithmus. Generell gilt, je länger die Nachhallzeit, desto schlechter wird die Trennungsleistung des BSS-Systems. [Asano et al. 2003] veröffentlichten eine Vorverarbeitungstufe für ein BSS-System, bei der durch eine Unterraumprojektion der Mikrofonsignale der in den Mikrofonsignalen vorhandene Nachhall reduziert wird. Es wurde gezeigt, dass sich dadurch die Trennungsleistung des BSS Algorithmus verbessert.

In dieser Arbeit soll die Unterraumprojektion nach [Asano et al. 2003] implementiert und als Vorverarbeitungsstufe für einen am LMS entwickelten BSS-Algorithmus in verschiedenen akustischen Szenarien mit unterschiedlicher Nachhallzeit evaluiert werden. Zur Evaluierung der Trennungsleistung sollen sowohl signalbasierte Qualitätsmaße als auch Spracherkennungsraten verwendet werden. Die Implementierung soll in MATLAB durchgeführt werden. Auf klar strukturierten und kommentierten Programmcode wird großer Wert gelegt.

Beginn: 21.04.2014

Ende: 20.09.2014

(Prof. Dr.-Ing. Kellermann)

Erklärung

Ich versichere, dass ich die vorliegende Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.

Erlangen, den 19. September 2014

Daniel Kiesel

Abstract

Blind source separation (BSS) is the approach to separate observed linear mixtures in order to receive the distinct source signals without knowledge of the original signals or the mixing system. The major problems however, that significantly decrease BSS performance, are ambient noise and reverberation when the system is used in a real acoustic environment. Therefore, in this thesis, the Subspace Method, an algorithm to suppress those undesired interferences, is employed and evaluated as a preprocessor for a BSS algorithm.

Preliminary investigations are conducted to verify the basic functionality of the Subspace Method for one speaker. Thereby, different room conditions with varying reverberation times and different white Gaussian noise levels are examined. The results show solid improvements in highly reverberant environments as well as for all scenarios where white noise has been applied. In contrast, in the BSS scenario with two active speakers all experiments lead to inferior results when the Subspace Method is used, either by itself or as a preprocessor for BSS, compared to the scenarios where only BSS is applied. It is assumed that these result occur due to inaccurate permutation in the frequency domain as well as decorrelation through the Subspace Method.

Kurzfassung

Blinde Quellentrennung (blind source separation, BSS) bezeichnet den Ansatz aus einer Mischung von Signalen die Quellensignale zu rekonstruieren. Dies geschieht ohne jegliches Wissen über die Originalsignale oder den Mischprozess. Dabei hängt die Qualität der Signaltrennung vor allem von Störquellen und der Nachhallzeit eines Raumes ab. Demzufolge wird in dieser Arbeit die Unterraummethode, ein Algorithmus, der diese unerwünschten Störgeräusche unterdrückt, vorgestellt und als Vorverarbeitungsstufe für ein BSS System eingesetzt.

Die Evaluation lässt sich in zwei unterschiedliche Szenarien gliedern für die jeweils verschiedene Nachhallzeiten und Rauschintensitäten untersucht wurden. Einerseits wird die Grundfunktionalität der Unterraummethode verifiziert, indem sie auf Sprachsignale eines einzelnen Sprechers angewandt wird. Im Ergebnis wird deutlich, dass vor allem für lange Nachhallzeiten und additives weißes gaußsches Rauschen deutliche Verbesserungen gegenüber dem ungefilterten Signal erzielt werden können. Andererseits wird das BSS Szenario mit zwei aktiven Sprechern untersucht. Dabei werden, neben dem Einsatz der Unterraummethode als Vorverarbeitungsstufe für BSS, zu Vergleichszwecken auch die Ergebnisse von der alleinigen Anwendung von BSS und der Unterraummethode betrachtet. Hierbei führen die Szenarien in denen die Unterraummethode eingesetzt wird zu schlechteren Ergebnissen als eine alleinige Nutzung des BSS Algorithmus. Es wird angenommen, dass dieses Verhalten aus dem, durch die Unterraummethode erzeugten, Permutationsproblem im Frequenzbereich resultiert und aus der Tatsache, dass die Unterraummethode die Ausgangssignale bereits dekorreliert.

Contents

Chapter 1

Introduction

One of the most prominent paradigms in audio signal processing is the so called cocktail-party problem. Thereby, two speakers are talking simultaneously in a room while microphones are observing the situation. The resulting microphone signals consequently contain a mixture of both speech signals. For applications such as speech recognition systems it is very desirable to recover the original source signals based on this observed mixture, as a clear source signal is required for an automatic speech recognizer (ASR). This signal separation can be achieved by applying a blind source separation (BSS) system. In this context, the biggest obstacles that real applications face are ambient noise and reverberation due to multipath propagation. These interferences dramatically reduce the performance of the BSS algorithm and thus, ultimately, the performance of the ASR [3].

Therefore, in this thesis, an algorithm to suppress ambient noise and reverberation, the Subspace Method in the frequency domain introduced by Asano et al. [3], is presented and employed as a preprocessing stage for a BSS algorithm in order to investigate whether applying the Subspace Method can improve BSS performance.

The thesis is organized in the following way: In Chapter 2, the fundamentals of BSS are outlined and an overview about reverberation is given. Furthermore the concept of vector spaces and subspaces is described, before in Chapter 3 the Subspace Method for reducing room reflections is presented. After that, the experimental results for different scenarios are evaluated in Chapter 4 based on the performance measures introduced in advance. Finally, a summary of the main results as well as an outlook on future work is given in Chapter 5.

Chapter 2

Fundamentals

In this chapter, the fundamentals that are necessary to understand the Subspace Method presented in Chapter 3 and the following evaluation in Chapter 4 are outlined. Therefore, the basic principle of BSS as well as the underlying signal model are described. Afterwards, the theory about room reflections is illustrated and before examining how the Subspace Method can reduce room reflections, the key aspect of this approach, namely the concept of vector spaces and subspaces, is regarded.

2.1 Blind Source Separation

The general idea of BSS is to separate observed linear mixtures in order to recover the original distinct source signals. The term "blind" implies that the algorithm does not require any knowledge about the original signals or the mixing system which is a necessary prerequisite as most real world applications, such as hearing aids or video conferencing, do not offer such information [21]. Thereby, the mixing system simulates the general signal propagation as well as signal interferences. The only prerequisite

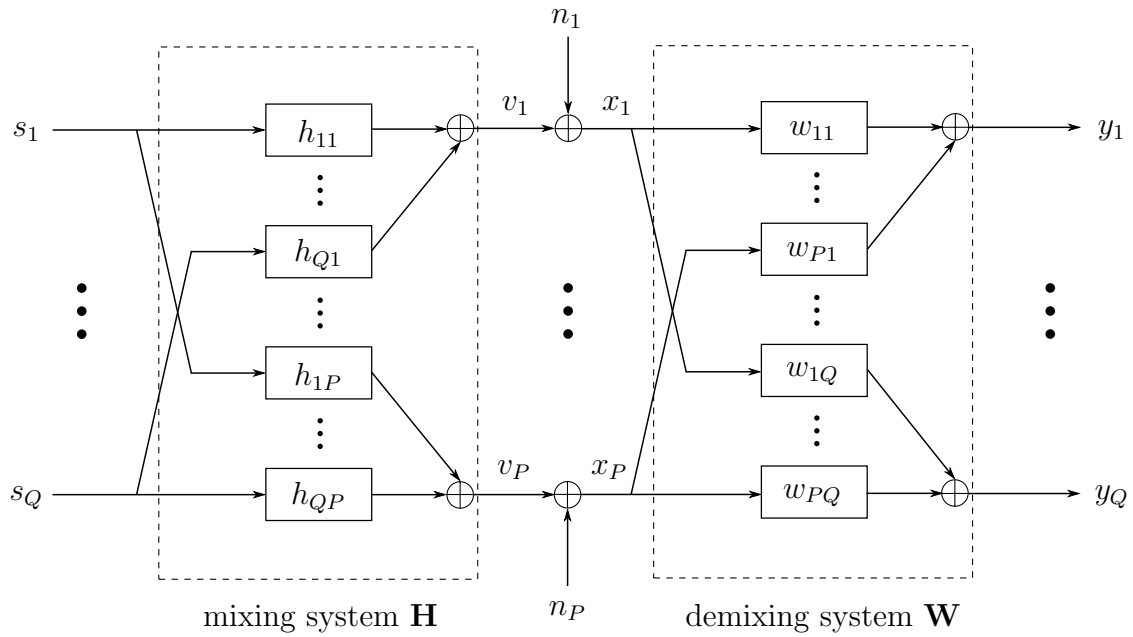


Figure 2.1: BSS signal model [7].

for BSS is statistical independence of the source signals. As already mentioned in the introduction, reverberation significantly reduces the performance of the BSS algorithm and, as a result, the performance of the following ASR. Before explaining the impact of reverberation on BSS, the general concept of the BSS algorithm is outlined in this chapter.

Figure 2.1 shows the basic BSS signal model used in this thesis, where s_q denotes the source signals and, after passing the mixing system and adding an optional noise component n_p, x_p denotes the recorded microphone signals. For the sake of simplicity, the number of microphones and sources is limited to $P = Q = 2$ hereafter. Note, that a basic assumption in this model is that each signal is filtered with an individual room impulse response and therefore the microphone signals x_p describe a linear convolutive mixture of the source signals such that

$$x_1 = s_1 * h_{11} + s_2 * h_{21} + n_1, \tag{2.1}$$

$$x_2 = s_2 * h_{22} + s_1 * h_{12} + n_2, \tag{2.2}$$

where h_{qp} denotes the impulse response from source q to microphone p. The aim of the BSS system is to find a demixing system in order to reverse the mixing process and to recover the initial source signals. This is done by filtering the microphone signals x_p with the coefficients of the demixing filters w_{pq}, analogously to the mixing process, as

$$y_1 = x_1 * w_{11} + x_2 * w_{21}, \tag{2.3}$$

$$y_2 = x_2 * w_{22} + x_1 * w_{12}. \tag{2.4}$$

The BSS algorithm used in this thesis is based on the "**TRI**ple-**N**-**I**ndependent component analysis for **CON**volutive mixtures" (**TRINICON**) framework, which simultaneously accounts for the three fundamental properties *nonwhiteness*, *nonstationarity* and *nongaussianity* [6, 8, 17].

Recall, that the fundamental assumption for the algorithm is statistical independence of the source signals. By minimizing the mutual information between the output signals separation of the independent sources can be achieved.

In this chapter, a matrix notation is used for the demixing filter $\mathbf{W}$ and the signals $(\mathbf{S}, \mathbf{X}, \mathbf{Y})$ which are described in [8] in detail. Throughout this thesis, TRINICON-BSS based on second-order statistics (SOS) is considered, which corresponds to a minimization of mutual information for the case of Gaussian distributions. Referring to the three properties the general TRINICON framework addresses, SOS-based TRINICON-BSS only exploits *nonwhiteness* and *nonstationarity*.

The cost function for SOS-based TRINICON-BSS, given as [6]

$$\mathcal{J}_{\mathrm{SOS}}(m) = \sum_{i=0}^{\infty} \beta(i,m)\{\log \det \hat{\mathbf{R}}_{\mathbf{ss}}(i) - \log \det \hat{\mathbf{R}}_{\mathbf{yy}}(i)\}, \tag{2.5}$$

is applied block-wise to blocks (index m) composed of several smaller blocks (index i) [21]. $\beta(i,m)$ defines a window function for online, block-online or offline realizations. The terms $\hat{\mathbf{R}}_{\mathbf{ss}}(i)$ and $\hat{\mathbf{R}}_{\mathbf{yy}}(i)$ describe the correlation matrices of the sources $\mathbf{s}_q$ and the output signals $\mathbf{y}_q$ as $\hat{\mathbf{R}}_{\mathbf{ss}}(i) = \mathbf{S}^H(i)\mathbf{S}(i)$ and $\hat{\mathbf{R}}_{\mathbf{yy}}(i) = \mathbf{Y}^H(i)\mathbf{Y}(i)$, respectively [1]. Hereby, superscript ^{H} denotes the conjugate transpose of a matrix.

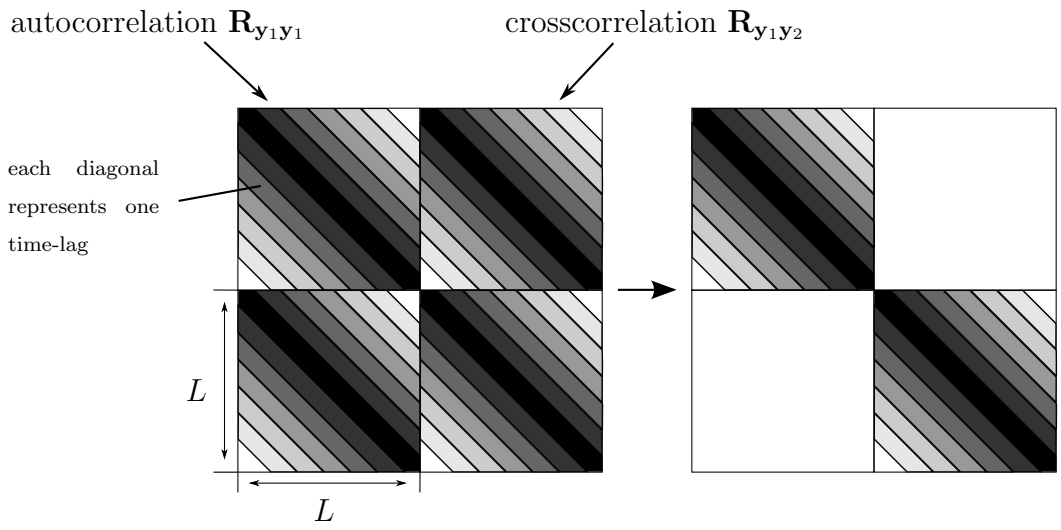


Figure 2.2: Desired score function for SOS [6].

The *nonwhiteness* property is exploited by minimizing the correlation between the two output channels for L output time lags. Simultaneously, the minimization of the correlation for several small blocks i between the two output channels exploits the *nonstationarity* of the signals [21].

The desired score function for SOS-based TRINICON-BSS can intuitively be expressed as

$$\hat{\mathbf{R}}_{\mathbf{ss}}(i) = \mathrm{bdiag}_L \hat{\mathbf{R}}_{\mathbf{yy}}(i) \tag{2.6}$$

and Figure 2.2 illustrates the ideal cancellation of all crosscorrelations taking into account L time-lags, while the autocorrelations stay untouched. Therefore, the structure of the individual signals is preserved [6]. The bdiag operation on a partitioned block matrix consisting of several submatrices sets all submatrices on the off-diagonals to zero [9] which, in this case, is equivalent to the initial assumption of statistically independent signals. The cost function reaches its minimum when the blocks on the off-diagonals vanish.

The update rule for SOS-BSS can be derived by taking the natural gradient of $\mathcal{J}_{\mathrm{SOS}}(m)$ with respect to the demixing filter matrix $\mathbf{W}(m)$

$$\Delta\mathbf{W} \propto \mathbf{W}\mathbf{W}^H \frac{\partial \mathcal{J}}{\partial \mathbf{W}^H}, \tag{2.7}$$

which yields the following generic coefficient update [6]:

$$\Delta\mathbf{W}(m) = 2\sum_{i=0}^{\infty} \beta(i,m)\mathbf{W}(i)\{\hat{\mathbf{R}}_{\mathbf{yy}} - \hat{\mathbf{R}}_{\mathbf{ss}}\}\hat{\mathbf{R}}_{\mathbf{ss}}^{-1}. \tag{2.8}$$

In the next step, the demixing filter matrix $\mathbf{W}$ is updated using the update term (2.8) weighted with the step size μ:

$$\mathbf{W}(m) = \mathbf{W}(m-1) - \mu\Delta\mathbf{W}(m). \tag{2.9}$$

Note, that since BSS employs the method of steepest descent and the gradient yields the direction of steepest ascent, the negative gradient has to be taken for the update in Equation (2.9). The correct choice of the step size is a crucial factor for the convergence speed and stability of the algorithm. Too small step sizes lead to slow convergence while too large step sizes could cause divergence or instability [21].

2.2 Room Reflections and Reverberation

The purpose of this thesis is to present an algorithm that suppresses reverberation and employ it prior to the BSS algorithm presented in Section 2.1. Therefore, as a basis, in this chapter the fundamentals of signal propagation are outlined in order to understand what reverberation is and how it affects BSS performance.

Intuitively speaking, reverberation can be explained as reflections of sound waves from walls. As Figure 2.3 depicts, there are different sound propagation paths from a sound source to the microphone that observes the sound signal. Firstly, there is a

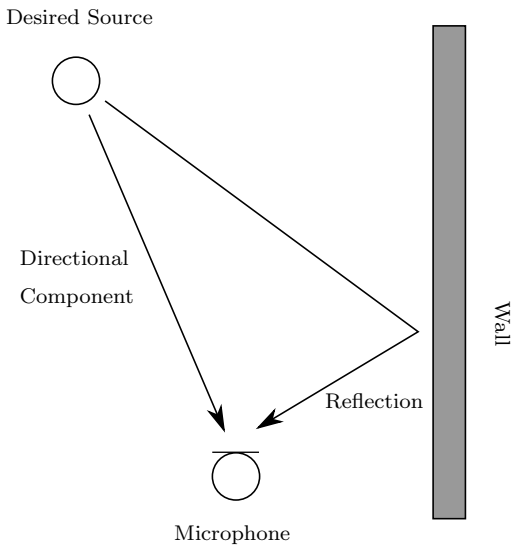


Figure 2.3: Illustration of the direct path and a single reflection from desired source to the microphone [12].

directional component which can be visualized by the shortest path between the sound source and the microphone. Note, that a direct component only exists when a free line of sight between the source and the receiver is given. Secondly, there are indirect components that approach the microphone over reflected paths from different angles and after covering different distances. Consequently, the microphone receives delayed and attenuated copies of the source signal which is called reverberation. In addition, early and late reverberation can be distinguished [12]. Thereby, early reverberation is not perceived as separate sound as long as the propagation time does not exceed $80 - 100$ ms and instead reinforces the direct sound. This phenomenon is termed precedence effect. In contrast, late reverberation describes larger delays and has a significant influence on speech intelligibility. As a summary, Figure 2.4 schematically depicts a measured room impulse response (RIR) divided into the different components.

In order to illustrate the influence of room reflections in the frequency domain, Figure 2.5 shows two spectrograms of a female voice speaking the words "once upon a time there was a little" in different environments. In Figure 2.5a the speech signal is anechoic while Figure 2.5b represents the same sentence in a room with a reverberation

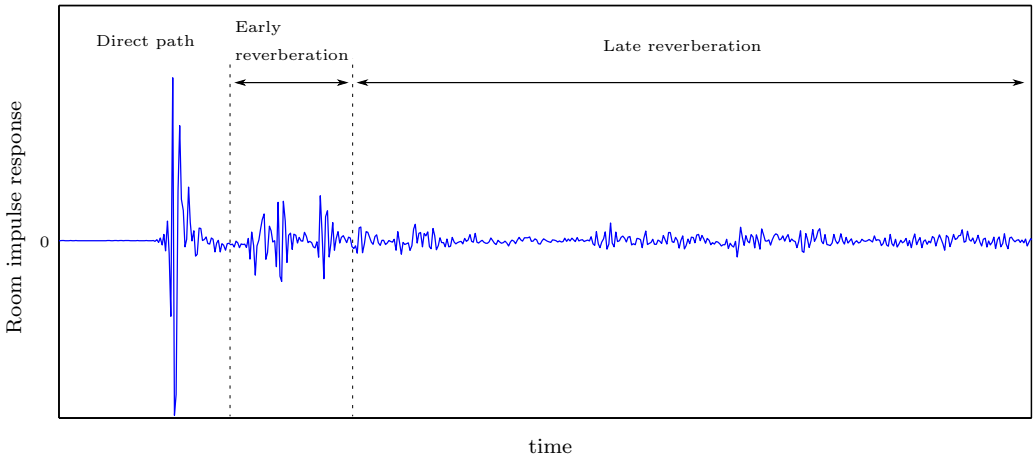
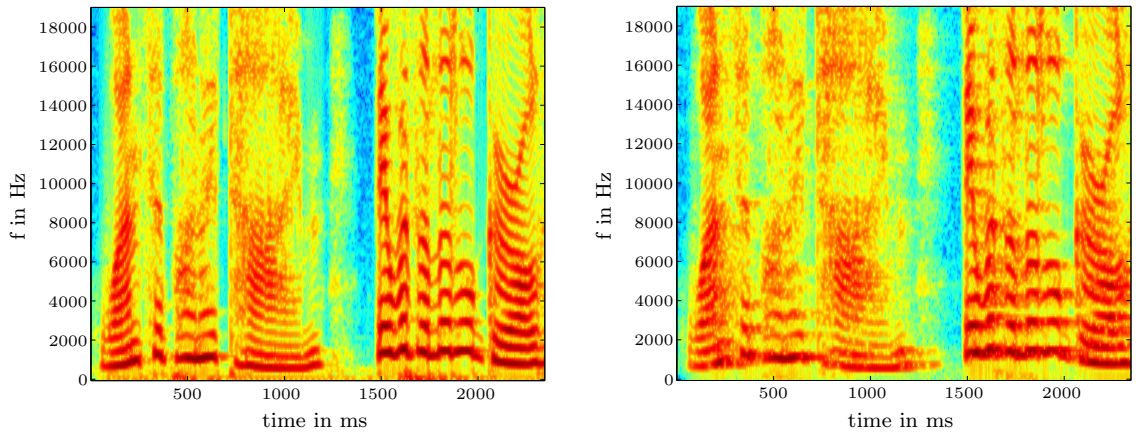


Figure 2.4: Schematic representation of a measured acoustic RIR.

time of $T_{60} = 400$ ms measured at a distance of 1 m. The term T_{60} denotes the time that is necessary for a 60 dB decay of the sound energy after switching of the sound source [12]. As can be seen from the figure, the spectrogram corresponding to the reverberated utterance is blurred and the empty spaces, representing speech pauses, are filled by reverberation causing audible differences in the speech signals.



(a) No reverberation **(b)** 400 ms reverberation

Figure 2.5: Spectrogram of non-reverberant and reverberant speech signal.

This blurring effect also explains the decrease in BSS performance described before, as the BSS system has to identify all different signal propagation paths from overlapping speech portions at a time instance in order to obtain a good separation performance. The number of degrees of freedom of the BSS system is limited due to the limited length of the demixing filter. Increased reverberation leads to many different propagation paths causing a reduction of BSS performance if not all propagation paths can be modeled sufficiently. In addition, the convergence speed of the BSS system decreases since more propagation paths need to be identified.

2.3 Vector Spaces and Subspaces

In order to suppress reverberation and ambient noise, several approaches, like employing a Delay-and-Sum (DS) beamformer [15] or using directional microphones [16], have been suggested. In this thesis, the Subspace Method presented by Asano et al. [3] is applied. It exploits the properties of vector spaces and subspaces which are described in the following.

The concept of vector spaces is illustrated by recalling the most common real vector spaces denoted by $\mathbb{R}^1$, $\mathbb{R}^2$ and $\mathbb{R}^3$ where $\mathbb{R}^1$ describes a line, whereas $\mathbb{R}^2$ describes the x-y plane. $\mathbb{R}^3$ denotes the three-dimensional space and, in general, the space $\mathbb{R}^n$ consists of all column vectors with n components. Within all vector spaces, vectors can be added and multiplied by scalars. In other words, linear combinations which lead to vectors in the same vector space can be taken. All operations have to fulfill the properties displayed in the following enumeration.

Vector Space properties

1. $x + y = y + x$
2. $x + (y + z) = (x + y) + z$
3. There is a unique "zero vector" such that $x + 0 = x$ for all x
4. For each x there is a unique vector $-x$ so that $x + (-x) = 0$
5. $1x = x$
6. $(c_1c_2)x = c_1(c_2x)$
7. $c(x + y) = cx + cy$
8. $(c_1 + c_2)x = c_1x + x_2x$

In summary, a real vector space is a set of vectors together with rules for vector addition and multiplication by real numbers [22]. In real applications, especially where the frequency domain is employed, the vector spaces are often complex. For the sake of simplicity and illustration the properties are presented for real vector spaces but the concept is equally applicable for complex spaces [22].

In Figure 2.6, a two-dimensional plane within a three-dimensional vector space is illustrated. This plane through $(0, 0, 0)$ is a vector space in its own right and it is called a **subspace** of the original space $\mathbb{R}^3$. Note, that the biggest subspace is $\mathbb{R}^3$ itself, while the zero vector forms the smallest, "zero-dimensional", subspace. The formal definition reads as follows: "A subspace of a vector space is a nonempty subset that satisfies the requirements for a vector space: Linear combinations stay in the subspace" [22]. This could also be described as the subspace being closed under addition and scalar multiplication. The properties outlined for the larger space are automatically satisfied in every subspace.

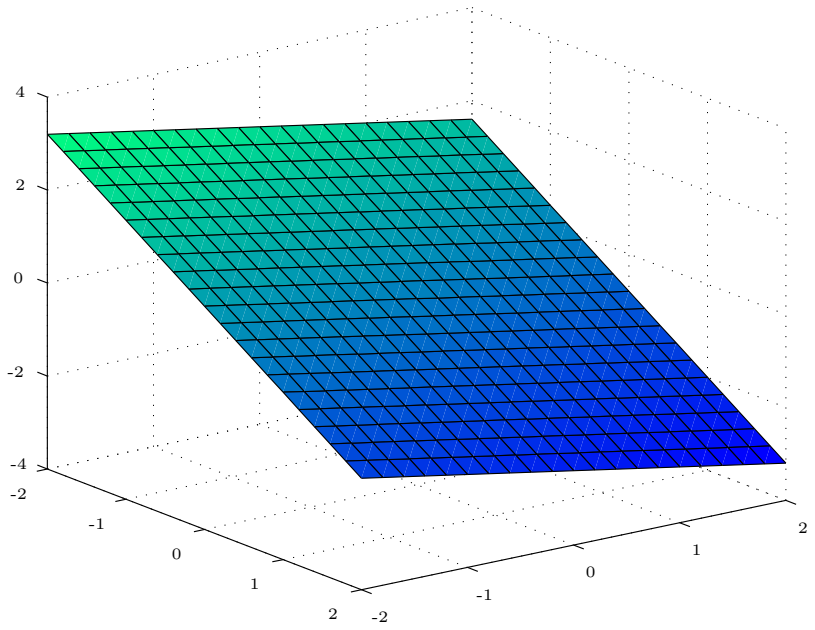


Figure 2.6: Two-dimensional subspace of three-dimensional vector space.

Chapter 3

Subspace Method

As mentioned in Section 2.3, there are several different approaches in audio signal processing to suppress noise and reverberation. This can be done by either suppressing the reverberation or enhancing the direct component. Conventional algorithms require knowledge of the array geometry or the sound field which is not given in the framework of BSS. Therefore, an algorithm that does not require previous knowledge is required in order to be used as a preprocessor for BSS. Prominent examples in this context are linear predictive coding (LPC) analysis or long term cepstral mean subtraction [23]. In the following, a method to reduce room reflections and reverberation, the Subspace Method in the frequency domain as described by Asano et al. [3], is presented. At this juncture, the general idea is to separate room reflections and directional components in the eigenvalue domain based on the spatial extent of the acoustic signal. By applying the subspace filter the ambient components are suppressed while the directional components are preserved.

3.1 Underlying Signal Model

Before proceeding to examine the Subspace Method, it will be necessary to introduce the signal model for the input signal as in [3]. Considering the observation of the sound field inside a room with M microphones and D sound sources, the input vector, representing the short-term Fourier transform (STFT) of the microphone inputs, yields

$$\mathbf{x}(\omega, t) = [X_1(\omega, t), ..., X_M(\omega, t)]^T \tag{3.1}$$

where t and ω denote the time and the frequency index, respectively, and $X_m(\omega, t)$ denotes the STFT of the m-th microphone signal. The symbol ^{T} refers to the transpose of a vector. As introduced in Section 2.2, in a noisy and reverberant environment there are directional components as well as less directional components such as room reflections and ambient noise. Correspondingly, in the following, the input signal is modeled as

$$\mathbf{x}(\omega, t) = \mathbf{A}(\omega)\mathbf{s}(\omega, t) + \mathbf{n}(\omega, t). \tag{3.2}$$

Similar to the input vector described in Equation (3.1), the vector $\mathbf{s}(\omega, t)$ consists of the source spectra $\mathbf{s}(\omega, t) = [S_1(\omega, t), ..., S_D(\omega, t)]^T$ where $S_n(\omega, t)$ describes the spectrum of the n-th source. Matrix $\mathbf{A}(\omega)$ is termed mixing matrix and its (m, n)-th element defines the transfer function from the n-th source to the m-th microphone as

$$A_{m,n}(\omega, t) = H_{m,n}(\omega)e^{-j\omega\tau_{m,n}}, \tag{3.3}$$

where $H_{m,n}(\omega)$ denotes the magnitude of the transfer function. The propagation time from the n-th source to the m-th microphone is denoted by $\tau_{m,n}$. The less directional components and ambient noise are combined in $\mathbf{n}(\omega, t)$ while the term $\mathbf{A}(\omega)\mathbf{s}(\omega, t)$ represents the directional component. In the following, the directional component will be called signal component while $\mathbf{n}(\omega, t)$ will be called ambient component.

3.2 Mathematical Approach

The fundamental idea of the Subspace Method is to reduce the ambient component, containing ambient noise as well as room reflections, and to preserve the direct component of a signal. This is done by employing the properties of vector spaces and subspaces as will be described in this chapter.

3.2.1 Spatial Correlation Matrix

The spatial correlation matrix of the microphone signals is defined as

$$\mathbf{R}(\omega) = \mathcal{E}\{\mathbf{x}(\omega, t)\mathbf{x}^H(\omega, t)\}. \tag{3.4}$$

$\mathcal{E}\{\cdot\}$ is generally understood to be the expectation value. Note, that the Subspace Method is conducted at each frequency independently. Therefore, the Fast Fourier Transform (FFT) length determines the number of distinct spatial correlation matrices for each block. To accommodate the nonstationarity of speech and audio signals, a recursive mean as in [25] is applied so that

$$\mathbf{R}(\omega) = \lambda\mathbf{R}(\omega - 1) + (1 - \lambda)\mathbf{R}(\omega), \tag{3.5}$$

where λ denotes the forgetting factor which is usually chosen close to one, to ensure that the influence of past values decreases over time. In order to simplify notation, the frequency index ω is omitted in the remainder of this chapter.

In the following, it is assumed that the signal and the ambient component are uncorrelated. This assumption holds to some extent in a practical sense when the window length of the STFT is short and the interval between the direct sound and the reflection exceeds this window length. Replacing $\mathbf{x}(\omega, t)$ in Equation (3.4) with its definition in Equation (3.2) and dismissing the crosscorrelation spectra containing

both signal and ambient components R can be written as

$$\mathbf{R} = \mathbf{A}\mathbf{Q}\mathbf{A}^H + \mathbf{K}, \tag{3.6}$$

with $\mathbf{Q} = \mathcal{E}\{\mathbf{s}(t)\mathbf{s}^H(t)\}$ being the cross-spectrum matrix of the sources and $\mathbf{K} = \mathcal{E}\{\mathbf{n}(t)\mathbf{n}^H(t)\}$ being the correlation matrix of the ambient components.

3.2.2 Properties of the Subspace Method

Applying the generalized eigenvalue decomposition of $\mathbf{R}$ as in [20] yields to

$$\mathbf{R} = \mathbf{K}\mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1} \tag{3.7}$$

with $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_M]$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_M)$ where $\mathbf{e}_m$ and λ_m are the eigenvectors and the eigenvalues, respectively. As $\mathbf{K}$ contains reflections and reverberation it cannot be observed separately and therefore in this thesis $\mathbf{K} = \mathbf{I}$ is assumed. This leads to the standard eigenvalue decomposition

$$\mathbf{R} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1}, \tag{3.8}$$

which is equivalent to the assumption that $\mathbf{n}(t)$ is spatially white.

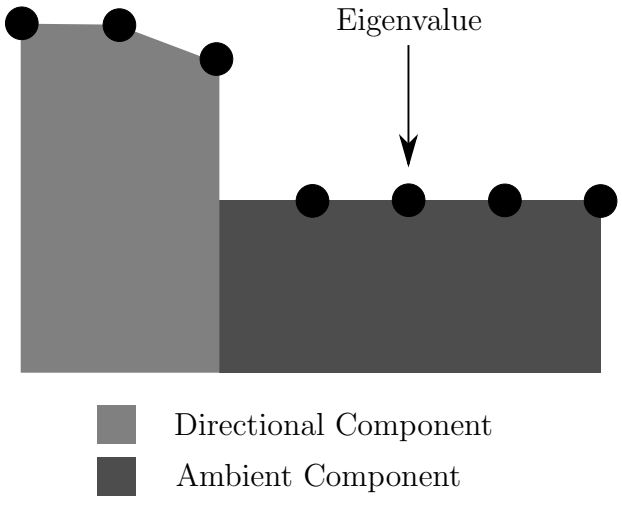


Figure 3.1: Typical eigenvalue distribution for $D = 3$ and $M = 7$ [3].

The above mentioned characteristics, as well as the structure of the spatial correlation matrix $\mathbf{R}$, entail four important properties of the Subspace Method [3, 10].

Property 1 The energy of D directional signals $\mathbf{s}(t)$ is concentrated on the D dominant eigenvalues.

Property 2 The energy of $\mathbf{n}(t)$ is equally spread over all eigenvalues.

Property 3 $\mathfrak{R}(\mathbf{A}) = \mathfrak{R}(\mathbf{E_s})$ where $\mathbf{E}_s = [\mathbf{e}_s, ..., \mathbf{e}_D]$ denotes the eigenvectors corresponding to the D dominant eigenvalues.

Property 4 $\mathfrak{R}(\mathbf{A}) = \mathfrak{R}(\mathbf{E_n})^{\perp}$ where $\mathbf{E}_n = [\mathbf{e}_{D+1}, ..., \mathbf{e}_M]$ denotes the eigenvectors corresponding to the remaining $M - D$ eigenvalues.

Properties 1 and 2 are illustrated in Figure 3.1 that shows an idealized eigenvalue distribution for three sound sources and seven microphones for one frequency bin. An example for a real eigenvalue distribution for two speech sources and six microphones in a room with short reverberation time ($T_{60} = 50$ ms) is depicted in Figure 3.2 where the concentration of the energy on the two dominant eigenvalues, visualized in dark red, can clearly be observed.

Properties 3 and 4 result from the fact that the spatial correlation matrix is an Hermitian matrix and therefore its eigenvectors form an orthogonal basis [10]. $\mathfrak{R}(A)$ describes the space spanned by the column vectors of the mixing matrix $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_D]$ which could also be written as $\text{span}(\mathbf{a}_1, ..., \mathbf{a}_D)$. The notation $\mathfrak{R}(\mathbf{E}_n)^{\perp}$ describes the orthogonal complement of $\mathfrak{R}(\mathbf{E}_n)$. The subspaces $\mathfrak{R}(\mathbf{E}_s)$ and $\mathfrak{R}(\mathbf{E}_n)$ are termed signal subspace and noise subspace, respectively. According to the properties mentioned above, signal and noise subspace are orthogonal to each other.

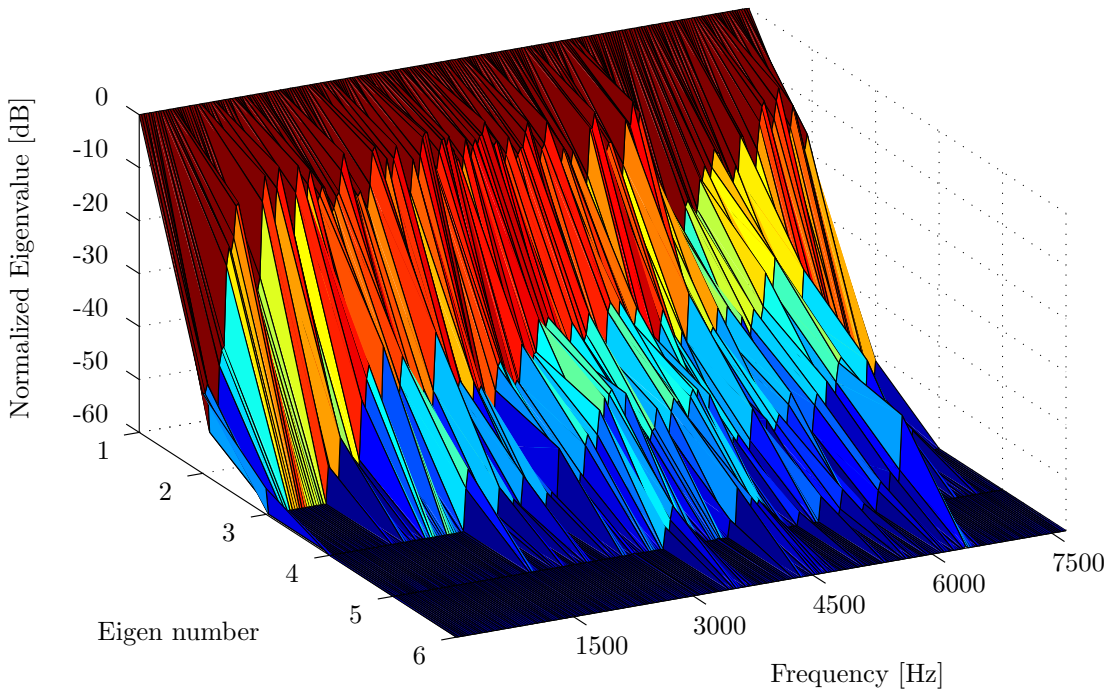


Figure 3.2: Eigenvalue distribution of a speech signal for $D = 2$, $M = 6$ and $f_s = 16$ kHz.

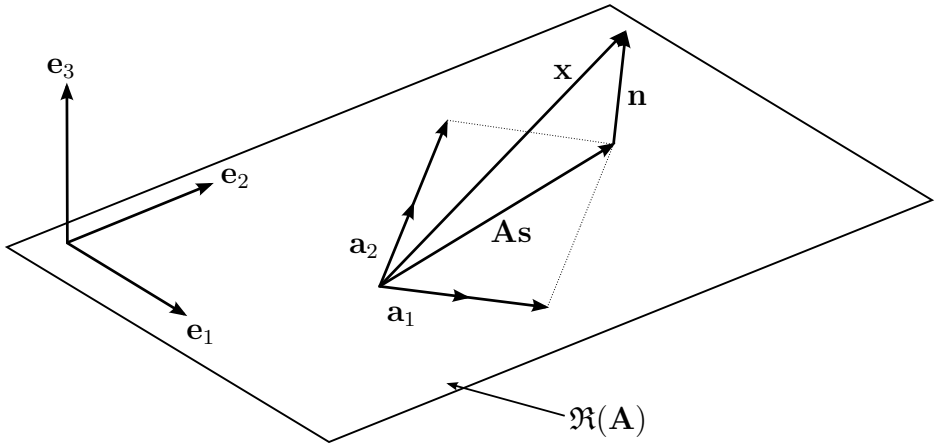


Figure 3.3: Relation of Vectors for $D = 2$ and $M = 3$ [3].

As displayed in Figure 3.3, the $D = 2$ eigenvectors $\mathbf{e}_1$ and $\mathbf{e}_2$ corresponding to the D dominant eigenvalues span the signal subspace $\mathfrak{R}(\mathbf{A}) = \mathfrak{R}(\mathbf{E}_s)$. Recall, that the direct component is represented by $\mathbf{As}$ and note that it is located within the signal subspace, while the ambient component $\mathbf{n}$ is located within the noise subspace that fills the remaining one-dimensional subspace. This is an additional visualization of Properties 1, 3 and 4 and shows, why the energy within the signal subspace distributes over the eigenvalues corresponding to $\mathbf{e}_1$ and $\mathbf{e}_2$. As the eigenvectors of $\mathbf{R}$ form an orthogonal basis, the number of microphones determines the order of the employed vector space while the number of sound sources determines the order of the signal subspace. Accordingly, the Subspace Method requires the number of microphones M to be considerably larger than the number of sources D. In this thesis, the dimension of the signal subspace $D = 2$ is assumed to be known. In real applications however, this dimension has to be estimated as, for instance, illustrated in [25].

3.2.3 Subspace Filter

Conceptually, the subspace filter projects the input signal $\mathbf{x}(t)$ onto the eigenvectors spanning the signal subspace in order to suppress the ambient component. This filter is defined as [3]

$$\mathbf{W} = \mathbf{\Lambda}_s^{-1/2}\mathbf{E}_s^H, \tag{3.9}$$

where $\mathbf{\Lambda}_s = \text{diag}(\lambda_1, ..., \lambda_D)$. The term $\mathbf{\Lambda}_s^{-1/2}$ is the same normalization factor as the one used in Principal Component Analysis (PCA) [18]. The conjugate transpose of the eigenvectors corresponding to the D dominant eigenvalues $\mathbf{E}_s^H$ plays a major role in the subspace filter to reduce the energy of $\mathbf{n}(t)$. As a proof of concept, recall Properties 1 and 3 which imply that the directional component $\mathbf{As}(t)$ can be expanded with the

subset of the basis vectors, $\{\mathbf{e}_1, ..., \mathbf{e}_D\}$, as

$$\mathbf{As}(t) = \sum_{i=1}^{D} \alpha_i(t)\mathbf{e}_i, \tag{3.10}$$

where $\alpha_i(t)$ is the projection coefficient of $\mathbf{As}(t)$ onto the basis vector $\mathbf{e}_i$. Similarly, according to Property 2, $\mathbf{n}(t)$ is expanded using all basis vectors $\{\mathbf{e}_1, ..., \mathbf{e}_M\}$ as

$$\mathbf{n}(t) = \sum_{i=1}^{M} \beta_i(t)\mathbf{e}_i, \tag{3.11}$$

where $\beta_i(t)$ is the projection coefficient of $\mathbf{n}(t)$ onto the basis vector $\mathbf{e}_i$. Written in a matrix vector notation this results in

$$\mathbf{As}(t) = \mathbf{E}_s\boldsymbol{\alpha}(t) \tag{3.12}$$

$$\mathbf{n}(t) = \mathbf{E}\boldsymbol{\beta}(t), \tag{3.13}$$

where $\boldsymbol{\alpha}(t) = [\alpha_1(t), ..., \alpha_D(t)]^T$ and $\boldsymbol{\beta}(t) = [\beta_1(t), ..., \beta_M(t)]^T$. As the noise energy equally distributes over all eigenvalues, there are ambient components in the signal subspace as well as in the noise subspace. Thus, the ambient component can be split as

$$\mathbf{n}(t) = \mathbf{n}_s(t) + \mathbf{n}_n(t), \tag{3.14}$$

where $\mathbf{n}_s(t) \in \mathfrak{R}(E_s)$ and $\mathbf{n}_n(t) \in \mathfrak{R}(E_n)$ and

$$\mathbf{n}_s(t) = \mathbf{E}_s(t)\boldsymbol{\beta}_s(t) \tag{3.15}$$

$$\mathbf{n}_n(t) = \mathbf{E}_n(t)\boldsymbol{\beta}_n(t) \tag{3.16}$$

with $\boldsymbol{\beta}_s(t) = [\beta_1(t), ..., \beta_D(t)]^T$ and $\boldsymbol{\beta}_n(t) = [\beta_{D+1}(t), ..., \beta_M(t)]^T$. Applying the subspace filter to the components mentioned in Equation (3.12), (3.15) and (3.16) and

using the properties of the eigenvectors, $\mathbf{E}_s^H(t)\mathbf{E}_s(t) = \mathbf{I}$ and $\mathbf{E}_s^H(t)\mathbf{E}_n(t) = 0$, leads to

$$\mathbf{WAs}(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_s\boldsymbol{\alpha}(t)) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\alpha}(t) \tag{3.17}$$

$$\mathbf{Wn}_s(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_s\boldsymbol{\beta}_s(t)) = \mathbf{\Lambda}^{-1/2}\boldsymbol{\beta}_s(t) \tag{3.18}$$

$$\mathbf{Wn}_n(t) = \mathbf{\Lambda}^{-1/2}\mathbf{E}_s^H(\mathbf{E}_n\boldsymbol{\beta}_n(t)) = 0. \tag{3.19}$$

Equation (3.19) indicates that by applying the Subspace Method the ambient component within the noise subspace is canceled. In contrast, the components in the signal subspace $\mathbf{As}(t)$ and $\mathbf{n}_s(t)$ are preserved. Consequently, when the number of microphones considerably exceeds the number of sound sources, the expectation is that a large amount of $\mathbf{n}(t)$ can be canceled by the Subspace Method.

Before applying each subspace filter, it is transformed into the time-domain as [3]

$$w_{n,m}(t) = \mathrm{IDFT}\{W_{n,m}(\omega)\}, \tag{3.20}$$

where $\mathrm{IDFT}\{\cdot\}$ denotes the inverse Discrete Fourier Transform (DFT). The (n, m)-th element of the frequency domain filter $\mathbf{W}(\omega)$ and its corresponding element in the time domain are denoted by $W_{n,m}(\omega)$ and $w_{n,m}(t)$, respectively. Finally, the filters are convolved with the corresponding time-domain microphone signals

$$y_d(t) = \sum_{i=1}^{M} w_{d,i}(t) * x_i(t) \tag{3.21}$$

in order to get the D output signals of the subspace filter which are summarized in the vector $\mathbf{y}(t) = [y_1(t), ..., y_D(t)]^T$.

The Subspace Method can be considered a self-organizing beamformer [3]. This can be explained by looking at a DS beamformer in the frequency domain focusing on the n-th target source which can be expressed as [19]

$$\mathbf{y}_{DS}(t) = \mathbf{w}_{DS}\mathbf{x}(t), \tag{3.22}$$

with the transfer function

$$\mathbf{w}_{DS}(t) = \frac{1}{M}[e^{j\omega\tau_{1,n}}, ..., e^{j\omega\tau_{M,n}}]. \tag{3.23}$$

It is assumed that $H_{1,n} = ... = H_{M,n} = 1$ in Equation (3.3). The vector representation of $\mathbf{w}_{DS}(t)$ yields [3]

$$\mathbf{w}_{DS}(t) = \frac{\mathbf{a}_n^H}{\mathbf{a}_n^H \mathbf{a}_n} \tag{3.24}$$

with $\mathbf{a}_n$ being the n-th column vector of $\mathbf{A}$ and $\mathbf{a}_n^H \mathbf{a}_n$ representing a normalization factor. The formula can be extended to a DS beamformer focusing on n target sources by deploying a matrix notation resulting in

$$\mathbf{W}_{DS}(t) = \frac{\mathbf{A}^H}{\mathbf{A}^H \mathbf{A}}. \tag{3.25}$$

Ultimately, applying this filter matrix to the noise component $\mathbf{n}_n(t)$ in Equation (3.16) as

$$\mathbf{W}_{DS}\mathbf{n}_n(t) = \frac{\mathbf{A}^H}{\mathbf{A}^H \mathbf{A}}\mathbf{E}_n \boldsymbol{\beta}_n(t) = 0 \tag{3.26}$$

leads to a total cancellation of the noise component because of Property 4 stating $\mathbf{A}^H \mathbf{E}_n = 0$. Consequently, the DS beamformer and the Subspace Method have the same noise reduction mechanism to cancel the noise component within the noise subspace. Thus, the subspace filter can be considered as a self-organizing beamformer [3].

Chapter 4

Objective Evaluation

In this chapter, the Subspace Method is evaluated using objective performance measures that are, along with other information about the experimental setup, described first. Subsequently, two scenarios are investigated. First, the basic functionality of the Subspace Method for a single speaker is verified for different room conditions. Second, the Subspace Method is employed as a preprocessing stage for a BSS algorithm. The underlying exact numbers for the various charts used to discuss the results of the experiments can be found in Appendix A.

4.1 Experimental Setup

Before presenting the different evaluation results, this section gives a brief overview of the experimental setup including the room environments, in which the experiments are conducted, as well as the microphone setup. Moreover, the employed ASR and the BSS framework are described and the underlying parameters for the Subspace Method are given. Eventually, the measures used as a performance criteria in the evaluation are explained.

4.1.1 Experimental Details

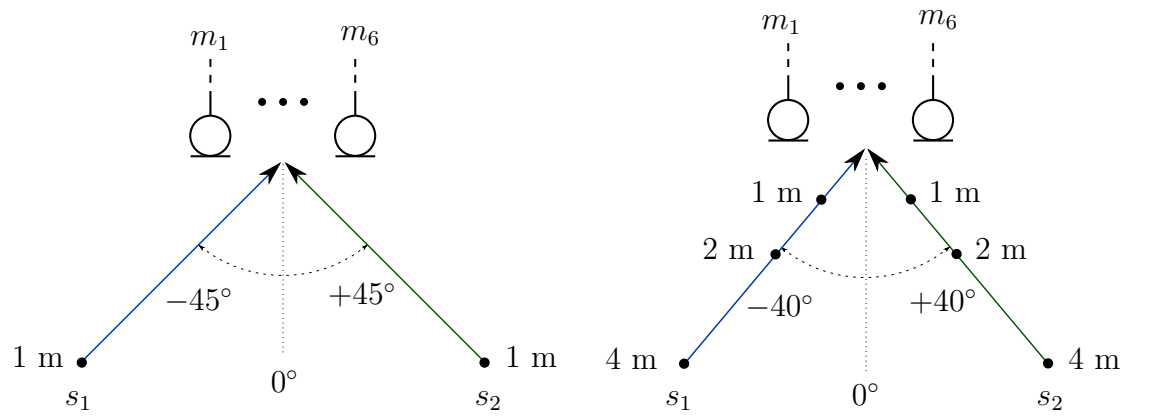
All experiments are conducted in the same way, with identical source signals in different reverberant environments. Table 4.1 shows the reverberation times T_{60} for the different room constellations termed A, B, C and D, where the subscript of the rooms displayed in the table refers to the distance of the speaker to the microphone array in m.

Table 4.1: Reverberation times T_{60} and distances from the speakers to the mircophone for four different rooms.

Room	reverberation time T_{60}	distances
A	50 ms	1 m
B_1 — B_2 — B_4	250 ms	1 m — 2 m — 4 m
C_1 — C_2 — C_4	400 ms	1 m — 2 m — 4 m
D_1 — D_2 — D_4	900 ms	1 m — 2 m — 4 m

The following setup describes the situation for two simultaneously active speakers, but for experiments with only one active source, as in Section 4.2, the same constellations are valid taking into account only one speaker. The angles of the speakers are chosen to be $+45°$ and $-45°$ in room A while, due to the given microphone setup, for room B, C and D the angles are chosen to be $+40°$ and $-40°$, as displayed in Figure 4.1a and Figure 4.1b.

The microphone array consists of $M = 6$ aligned microphones with an inter-element spacing of 0.042 m, shown in Figure 4.2. The microphone signals are obtained by convolving the source signals with M individually measured RIRs for each speaker position.



(a) Speaker positions for room A. **(b)** Speaker positions for rooms B, C and D.

Figure 4.1: Speaker positions for different rooms with different reverberation times.

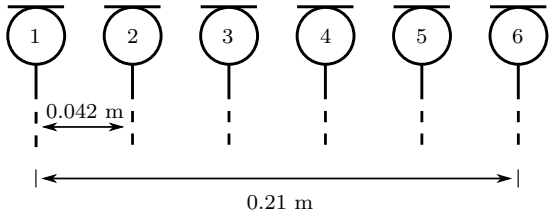


Figure 4.2: Mircophone constellation [4].

The Subspace Method used in this thesis, introduced in Chapter 3, is employed in the frequency domain. Throughout the entire evaluation, a block-offline structure is used. The method is performed block by block and each frame is weighted with a von Hann window. For each frequency bin, the spatial correlation matrix is calculated. As described in Equation (3.5), by applying the recursive mean, the spatial correlation matrices are then updated for every block. Consequently, the complete signal is used for the estimate. Afterwards, the eigenvalues and eigenvectors of the resulting spatial correlation matrices are calculated in order to receive the filter matrices, see Equation (3.9). Ultimately, the filter matrices are transformed into the time domain in order to avoid time-domain aliasing [3], resulting in M time-domain filters. Each microphone signal is then filtered with the corresponding time-domain filter and by adding all filtered signals the output signal, as in Equation (3.21), is obtained. A complete list of parameters used for the Subspace Method can be found in Table 4.2. All algorithms used in this thesis are implemented and evaluated using MATLAB.

Table 4.2: Parameters used for the block-offline realization of the Subspace Method.

Parameters for the Subspace Method	
Sampling frequency	16 kHz
Shift of STFT	16
Window length	512
Window	von Hann
Forgetting factor, λ	0.999
Number of microphones, M	6
Number of sources, D	1-2

In Section 4.3, an offline version of the BSS algorithm characterized in Section 2.1 is used to achieve signal separation. Thereby, the length of the demixing filter is chosen to be 1024 and a stepsize of $\mu = 0.00001$ is applied, see Equation (2.9). Furthermore, the number of iterations is set to be 500. The output signals of the BSS algorithm are then provided as input arguments for the ASR that is described in the following section.

4.1.2 Performace Measures

Automatic speech recognition rate

The central performance measure for the evaluation is the automatic speech recognition rate (ASR rate) of an ASR which evaluates the overall signal quality. The ASR engine PocketSphinx [24] is used with an acoustic model trained on clean speech from the GRID corpus [11]. The source signals consist of utterances of the form "command - color - preposition - letter - number - adverb", but, as in the PASCAL CHiME challenge [5], for the computation of the ASR rate only the letter and the number are taken into account. For the evaluation of the Subspace Method for a single speaker, as in

Section 4.2, a signal consisting of 100 utterances, corresponding to an input signal of approximately 3 minutes, is used. In order to evaluate the Subspace Method as a preprocessor for BSS, which requires a mixture of two signals, another 100 utterances are taken into account. Thereby, the two source signals are of equal energy. All recordings are sampled at a rate of 16 kHz. Any additive noise is white Gaussian noise and used to model microphone self-noise. The ASR requires the endpoints of each utterance and therefore accurate timing is required. Accordingly, in order to accommodate delays from RIRs or filters, the crosscorrelation of the output signal and the original source signal is determined and the corresponding shift is considered to obtain precise ASR rates. In order to increase the reliability of the measures for unfiltered signals, the ASR rates for the M microphone input signals are individually determined and averaged.

Frequency-weighted segmental signal-to-noise ratio

The second performance measure, taken to determine the impact of the Subspace Method in terms of noise suppression, is the frequency-weighted segmental signal-to-noise ratio (fwsegSNR) which is calculated as [13]

$$\text{fwsegSNR} = \frac{10}{M} \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) \log_{10} \frac{|X(j,m)|^2}{(|X(j,m)|-|\hat{X}(j,m)|)^2}}{\sum_{j=1}^{K} W(j,m)}, \tag{4.1}$$

where K denotes the number of bands and $W(j,m)$ is the weight placed on the j-th frequency band as in [13]. M denotes the total number of frames in the signal and $|X(j,m)|$ is the signal spectrum in the j-th frequency at the m-th frame weighted by a Gaussian-shaped window. $|\hat{X}(j,m)|$ refers to the weighted reference signal spectrum in the same band. The fwsegSNR is a better match to auditory perception as perceivable frequencies are weighted stronger. Note, that the noise component also contains reverberation. Consequently, when no additional noise is applied, this measure could also be described as a frequency-weighted segmental signal-to-reverberation ratio. To con-

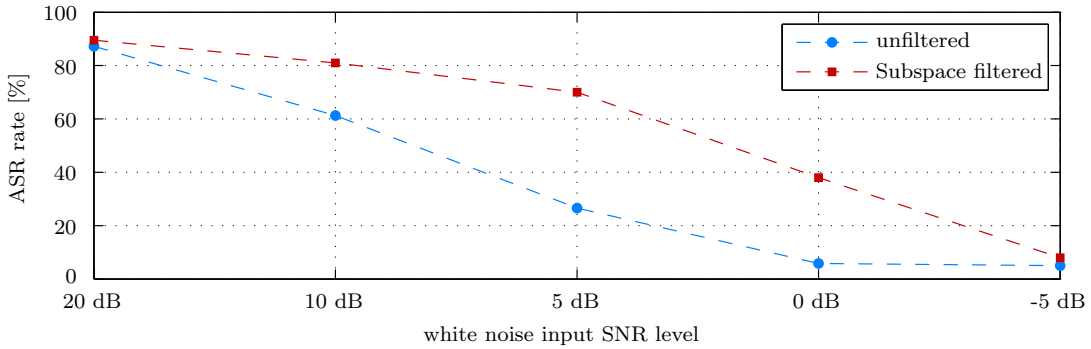
sider reverberation as a part of the noise component, the signal $\hat{X}$ in Equation (4.1) only consists of the direct component, whereas the signal X that is analyzed additionally contains room reflections and noise. This direct component is obtained by extracting the direct path from each individual RIR, as displayed in Figure 2.4, and convolving it with the source signal. In order to obtain the reference signal to measure the fwsegSNR after filtering through the Subspace Method, the direct component is filtered with the same subspace filters as the microphone signals. High values are desired for the fwsegSNR as the difference $|X(j,m)| - |\hat{X}(j,m)|$ in the denominator in Equation (4.1) describes the noise component which is aimed to be small.

4.2 Evaluation of the Subspace Method

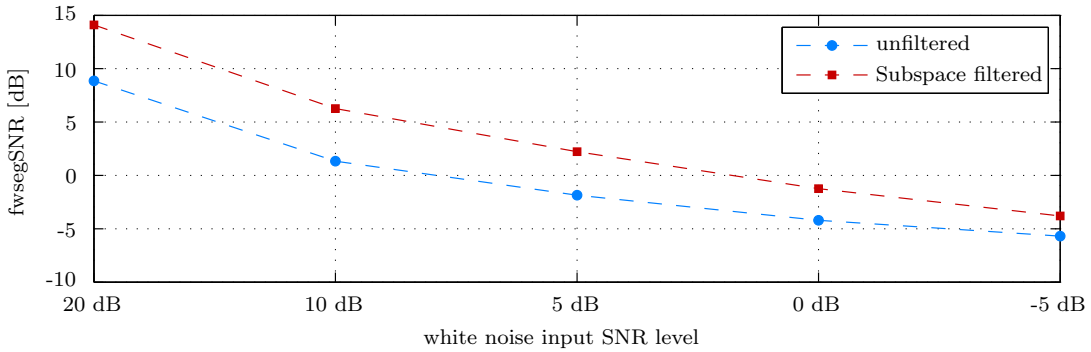
The first set of analyzes verifies the functionality of the Subspace Method for one speaker regarding the performance measures presented in Section 4.1.2, before proceeding to examine the Subspace Method as a preprocessor for BSS. Therefore, three different scenarios are analyzed. In the first scenario, different white Gaussian noise levels in the least reverberant environment, room A, are investigated. In contrast, the second scenario deals with sole reverberation in rooms B, C and D for different speaker distances as given in Table 4.1. Ultimately, in the third scenario, the reverberation scenario is extended by adding 20 dB white Gaussian noise. The results obtained from the simulations are shown in Figures 4.3 to 4.5. Exact numbers for the charts are provided in Tables A.1 to A.3.

In the first scenario, white Gaussian noise at different signal-to-noise ratio (SNR) levels is added to the source signal with SNR = {20, 10, 5, 0, -5} dB. Figure 4.3a compares the results obtained from the ASR before and after subspace filtering. As a reference, the ASR rate of the original source signal was measured to be 93 %. With successive decrease of the SNR level, the recognition rates also decrease in both cases.

However, the subspace filtered signal constantly shows higher ASR rates, especially between 10 dB and 0 dB and therefore proves robustness against white Gaussian noise in terms of ASR rate. It is striking, that for the lowest SNR level no significant improvement was achieved. This could result from the fact, that the noise energy is exceeding the energy of the direct components, causing the Subspace Method to choose wrong eigenvectors that are located in the noise subspace. Figure 4.3b shows a similar development for the fwsegSNR, where the subspace filtered signal outperforms the unfiltered signal resulting in gains between 1.9 and 5.24 dB. It can be concluded that the subspace filtered signal shows superior performance and robustness against white Gaussian noise compared to the unfiltered signal. As mentioned in Chapter 3, one assumption of the Subspace Method is, that the noise component is white, which is basically fulfilled in this setup.



(a) Comparison of the ASR rates.

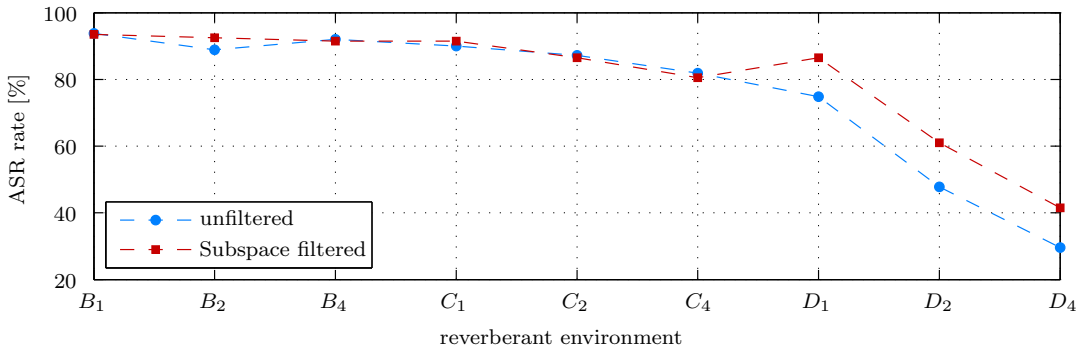


(b) Comparison of the fwsegSNR values.

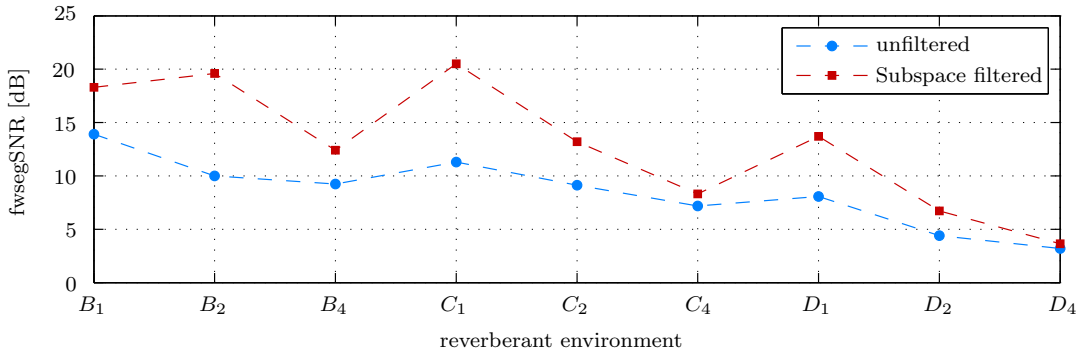
Figure 4.3: Comparison of the ASR rates and fwsegSNR values for one speaker in room *A* for different additive white Gaussian noise levels before and after filtering through the Subspace Method.

In the second scenario, depicted in Figure 4.4, no additional noise is added and therefore the noise component only consists of reverberation. It can be seen from Figure 4.4a that there is no significant difference between the unfiltered and the subspace filtered signal for different distances in room B and C. For room D however, ASR rates significantly drop with increasing distances and constant gains around 12 percentage points are achieved through filtering. The results indicate, that the ASR system is not very susceptible to reverberation times T_{60} under 500 ms as the results are still close to those of the clean source signal. This might result from the very limited amount of words the ASR can distinct and better filtering results are expected for ASRs with a larger database. In addition, the initial assumption of the ambient component being completely uncorrelated to the direct component is not entirely fulfilled for reverberation, especially for shorter reverberation times, causing worse results than for white Gaussian noise. It is apparent, that for the fwsegSNR, displayed in Figure 4.4b, the subspace filtered signal shows a significantly higher performance than the unfiltered signal which underlines the robustness of the speech recognizer against reverberation. In rooms C and D, for larger distances within the same room environment, the fwsegSNR gains drop since for higher distances the estimation of the signal subspace might be less accurate.

Finally, Figure 4.5 shows the results for adding 20 dB white Gaussian noise to the second scenario displayed in Figure 4.4. While the subspace filtered signal only exhibits small decreases in ASR performance compared to the second scenario, the impact on the unfiltered signal is much bigger, resulting in good improvements. The overall tendency of the fwsegSNR shows a similar development as in Figure 4.5b but the line is shifted downwards due to the added noise.



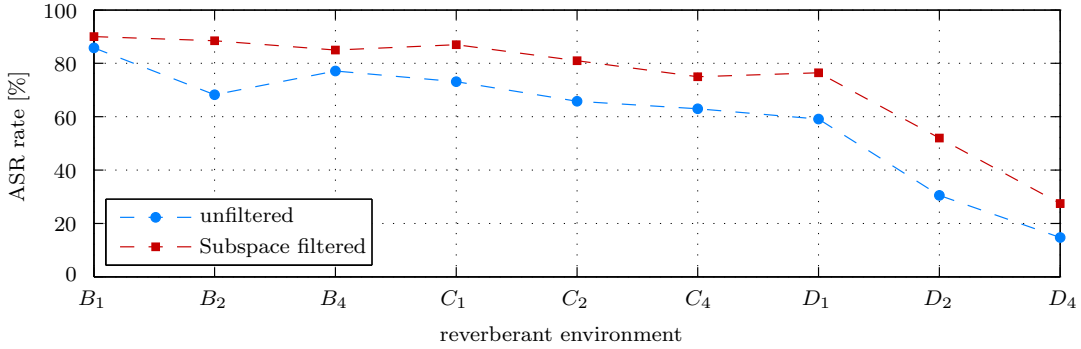
(a) Comparison of the ASR rates.



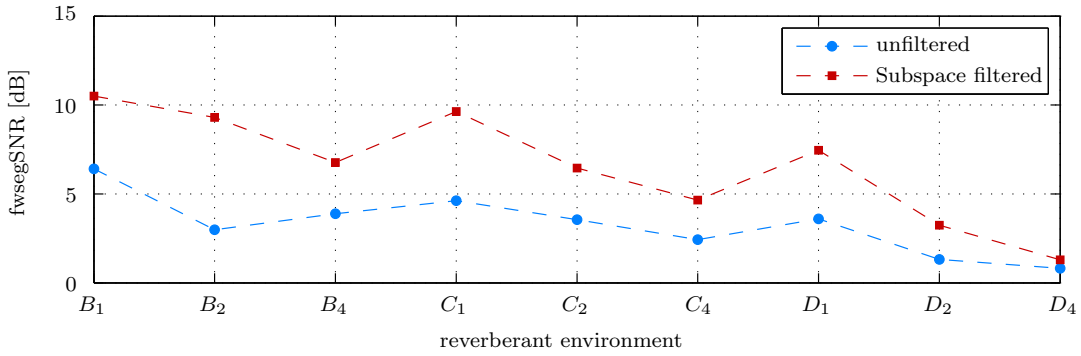
(b) Comparison of the fwsegSNR values.

Figure 4.4: Comparison of the ASR rates and fwsegSNR values for one speaker in different reverberant environments before and after filtering through the Subspace Method.

In summary, these results show that by applying the Subspace Method for one speaker, improvements regarding both ASR rates and fwsegSNR can be achieved. Hereby, the best performance increases are achieved for signals with high reverberation times or additive white Gaussian noise.



(a) Comparison of the ASR rates.



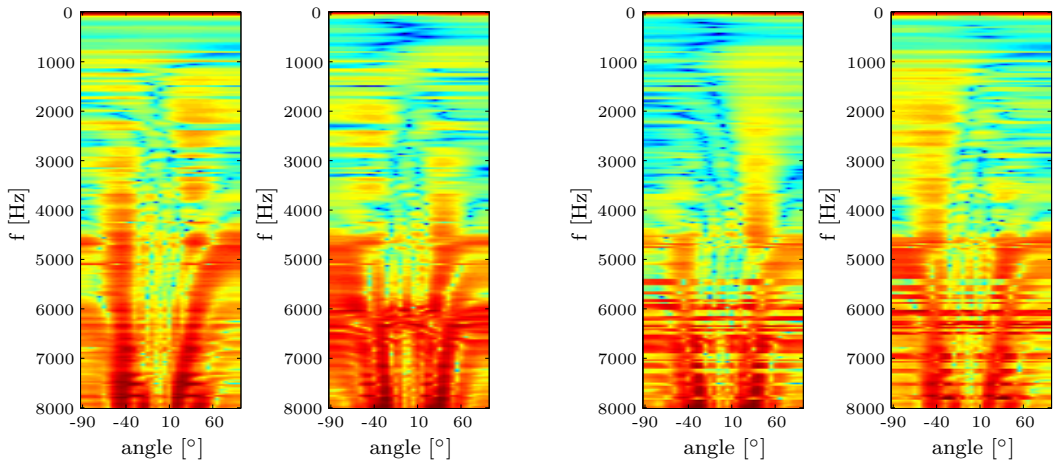
(b) Comparison of the fwsegSNR values.

Figure 4.5: Comparison of the ASR rates and fwsegSNR values for one speaker in different reverberant environments with 20 dB additive white Gaussian noise before and after filtering through the Subspace Method.

4.3 Evaluation in the Blind Source Separation Scenario

Before investigating the results for the BSS scenario, one of the occurring obstacles for two or more speakers, namely the permutation problem, is described. Therefore, recall Property 3 of the Subspace Method, explained in Chapter 3, which defined the signal subspace as $\mathbf{E}_s = [\mathbf{e}_1, ..., \mathbf{e}_D]$. Thereby, it was stated that $\mathbf{E}_s$ consists of the eigenvectors corresponding to the D dominant eigenvalues. However, the order of those eigenvectors is dependent on the eigenvalues and as a result arbitrary permutation within the filter matrix $\mathbf{W} = \mathbf{\Lambda}_s^{-1/2}\mathbf{E}_s^H$ arises. Consequently, as the Subspace Method acts as a self-organizing beamformer, for different frequencies different speaker positions are focused. This phenomenon is visualized in the beampattern displayed in Figure 4.6a which shows the angles for which the Subspace Filter attenuates or amplifies signal components. Note, that especially the frequencies below $4 - 5$ kHz are of interest because most of the speech energy is concentrated in this range. The approach to solve permutation chosen in this thesis is sorting the subspace filters such that for each of the $D = 2$ beampatterns the maximum in each frequency points to the same direction. The result of this approach is depicted in Figure 4.6b, showing decent performance below 4.5 kHz, where the beams at $+45°$ and $-45°$ are clearly visible.

In the following, four different scenarios are investigated. Firstly, as a reference, the mixed microphone signal before applying any filtering or separation is regarded. As the BSS only requires two input signals, in this case the microphone signals 1 and 6 are taken. Moreover, the BSS performance is investigated with and without employing the Subspace Method as a preprocessor. Finally, as mentioned in Chapter 3, the Subspace Method works as a self-organizing beamformer and therefore the impact of only applying the Subspace Method to the mixture of signals is taken into account as well. The BSS system has two output channels and ideally each output only contains



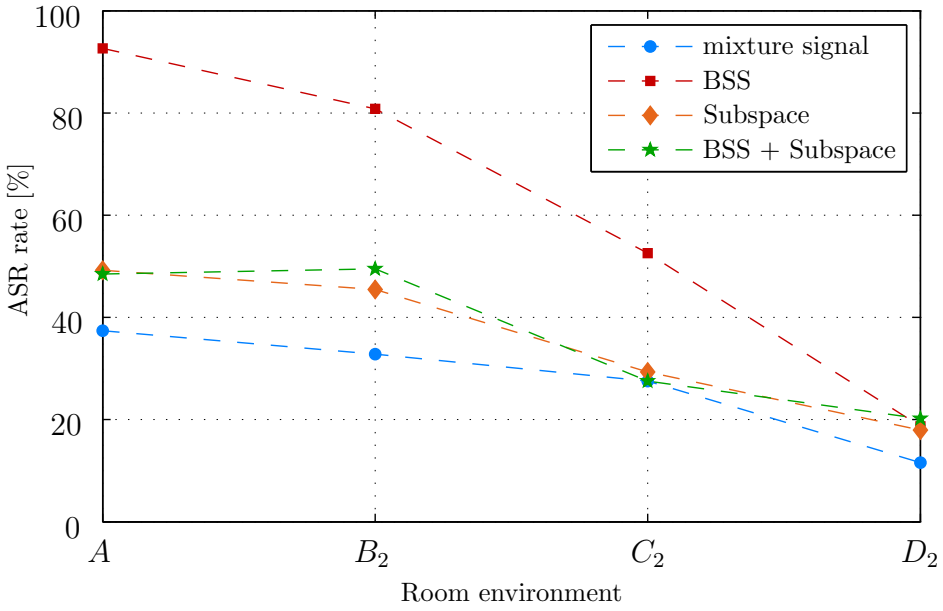
(a) Unsorted filter matrices.

(b) Sorted filter matrices.

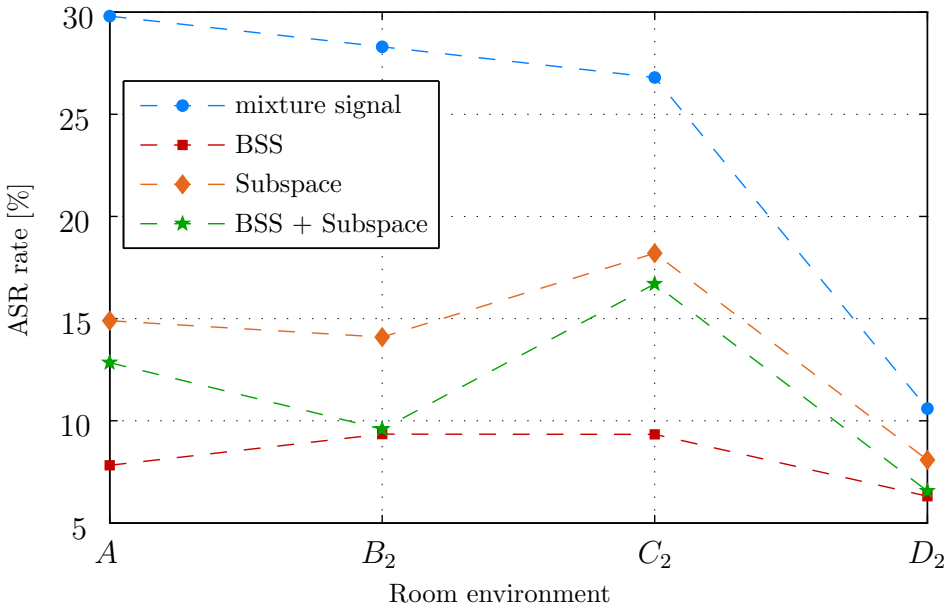
Figure 4.6: Beampattern for subspace filter **W** in room *A*.

one of the desired source signals. In order to verify the suppression, in addition to measuring the ASR rates of the desired source signal, the ASR rates of the undesired source are also regarded. In contrast to the desired source, the ASR rates for the undesired source are aimed to be small. The following charts depict the average ASR rates of both output channels for the desired and the undesired source, respectively. The exact numbers for each output channel can be found in Tables A.4 to A.7.

Figure 4.7 shows the measurements of the ASR rates for different room environments. First of all, referring to Section 2.2, Figure 4.7a visualizes the impact of reverberation on BSS performance as the ASR rates show a tendency to fall with increasing reverberation times. Interestingly, the sole BSS algorithm clearly outperforms the Subspace Method as well as the combination of the Subspace Method and the BSS. Furthermore, it is striking that the two last-named approaches, both superior to the mixed signal, do not show any significant difference in ASR rates to each other. Only for room *D* all filtering approaches result in equal results. A similar tendency can be observed in Figure 4.7b regarding the undesired sources where the BSS again shows the best signal suppression resulting in the lowest ASR rates, followed by the combination of BSS and Subspace Method.



(a) Average ASR rates for the desired source.



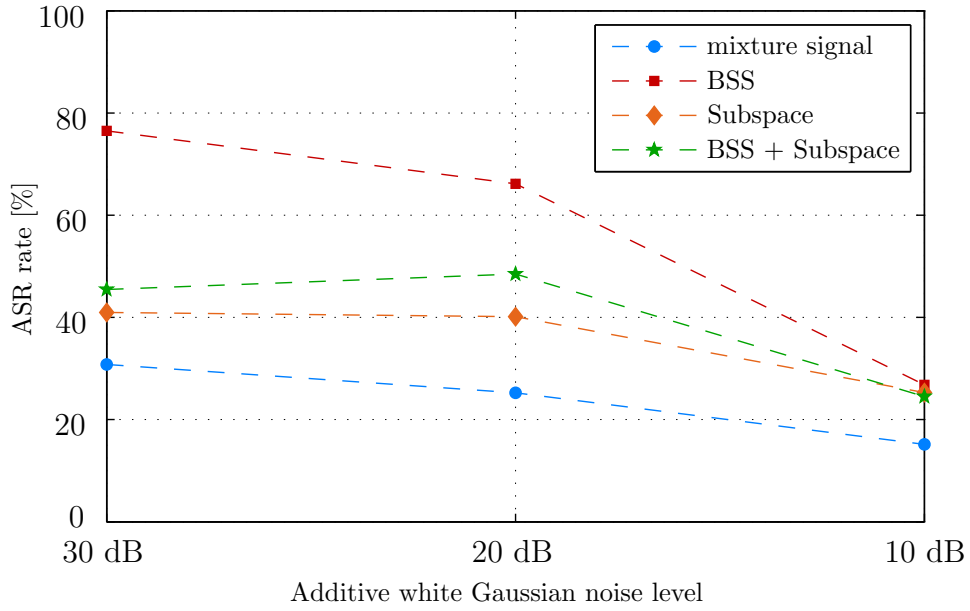
(b) Average ASR rates for the undesired source.

Figure 4.7: Average ASR rates for desired and undesired sources in different room environments.

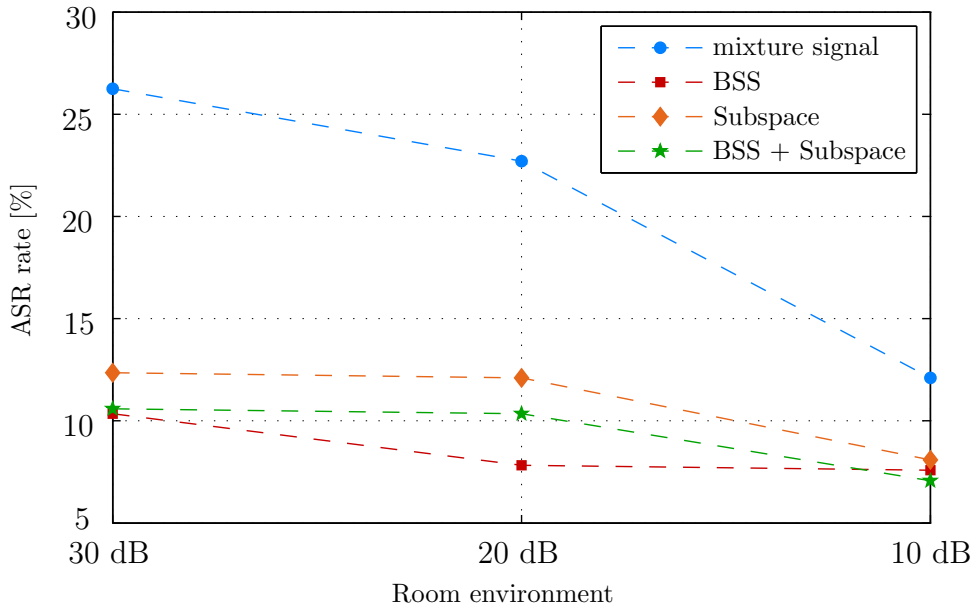
In Figure 4.8 the results for the experiments in room C with different additive white Gaussian noise levels are illustrated. Again, the ASR rates of the desired source signals for the Subspace Method and the combined approach are inferior to those of the BSS algorithm, while the combined approach outperfoms the Subspace Method in this setup. This order also counts for the suppression of the undesired signal as depicted in Figure 4.8b.

One explanation for the worse performance of the Subspace Methode compared to the BSS algorithm might be the permutation problem introduced beforehand. The BSS requires directional information contained in the signals and permutation causes different frequencies approaching from different directions. Even though sorting the beampattern showed decent solutions for frequencies below 4.5 kHz, further research has to show whether more sophisticated approaches can improve the results. Another outstanding observation is the very small performance increase for the Subspace Method as a preprocessor for BSS compared to only using the Subspace Method. As the Subspace Method generally does a PCA by projecting the signal vector onto orthogonal eigenvectors, the output signals of the Subspace Method are already subject to decorrelation. Consequently, as the BSS uses the correlation between the two signals in order to decorrelate them, the improvements through BSS are minimal.

Overall, these results indicate that employing the Subspace Method as a preprocessor for time-domain BSS cannot improve BSS performance with the methods presented in this thesis.



(a) Average ASR rates for the desired source.



(b) Average ASR rates for the undesired source.

Figure 4.8: Average ASR rates for desired and undesired sources in room C_2 for different additive white Gaussian noise levels.

Chapter 5

Conclusions & Future Work

After many years of research, BSS is still a very popular approach to separate observed mixtures in order to receive the distinct source signals. One of the major problems however, that significantly effects BSS performance, are reverberation and ambient noise when the system is used in a real acoustic environment.

In this thesis, the Subspace Method, an algorithm to suppress those undesired interferences, has been employed and evaluated as a preprocessor for a BSS algorithm. Preliminary investigations have been conducted to verify the basic functionality of the Subspace Method for one speaker. Thereby, different room conditions with varying reverberation times and different white Gaussian noise levels have been tested. The results show solid improvements in terms of ASR rate for highly reverberant environments as well as for all scenarios where white noise has been applied. The fwsegSNR increased significantly for all of the regarded scenarios. Afterwards, the BSS scenario has been investigated and it has been shown that the Subspace Method itself as well as the Subspace Method applied as a preprocessor for BSS lead to limited signal separation. It should be noted, that the latter only caused slight improvements compared to

the sole Subspace Method most likely because the Subspace Method already decorrelates the signals. However, both scenarios show clearly inferior performance compared to only applying the BSS algorithm. It is assumed that the main problem causing the lower ASR rates is the permutation problem which has partially been solved by sorting the beampattern.

In summery, the findings indicate that the Subspace Method in the frequency domain does not improve the performance of the time-domain BSS algorithm with the presented methods and is therefore not applicable as preprocessor. Further research has to show whether a more precise solution for solving the permutation problem leads to better results. Moreover, another approach could be to implement the Subspace Method in the time domain in order to avoid the permutation problem.

Appendix A

Tables of Evaluation Results

Table A.1: ASR rates and fwsegSNR values for one speaker in room A for different additive white Gaussian noise levels.

Scenario	ASR rate [%]			fwsegSNR [dB]		
	unfiltered	subspace filtered	gain	unfiltered	subspace filtered	gain
$A + 20$ dB	87,2	89,5	2,3	8,86	14,10	5,24
$A + 10$ dB	61,3	81,0	19,7	1,34	6,26	4,92
$A + 5$ dB	26,6	70,0	43,4	-1,85	2,23	4,08
$A + 0$ dB	5,8	38,0	32,2	-4,21	-1,25	2,96
$A + -5$ dB	5,0	8,0	3,0	-5,70	-3,80	1,90

Table A.2: ASR rates and fwsegSNR values for one speaker with varying reverberation times at different distances.

Scenario	ASR rate [%]			fwsegSNR [dB]		
	unfiltered	subspace filtered	gain	unfiltered	subspace filtered	gain
B_1	93,8	93,5	-0,3	13,90	18,30	4,40
B_2	88,9	92,5	3,6	10,00	19,60	9,60
B_4	92,0	91,5	-0,5	9,24	12,40	3,16
C_1	90,0	91,5	1,5	11,30	20,50	9,20
C_2	87,2	86,5	-0,7	9,13	13,20	4,07
C_4	81,9	80,5	-1,4	7,19	8,31	1,12
D_1	74,8	86,5	11,7	8,08	13,70	5,62
D_2	47,8	61,0	13,2	4,40	6,72	2,32
D_4	29,6	41,5	11,9	3,19	3,65	0,46

Table A.3: ASR rates and fwsegSNR values for one speaker with varying reverberation times and 20 dB additional white Gaussian noise at different distances.

Scenario	ASR rate [%]			fwsegSNR [dB]		
	unfiltered	subspace filtered	gain	unfiltered	subspace filtered	gain
$B_1 + 20$ dB	85,8	90,0	4,2	6,41	10,50	4,09
$B_2 + 20$ dB	68,2	88,5	20,3	2,98	9,30	6,32
$B_4 + 20$ dB	77,1	85,0	7,9	3,89	6,77	2,88
$C_1 + 20$ dB	73,1	87,0	13,9	4,62	9,64	5,02
$C_2 + 20$ dB	65,8	81,0	15,2	3,55	6,46	2,91
$C_4 + 20$ dB	63,0	75,0	12,0	2,43	4,65	2,22
$D_1 + 20$ dB	59,1	76,5	17,4	3,60	7,45	3,85
$D_2 + 20$ dB	30,5	52,0	21,5	1,32	3,24	1,92
$D_4 + 20$ dB	14,7	27,5	12,8	0,82	1,29	0,47

Table A.4: Scenario: no BSS or Subspace Method.
Distinct ASR rates [%] of desired and undesired sources for two simultaneously active speakers for different reverberant and noisy scenarios.

Scenario	desired signal		average	undesired signal		average
	source 1	source 2		source 1	source 2	
A	35,9	38,9	37,4	31,3	28,3	29,8
B_2	33,8	31,8	32,8	25,3	31,3	28,3
C_2	27,3	27,8	27,55	26,3	27,3	26,8
D_2	12,1	11,1	11,6	10,1	11,1	10,6
B_2 + 30 dB	32,8	28,8	30,8	23,7	28,8	26,25
B_2 + 20 dB	25,8	24,7	25,25	23,2	22,2	22,7
B_2 + 10 dB	14,1	16,2	15,15	11,6	12,6	12,1

Table A.5: Scenario: BSS only.
Distinct ASR rates [%] of desired and undesired sources for two simultaneously active speakers for different reverberant and noisy scenarios.

Scenario	desired signal		average	undesired signal		average
	source 1	source 2		source 1	source 2	
A	92,9	92,4	92,65	8,08	7,58	7,83
B_2	77,8	83,8	80,8	9,09	9,6	9,345
C_2	49,0	56,1	52,55	10,6	8,08	9,34
D_2	17,2	20,2	18,7	7,58	5,05	6,315
B_2 + 30 dB	74,7	78,3	76,5	10,6	10,1	10,35
B_2 + 20 dB	63,6	68,7	66,15	7,58	8,08	7,83
B_2 + 10 dB	28,3	25,3	26,8	9,6	5,56	7,58

Table A.6: Scenario: Subspace Method only.
Distinct ASR rates [%] of desired and undesired sources for two simultaneously active speakers for different reverberant and noisy scenarios.

Scenario	desired signal		average	undesired signal		average
	source 1	source 2		source 1	source 2	
A	54,5	43,9	49,2	16,2	13,6	14,9
B_2	43,9	47,0	45,45	13,6	14,6	14,1
C_2	26,3	32,3	29,3	17,7	18,7	18,2
D_2	17,2	18,7	17,95	7,6	8,58	8,09
B_2 + 30 dB	46,0	35,9	40,95	13,1	11,6	12,35
B_2 + 20 dB	40,9	39,4	40,15	9,6	14,6	12,1
B_2 + 10 dB	25,3	25,3	25,3	7,58	8,59	8,085

Table A.7: Scenario: Subspace Method as a preprocessor for BSS.
Distinct ASR rates [%] of desired and undesired sources for two simultaneously active speakers for different reverberant and noisy scenarios.

Scenario	desired signal		average	undesired signal		average
	source 1	source 2		source 1	source 2	
A	51,5	45,5	48,5	12,6	13,1	12,85
B_2	43,4	55,6	49,5	8,59	10,6	9,595
C_2	26,8	28,3	27,55	16,7	16,7	16,7
D_2	21,7	18,7	20,2	6,06	7,07	6,565
B_2 + 30 dB	53,0	37,9	45,45	13,6	7,58	10,59
B_2 + 20 dB	47,0	50,0	48,5	10,1	10,6	10,35
B_2 + 10 dB	22,7	26,3	24,5	7,07	7,07	7,07

Appendix B

Abbreviations and Acronyms

ASR	automatic speech recognizer
ASR rate	automatic speech recognition rate
BSS	blind source separation
DS	Delay-and-Sum
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
fwsegSNR	frequency-weighted segmental signal-to-noise ratio
LPC	linear predictive coding
MATLAB	MATrix LABoratory (C) The Mathworks, Inc.
PCA	Principal Component Analysis
RIR	room impulse response
SNR	signal-to-noise ratio
SOS	second-order statistics
STFT	short-term Fourier transform
TRINICON	Triple-N-Independent component analysis for convolutive mixtures

Appendix C

Notation

C.1 Notation in General

f_s	sampling frequency
ω	normalized frequency
t	block index
T_{60}	Reverberation time
$\mathfrak{R}(\mathbf{A})$	vector space spanned by the column vectors of the matrix $\mathbf{A}$

C.2 Conventions

In this thesis the following conventions are used:

- Vectors are denoted by lower-case, bold-faced symbols.
- Matrices are denoted by upper-case, bold-faced symbols.
- Estimates are denoted by $\hat{(\cdot)}$, e.g., $\hat{x}$.

C.3 Mathematical Operators

$*$	Time-domain convolution operator
$(\cdot)^T$	Transpose of $(\cdot)$
$(\cdot)^H$	Conjugate transpose of $(\cdot)$
$(\cdot)^{-1}$	Inverse of $(\cdot)$
$\frac{\partial}{\partial \mathbf{W}}$	Partial derivative of $(\cdot)$ with respect to $\mathbf{W}$
bdiag	sets all submatrices on the off-diagonals of a matrix to zero
$\det\{\cdot\}$	Determinant of a matrix
$\text{IDFT}\{\cdot\}$	Inverse Discrete Fourier Transform

Bibliography

[1] R. Aichner, H. Buchner, and W. Kellermann, "Comparison and a theoretical link between time-domain and frequency-domain blind source separation," 2003.

[2] R. Aichner, H. Buchner, F. Yan, and W. Kellermann, "A real-time blind source separation scheme and its application to reverberant and noisy acoustic environments," *Signal Processing*, vol. 86, no. 6, pp. 1260–1277, 2006.

[3] F. Asano, S. Ikeda, M. Ogawa, H. Asoh, and N. Kitawaki, "Combined approach of array processing and independent component analysis for blind separation of acoustic signals," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 204–215, 2003.

[4] H. Barfuss, "Evaluation of geometrically constrained independent component analysis as alternative to lcmv," Master thesis, University of Erlangen-Nuremberg, 2013.

[5] J. Barker, E. Vincent, N. Ma, H. Christensen, and P. D. Green, "The PASCAL chime speech separation and recognition challenge," *Computer Speech & Language*, vol. 27, no. 3, pp. 621–633, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2012.10.004

[6] H. Buchner, R. Aichner, and W. Kellermann, "TRINICON: a versatile framework for multichannel blind signal processing," in *2004 IEEE International Conference*

on Acoustics, Speech, and Signal Processing, vol. 3. IEEE, 2004, pp. iii–889–92. [Online]. Available: http://dx.doi.org/10.1109/icassp.2004.1326688

[7] H. Buchner, R. Aichner, and W. Kellermann, "The trinicon framework for adaptive mimo signal processing with focus on the generic sylvester constraint," in *Proc. ITG Conference on Speech Communication*, Aachen, Germany, October 2008.

[8] H. Buchner, R. Aichner, and W. Kellermann, "Blind source separation for convolutive mixtures: A unified treatment," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y. Huang and J. Benesty, Eds. Springer US, 2004, pp. 255–293. [Online]. Available: http://dx.doi.org/10.1007/1-4020-7769-6_10

[9] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[10] A. Cantoni and L. C. Godara, "Resolving the direction of sources in a correlated signal field incident on an array," *Journal of the Acoustical Society of America*, vol. 67, pp. 1247–1255, Apr. 1980.

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

[12] E. A. P. Habets, "Single- and multi-microphone speech dereverberation using spectral enhancement," Ph.D. dissertation, Technische Universiteit Eindhoven, 2007. [Online]. Available: http://alexandria.tue.nl/extra2/200710970.pdf

[13] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement." *IEEE Transactions on Audio, Speech & Language Processing*,

vol. 16, no. 1, pp. 229–238, 2008. [Online]. Available: http://dblp.uni-trier.de/db/journals/taslp/taslp16.html#HuL08

[14] A. Hyvaerinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608000000265

[15] D. H. Johnson and D. E. Dudgeon, *Array signal processing : concepts and techniques / Don H. Johnson, Dan E. Dudgeon.* P T R Prentice Hall Englewood Cliffs, NJ, 1993.

[16] M. Kawamoto, A. Barros, K. Matsuoka, and N. Ohnishi, "A method of real-world blind separation implemented in frequency domain," *ICA 2000*, pp. 267–272, 2000.

[17] W. Kellermann, "Speech and Audio Signal Processing," Lecture notes, University of Erlangen-Nuremberg, Summer term 2014.

[18] T.-W. Lee, "Independent component analysis," in *Independent Component Analysis.* Springer US, 1998, pp. 27–66. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-2851-4_2

[19] S. Pillai and C. Burrus, *Array signal processing*, ser. Signal Processing and Digital Filtering. Springer-Verlag, 1989.

[20] R. Roy and T. Kailath, "Esprit-estimation of signal parameters via rotational invariance techniques," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 7, pp. 984–995, Jul 1989.

[21] A. Schwarz, "Blind performance measures for a bss-based acoustic source localization and separation scheme," Seminar paper, University of Erlangen-Nuremberg, 2008.

[22] G. Strang, *Linear Algebra and its applications.* Harcourt Brace Jovanovich Publishers, 1988.

[23] I. Tashev, "Revereberation reduction for improved speech recognition," in *in Proc. of HSCMA*, 2005.

[24] W. Walker, P. Lamere, P. Kwok, B. Raj, R. Singh, E. Gouvea, P. Wolf, and J. Woelfel, "Sphinx-4: A flexible open source framework for speech recognition," Tech. Rep., 2004.

[25] T. Wurzbacher, "Eigenraum-beamforming und adaptive realisierungen," Diploma thesis, University of Erlangen-Nuremberg, 2003.